Author:
Madley-Dowd, Paul C

Title:
Assessing causality of the association between maternal smoking during pregnancy and offspring intellectual disability

Assessing causality of the association between maternal smoking during pregnancy and offspring intellectual disability

Paul Madley-Dowd

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Health Sciences, Bristol Medical School, October 2020.

Word count: 56,275

Abstract

Maternal smoking has known adverse effects on fetal development. However, research on the association between maternal smoking during pregnancy and offspring intellectual disability (ID) is limited, and whether any associations are due to a causal effect or residual confounding is unclear.

The association was investigated using two intergenerational cohorts, each with over 1 million individuals with data held in the Danish and Swedish registers, and using a prospectively collected pregnancy cohort of approximately 15,000 mother and child pairs in the United Kingdom. Observational analyses were performed in each cohort using regression analyses adjusted for potential confounders. Sibling comparisons, negative control analyses, and Mendelian randomisation (MR) were used to provide evidence about the causal nature of the association. Simulation studies were conducted to investigate the nature of biases arising from assortative mating in the negative control design and from the proportion of missing data in multiple imputation analyses, a method to account for missing data.

Observational analyses provided evidence for an increased risk of ID in children of mothers who smoked during pregnancy. Sibling comparison models decomposed this population averaged effect to reveal a null within-family effect while negative control analyses showed comparable effect estimates for maternal and paternal smoking during pregnancy. MR analyses also did not provide evidence of a causal effect.

The results of analyses contained within this thesis are not consistent with a causal effect of maternal smoking during pregnancy on offspring risk of ID. By combining evidence across different analysis methods, results suggest that prior observational associations were the result of unmeasured genetic or environmental characteristics of families in which the mother smokes during pregnancy.

Dedication and acknowledgements

I would like to start by thanking my supervisors. Dheeraj, your encouragement and optimism throughout have made this journey far easier than it otherwise would have been. Your ability to sift the good ideas from the bad has helped sway me from many a rabbit hole and is something I hope to eventually learn. Jon and Stan, I'd like to thank you for giving so much of your time, knowledge and patience. In this you allowed me to develop some very rough and scattered ideas into fully formed projects which I am now very proud of. It has been a pleasure working with all of you these last four years.

To Rachael Hughes and Kate Tilling, thank you for your work, guidance and support on my first publication. This was my first real taste of research and I thoroughly enjoyed it. I will always be blown away by the detail of your knowledge about data that aren't there. To Diana Schendel and Amy Kalkbrenner, thank you for making me feel so at home in Aarhus and for all the work you put in to making the research visit happen. To Cecilia Magnusson, Micke Lundberg and Christina Dalman, thank you for your hospitality in Stockholm and for jumping through hoops to get me access to the Swedish registry data. To Richard Thomas and Andrew Boyd, thank you for your work and persistence in making the ALPSAC data linkage a reality.

I'd like to thank many colleagues and fellow PhD students including Alexandra Creavin, Amanda Ly, Christina Dardani, Doretta Caramaschi, Hannah Jones, Hannah Sallis, Hein Heuvelman, Jamie Robinson, Jazz Croft, Kayleigh Easey, Maki Morinaga, Ryan Langdon, Tom Battram, and Viktor Ahlqvist. I have learned so much from each of you.

Thank you to everyone at the University of Bristol Spartans Basketball Club, particularly Sam Davis, for providing me with something to focus on outside of work. To the Junto – Cameron, James and Shaun – thank you for listening and for your balanced advice. To Emily Sanderson, Jess Laidlaw and Rob Larson, you are three of the kindest people I know. Thank you for always being that way. To Das Monkeys – Cameron, Harvey, Matt, Scott and Tom – our trips simultaneously delayed the completion of this work and kept it going. I am very grateful to have known each of your for so long. Paul Moore requested to feature in this thesis. He contributed the occasional pint and pretended to fall asleep when I described my work to him. It's been a pleasure knowing him also.

Finally, and most importantly, to Mum, Dad and Lauren, I hope you already know how grateful I am to each of you for everything along the way. Thank you for reading through my work and for listening when it wasn't going all that well. Thank you for helping me get to where I am today.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Paul Madley-Dowd DATE: 1st October 2020

Publications

The work carried out as part of this thesis has contributed to the following publications:

Published

- **Madley-Dowd P.**, Hughes R., Tilling K., and Heron J., The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of clinical epidemiology, 2019. 110: p. 63-73. DOI: https://doi.org/10.1016/j.jclinepi.2019.02.016
- **Madley-Dowd P.**, Rai D., Zammit S., and Heron J., Simulations and directed acyclic graphs explained why assortative mating biases the prenatal negative control design. Journal of clinical epidemiology, 2020. 118: p. 9-17. DOI: https://doi.org/10.1016/j.jclinepi.2019.10.008
- **Madley-Dowd P.**, Kalkbrenner A.E., Heuvelman H., Heron J., Zammit S., Rai D., and Schendel D., Maternal smoking during pregnancy and offspring intellectual disability: sibling analysis in an intergenerational Danish cohort. Psychological medicine, 2020: p. 1-10. DOI: https://doi.org/10.1017/S0033291720003621

Being revised following review

- **Madley-Dowd P.**, Lundberg M., Heron J., Zammit S., Ahlqvist V.H., Magnusson C., and Rai D., Maternal smoking and smokeless tobacco use during pregnancy and offspring development: sibling analysis in an intergenerational Swedish cohort. International journal of epidemiology.

Planned papers

- **Madley-Dowd P.**, Zammit S., Heron J., Boyd A., Thomas R., and Rai D., Intellectual disability in the children of the Avon longitudinal study of parents and children.
- **Madley-Dowd P.**, Zammit S., Heron J., Boyd A., Thomas R., and Rai D., Maternal smoking in pregnancy and offspring intellectual disability: a causal analysis in the Avon Longitudinal Study of Parents and Children

Additional publications not included in this thesis

- Kalkbrenner A.E., Meier S.M., **Madley-Dowd P.**, Ladd-Acosta C., Daniele Fallin M., Partner E., and Schendel D., Familial confounding of the association between maternal smoking in pregnancy and autism spectrum disorder in offspring. Autism research, 2020. 13: p. 134-144. DOI: https://doi.org/10.1002/AUR.2196

Contributions to publications with content used in this thesis

The proportion of missing data should not be used to guide decisions on multiple imputation.

CRediT authorship contribution statement from article:

- *Paul Madley-Dowd: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. Rachael Hughes: Conceptualization, Methodology, Writing - review & editing, Supervision. Kate Tilling: Conceptualization, Methodology, Writing- review & editing, Supervision. Jon Heron: Conceptualization, Methodology, Writing - review & editing, Supervision.*

Simulations and directed acyclic graphs explained why assortative mating biases the prenatal negative control design.

CRediT authorship contribution statement from article:

- *Paul Madley-Dowd: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Visualization. Dheeraj Rai: Conceptualization, Methodology, Writing - review & editing, Supervision. Stanley Zammit: Conceptualization, Methodology, Writing - review & editing, Supervision. Jon Heron: Conceptualization, Methodology, Writing - review & editing, Supervision.*

Maternal smoking during pregnancy and offspring intellectual disability: sibling analysis in an intergenerational Danish cohort.

CRediT authorship contribution statement:

- *Paul Madley-Dowd: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Visualization. Amy E. Kalkbrenner: Conceptualization, Methodology, Writing - review & editing, Supervision. Hein Heuvelman: Methodology, Writing - review & editing, Supervision. Jon Heron: Methodology, Writing - review & editing, Supervision. Stan Zammit: Methodology, Writing - review & editing, Supervision. Dheeraj Rai: Conceptualization, Methodology, Writing - review & editing, Supervision. Diana Schendel: Conceptualization, Methodology, Writing - review & editing, Supervision.*

Table of contents

List of tables

List of figures

List of Abbreviations

ADHD	Attention deficit hyperactivity disorder
ALSPAC	Avon Longitudinal Study of Parents and Children
AOI	Association of interest
ASD	Autism spectrum disorder
BMI	Body mass index
CAG	Confidential advisory group
CCA	Complete case analysis
CI	Confidence interval
DAG	Directed acyclic graph
DPR	Danish psychiatric registry
DSM-5	Diagnostic And Statistical Manual Of Mental Disorders, 5th Edition
DSM-IV	Diagnostic And Statistical Manual Of Mental Disorders, 4th Edition
FCS	Fully conditional specification
FMI	Fraction of missing information
GEE	Generalised estimating equation
GP	General practitioner
GSCAN	GWAS and sequencing consortium of alcohol and nicotine
GWAS	Genome wide association study
HES	Hospital episode statistics
ICD	International classification of disease
ICD-10	International classification of diseases, version 10
ICD-8	International classification of diseases, version 8
ID	Intellectual disability
IDI	Identification of developmental impairments
InSIDE	Instrument strength independent of direct effect
iPSYCH	Integrative psychiatric research consortium
IQ	Intelligence quotient
IRAS	Integrated research application system
IVW	Inverse variance weighted
LISA	Longitudinal integration database for health insurance and labour market studies
MAR	Missing at random
MBR	Medical birth registry
MCAR	Missing completely at random
MCS	Millennium cohort study

MGR	Multi-generation register
MHSDS	Mental health services data set
MI	Multiple imputation
MICE	Multivariate imputation using chained equations
MNAR	Missing not at random
MR	Mendelian randomisation
nAChR	Nicotinic acetylcholine receptors
NCA	Negative control association
NCDS	National child development study
NPR	National patient registry
OR	Odds ratio
PDD	Pervasive developmental disorders
PLASC	Pupil level annual school census
PRS	Polygenic risk score
SD	Standard deviation
SE	Standard error
SEN	Special educational needs
SES	Socioeconomic status
SGA	Small for gestational age
SNP	Single nucleotide polymorphism
SUTVA	Stable unit treatment value assumption
UKSeRP	UK secure eResearch platform
WAIS	Weschler abbreviated scale of intelligence
WISC	Weschler intelligence scale for children

Chapter 1 Introduction

Smoking in pregnancy is reported in at least 10% of pregnancies in Europe as of 2010 [1]. It has a well-established causal relationship with low birthweight [2-9] and a more tentative association with other poor pregnancy and offspring health outcomes such as pregnancy complications [10] and sudden infant death syndrome [11]. Establishing which offspring health outcomes are caused by maternal smoking in pregnancy may (i) provide insight as to which disease burdens may be reduced through smoking cessation initiatives, (ii) aid in understanding the mechanisms by which these conditions occur and (iii) create the opportunity for mothers to have an informed choice about the potential consequences of deciding to, or not to, give up smoking during pregnancy.

An association between maternal smoking during pregnancy and offspring risk of intellectual disability has been suggested in the literature. In this chapter I define intellectual disability, describe the biological plausibility of the association and review the body of epidemiological evidence. As will be shown, the causal nature of such an association is unclear and requires further study.

1.1 – Intellectual disability

1.1.1 – Definition

The outcome of interest in this thesis is offspring intellectual disability (ID). ID is defined as having an arrested or incomplete development of the mind alongside functional impairment in facets that contribute to overall intelligence such as cognition, language and social ability [12]. ID manifests during the developmental period and is not the result of later changes to the brain as a result of injury or disease.

There are several challenges in defining ID in practice, particularly in relation to the language used. Several terms are used in the UK including learning disability, learning difficulties, developmental disorder (or delay) and special educational needs [13]. Confusion can arise as these phrases are components of other, separate concepts. For example specific learning disability refers to dyslexia or dyscalculia, while learning difficulty can refer to intellectual

disability or a specific learning disability. It is important to note that those with ID may also have a specific learning disability. Further challenges arise in the definitions used between studies based in different global regions. In the USA the phrase “intellectual disability” carries the same meaning as “learning disability” in the UK, while use of the phrase “learning disability” in the USA refers to what would be described as a “specific learning disability” in the UK.

In a healthcare setting, several diagnostic criteria including the International Classification of Diseases, Version 10 (ICD-10) [12] and Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) [14] define ID using an intelligence quotient (IQ) score of less than 70, equivalent to 2 standard deviations (SD) less than the population average of 100, alongside functional impairments. The Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) [15], states that IQ tests will generally be measured with an error of around 5 points and therefore scores between 65 and 75 may indicate ID. The definition used will greatly affect the prevalence of ID in studies. Cooper et al. [16] note that statistically, 2.5% of people in the population would be expected to lie in the region of IQ scores between 70 and 75 which is greater than the 2.28% prevalence of ID expected using a cut off that is two SDs below the population average. The educational system in the UK uses an even less stringent cut off, IQ less than 85 (equivalent to 1 standard deviation lower than the population average), to indicate “mild learning difficulty” [17, 18].

It has been argued that ID should not be defined on the basis of IQ test scores alone [18, 19] due to the instability of the measure on the basis of mood and fatigue, potential to be influenced by learning or rehearsal, and tests that are largely centred around Western cultural understanding that may have important implications, particularly for migrants. The ICD-10 and DSM-5 also use social functioning and age of onset for diagnosis. Those who have an IQ less than 70 but are able to function without assistance by this definition are not considered to have ID in relation to clinical services. Cooper et al. [16] provide examples such as living independently and holding a job as meeting this criteria of functioning without assistance. Such a definition means that ID is not necessarily stable throughout the life. Those with ID do learn throughout the lifetime, and some of those who may require significant support during school age years may go on to learn to live independently.

There is inherent heterogeneity in the concept of ID which may differ by the way that ID has been defined. As a result investigations into the aetiology of ID may not be comparable where different definitions have been used.

1.1.2 – Epidemiology of ID

The global prevalence of ID is approximately 1% according to a meta-analysis of 52 population based studies [20]. As described above, this estimate is likely to reflect people with a severe enough ID to receive clinical services. Prevalence differs by country and is strongly influenced by socioeconomic development and methods of data capture, though whether this is the result of differing distributions of the determinants of ID or differences in diagnostic practice is unclear. People with ID may require different pathways of care, such as health care and social care, and therefore multi-source case ascertainment may help capture the full population of ID as in other neurodevelopment conditions [21, 22].

Individuals with ID have been found to suffer from poor long-term outcomes and inequalities compared to the general population including socioeconomic disadvantage [23, 24], increased mortality [25] and worse access to, and effectiveness of, health care [26-29]. Comorbidities are common, particularly with other neurodevelopmental conditions such as Attention Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD). Recent evidence has suggested that severe ID may represent a distinct condition to mild and moderate ID, which in turn may represent the extreme low of the normal range of the IQ spectrum [30].

1.2 – Biological plausibility of the association between maternal smoking during pregnancy and offspring ID

The key biological mechanisms by which maternal smoking during pregnancy may cause offspring intellectual disability include: (i) hypoxia-ischemia induced brain damage and (ii) alterations to normal fetal brain development following changes in the activation of nicotinic acetylcholine receptors (nAChR). A reduction in the availability of oxygen in the blood (hypoxia) can lead to tissue damage [31]. Cells in the brain are particularly susceptible to hypoxic damage in comparison to those in the rest of the body. A restriction in blood flow (ischemia) will lead to hypoxia, however hypoxia can also result from other causes. Smoking has been found to be associated with reduced uterine blood flow [32, 33] and increases in fetal blood concentrations of carboxyhaemoglobin [34], the result of carbon monoxide binding to haemoglobin which prevents oxygen from being transported. Exposure to maternal smoking during pregnancy may therefore lead to hypoxic damage to fetal brain tissue via two separate mechanisms.

Nicotine, the psychoactive component in tobacco smoke, has been shown to cross the placenta and expose the fetus at higher concentrations than the mother [35, 36]. Once nicotine has reached the fetus it acts on nAChR which, in animal models, have been shown to influence developmental processes in the brain including neurogenesis, migration, differentiation, and synaptogenesis (see the review by Dwyer et al. [37]). Nicotine exposure has been found to affect fetal brain development even in the absence of fetal growth restriction [38] suggesting that the influence of nicotine is targeted to the brain.

Chan and colleagues [39], however, state that the field is unclear as to whether the biological consequences of maternal smoking in pregnancy are due to nicotine or to the effects of cigarette smoke. They argue that of the many chemical compounds in cigarette smoke, no single component is likely to be responsible for all pathology. Oxidative stress in the brain, known to be caused by cigarette smoking, has been shown to block activation of nAChRs [40]. Nicotine, in contrast, may activate or inactivate nAChRs dependent on the dose, duration and type of receptor [37]. The mechanism by which exposure to cigarette smoking in pregnancy influences changes in brain development may therefore be a complex interplay of activation and inactivation of nAChRs by the components and consequences of smoking and via other yet unidentified pathways.

The evidence supporting an influence of maternal smoking during pregnancy on changes to fetal brain development largely comes from animal models. The generalisability of this evidence to an effect in humans is limited due to: (i) differences in the exposure administered in experiments versus that seen in practice among smokers (i.e. continuous versus intermittent exposure), (ii) differences in the brain structure of animals from humans and (iii) differing developmental periods of the brain, for example the third trimester of human fetal development corresponds to the postnatal period in rats [41].

In the human based literature, a number of studies have shown that smoking in pregnancy is associated with reduced head circumference at birth (collated in the review by Ekblad and colleagues [42]), though this is a particularly crude measure of brain changes as a result of smoking during pregnancy. Evidence for structural and functional smoking related brain changes has been collated in a literature review by Bublitz and Stroud [43]. Reductions in fetal brain volumes associated with maternal smoking in pregnancy were found in areas such as the cerebellum [44, 45], frontal lobe [44, 46], parietal lobe [46], ventricular system [45], corpus callosum [47-49] and cerebral cortex in general [50]. Analysis of functional changes among people exposed to smoking during pregnancy showed increased activation in the brainstem in response to auditory stimuli [51, 52] as well as greater activation in the frontal [53], temporal [53-55], parietal [53] and occipital lobes [54] and the cerebellum [53] in response to executive function tasks involving inhibition, attention and memory. Given that the age of brain

measurement in some of these studies was in adolescence or young adulthood, these findings show that smoking during pregnancy can lead to persistent functional brain changes [46-50, 53-55].

Brain imaging work has suggested a particular role of the frontal lobe in contributing to intelligence and perceptual impairment in individuals with ID [56, 57] while a substantial body of work has suggested that ID results from changes to synaptic function [58] which would implicate a wider range of brain areas in the development of ID. Exposure to maternal smoking during fetal brain development may therefore lead to widespread and propagating changes in the brain that may functionally manifest themselves as ID or other neurodevelopmental disorders. Fully understanding the mechanisms by which maternal smoking during pregnancy influences risk of offspring ID is not necessary to establish whether an association reflects a causal effect, though if causality exists, knowledge of the mechanisms may be useful in guiding treatment in order to reduce the effects.

1.3 – Epidemiological literature

The evidence for an association between exposure to maternal smoking during pregnancy and persistent differences in brain structure and function has led to a great scientific interest in whether smoking in pregnancy is also associated with changes in offspring cognition and behaviour. Systematic reviews of the neurodevelopmental literature [59, 60] have shown consistent findings for associations of exposure with poor academic achievement [61-65] and behavioural problems in children [66-68], in particular ADHD [69]. Inconsistent effects of exposure to smoking during pregnancy have been found for offspring intelligence [61, 70-76], memory [66, 72, 73, 77-79], attention and executive function [66, 71, 72, 77, 79-82]. The causal nature of these associations remains unclear and may be accounted for by bias in effect estimates. In recent years evidence has been collated that suggests the absence of an association between smoking in pregnancy and offspring ASD [83-87].

The majority of epidemiological research in this area to date has focused on variations in IQ within the normal range [61, 70-76]. Very low IQ has at times been used to exclude participants in smoking studies [e.g. 88, 89], thus preventing exploration of association in this region of the distribution of intelligence. Less is therefore known about the potential impact of smoking during pregnancy on the risk of more severe and debilitating cognitive impairments, such as those present in ID. Below, I provide a review of the available evidence, specific to the association between maternal smoking in pregnancy and offspring risk of ID. This is followed by a gathering of evidence for the potential confounders for the association.

Table 1-1: Summary of previous studies on the association between maternal smoking during pregnancy and offspring ID

Study	Study design	Population	Sample size	Exposure definition	Definition of ID	Definition of control	Estimate (95% CI)	Estimate measure
Drews et al. 1996	Case-control	Children aged 10, born in the Metropolitan Atlanta area between 1975 and 1976	221 cases, 400 controls	Retrospectively collected smoking at any time during pregnancy	IQ < 70 on the most recent psychometric test	Random selection of children of the same age and area attending public school	1.74 (1.21-2.50) unadjusted 1.56 (1.01-2.41) adjusted for confounders	Odds ratio
Roeleveld et al. 1992	Case-control	Referrals to the Paediatric or Child Neurology departments of the Nijmegen University Hospital (Netherlands) or to nearby rehabilitation centres	306 cases, 322 controls	Retrospective collection of average number of cigarettes smoked per day in pregnancy	ICD-9 code 317-319 with IQ less than 80 and no specific cause	Children with congenital physical handicaps of known cause	1.1 (0.8-1.5) unadjusted	Odds ratio
Mann et al. 2009	Prospective birth cohort	Births between 1996 and 2002 in the South Carolina Medicaid billing records	134,596 births (1089 cases)	Tobacco use in pregnancy recorded on birth certificate	(i) ICD-9 code 317-319 in the Medicaid billing records, or (ii) record of mental handicap in data file from the department of education, or (iii) enrolment in services from the department of disabilities and special needs		1.00 (0.94-1.07) unadjusted	Odds ratio

Study	Study design	Population	Sample size	Exposure definition	Definition of ID	Definition of control	Estimate (95% CI)	Estimate measure
Mann et al. 2013	Prospective birth cohort	Births between 2000 and 2007 in the South Carolina Medicaid billing records	165,311 births (629 cases)	Tobacco use in pregnancy recorded on birth certificate	(i) ICD-9 code 317-319 on at least 5 unique occasions in the Medicaid billing records, or (ii) record of mental handicap in data file from the department of education, or (iii) enrolment in services from the department of disabilities and special needs		1.14 (1.07-1.22) unadjusted	Odds ratio
Huang et al. 2016	**Meta analysis**	**Drews et al., Roeleveld et al., Mann et al. and Mann et al.**	**-**	**-**	**-**	**-**	**1.10 (1.06, 1.15)**	**Odds ratio**
Braun et al. 2009	Case-control	Children living within the US based Centres for Disease Control and Prevention's Autism and Developmental Disability Monitoring Network that were age 8 in 2002 or 2004	965 cases, 104,607 controls	Tobacco use in pregnancy recorded on birth certificate	IQ < 70 in health or education records or written letter from clinician based on previous psychometric testing if no IQ data available	Children of the same age within the surveillance network	1.52 (1.27, 1.83) unadjusted 1.12 (0.92-1.36) adjusted for confounders	Risk ratio

Study	Study design	Population	Sample size	Exposure definition	Definition of ID	Definition of control	Estimate (95% CI)	Estimate measure
Lundberg et al. 2010	Prospective birth cohort	Singleton males born between 1883 and 1988 with Nordic mothers	205,777 births (~15,000 cases)	Smoking in pregnancy (1-9 cigarettes per day, 10+ cigarettes per day) recorded in the Swedish medical birth register	Score <= 2 on the Swedish conscript intellectual performance test		1.70 (1.63-1.78) unadjusted 1.27 (1.19-1.34) adjusted for confounders	Odds ratio for 1-9 cigarettes per day
Hirvonen et al. 2017	Prospective birth cohort	Finnish infants born alive between 1991 and 2008 who were alive at one year of age and had no major congenital anomalies	1,018,256 infants (3814 cases)	Smoking in pregnancy recorded in the Finnish medical birth registry	ICD-9 (codes 317-319) or ICD-10 (code F70-F79) in the hospital discharge register or reimbursement register of the social insurance institution by the age of 7 years old		1.40 (0.84-2.31) <32 weeks 0.22 (0.03-1.64) 32-33 weeks 1.94 (1.31-2.86) 34-36 weeks 1.31 (1.13-1.47) 37-41 weeks 1.03 (0.60-1.78) ≥42 weeks	Hazard ratio for each gestational length category (<32 weeks, 32-33 weeks, 34-36 weeks, 37-41 weeks, ≥42 weeks)

1.3.1 – Maternal smoking in pregnancy and offspring risk of ID

A limited number of studies have investigated the association between maternal smoking during pregnancy and offspring ID (see Table 1-1 for a summary). In this section I will critically appraise each of these studies with particular emphasis on evaluating potential epidemiological biases.

Early studies, including Drews et al. [90] and Roeleveld et al. [91], used a case-control design to assess the association. Drews et al. used the Metropolitan Atlanta Developmental Disabilities Study [92] to identify 221 cases and separately sampled 400 controls randomly from the same population. Roeleveld et al. selected 306 cases and 322 controls from referrals to the Paediatric or Child Neurology departments of the Nijmegen University Hospital (Netherlands) or to nearby rehabilitation centres. Drews et al. found increased odds of ID for those exposed to maternal smoking during pregnancy (unadjusted OR = 1.74; 95% CI = 1.21-2.50; adjusted OR = 1.56; 95% CI = 1.01-2.41) while Roeleveld et al. found no association between exposure and outcome (unadjusted OR = 1.1; 95% CI = 0.8-1.5). Roeleveld et al. did, however, find increased odds of ID among children of fathers who smoked a cigar or pipe during pregnancy (OR = 2.4; 95% CI = 1.2-5.1), potentially suggesting a role of second-hand smoke, though the definition of the unexposed group is unclear in their writeup.

Both studies made use of retrospective collection of exposure information through parental interview. This method of collection leads to a high risk of information bias as a result of case status influencing the error in the exposure measurement (commonly referred to as recall bias). This would occur when parents of affected children think more carefully about the different exposures and behaviours that they believe may have had an influence on disease development than parents of unaffected children. Roeleveld and colleagues argue that by using controls who had congenital disorders with known cause this would reduce information bias as the births and upbringing of case and control children were more likely to be similar than if healthy controls were used. This argument may not hold as, if a known cause has been identified for the congenital anomaly, then the parents may be less likely to thoroughly examine the possibility of other causes. Further, sampling both cases and controls from hospital referrals may lead to the possibility of selection bias, specifically in the form described in Berkson's seminal work [93]. Snoep et al. [94] note that if the exposure (here maternal smoking during pregnancy) does not itself lead to a hospital referral, and it seems unlikely that it would, then the resulting bias will be eliminated. However, they also state that indirect Berkson's bias may occur if the exposure causes the disease used to define controls or if the exposure has a common cause with the diseases used to define controls. Smoking in pregnancy has been associated with pregnancy complications [10] which may themselves lead to congenital anomalies. It is therefore possible that Roeleveld's study suffers from Berkson's bias that makes the cases and controls more

similar, thereby masking any effect. In contrast, Drews et al.'s sampling strategy meant that controls were sampled from the population that gave rise to the cases, thereby preventing selection bias. A final advantage of the study by Drews et al. was that they measured socioeconomic status (SES) by census block group of the home address reported on the birth certificate. This would reduce information bias in this confounding variable.

Two studies investigating the aetiology of ID used overlapping birth cohorts obtained from the South Carolina Medicaid billing health services [95, 96]. The first study [95] took 134,596 births between 1996 and 2002 to investigate the effect of trichomoniasis in pregnancy on ID while the second [96] took 165,311 births between 2000 and 2007 to investigate the effect of maternal diabetes mellitus at any time on ID. Neither study specifically investigated the effect of smoking in pregnancy on ID, however, both had data on prospectively collected tobacco use in pregnancy. The cell counts in the descriptives table of each study can be used to obtain unadjusted odds ratios of 1.00 (95% CI = 0.94-1.07) and 1.14 (95% CI = 1.07-1.22) respectively. Interpretation of these effect estimates should be made with caution given that no effort has been made to adjust for confounding factors. The differing effect sizes in the two very similar cohorts may be the result of slightly different definitions of ID (see Table 1-1).

The combined evidence of all four of the above studies was compiled in a meta-analysis [97] which suggested a small increase in the risk of ID with exposure to maternal smoking during pregnancy (OR = 1.10; 95% CI = 1.08-1.23). The bias in each of the above studies is not addressed by combining the statistical associations in a meta-analysis. The conclusion of this meta-analysis therefore must also be taken with caution and further evidence is required. Better evidence has been obtained from more recent studies that suggested an influence of confounding in the association.

Braun et al. [98] performed a large US-based case-control study of 8 year olds where controls were sampled from the population under surveillance for ID and maternal smoking status during pregnancy was ascertained prospectively. In this study an elevated risk of ID (defined as an IQ<70 on recent psychometric testing; unadjusted OR = 1.52; 95% CI = 1.27-1.83) was strongly attenuated after adjustment for the confounders maternal age, race, marital status, education, parity and child sex (adjusted OR = 1.12; 95% CI = 0.92-1.36). An elevated risk for males whose mothers smoked more than 20 cigarettes a day was found following adjustment for confounders (OR = 1.77; 95% CI = 1.20- 2.62); no other adjusted associations were suggested. This may suggest that offspring sex modifies any effect of smoking in pregnancy on risk of ID. However, using sensitivity analyses the authors suggest that this association may be the result of unmeasured confounding. In a sub-cohort of children from North Carolina, Braun et al. compared the odds of ID in children exposed to 20+ cigarettes per day during pregnancy to the odds in children of non-smokers. Here they showed an adjusted association between smoking

20+ cigarettes per day and offspring risk of ID (OR = 1.75; 95% CI = 1.12-2.74). Further adjustment for linked area level income at the time of birth lead to attenuation of the association (OR = 1.57; 95% CI = 1.00-2.46). The authors use this finding to suggest that the associations found for high dosages may still be subject to confounding bias that has not been adequately controlled for.

Another study attempted to make use of the genetics and environment shared between siblings to account for unmeasured confounding. In a cohort of over 160,000 male Swedish conscripts, Lundberg et al. [99] assessed the association between prospectively reported maternal smoking in pregnancy and Swedish military conscript intelligence test score (mean=5, SD=2). A score of ≤2 was defined as poor intellectual performance. The authors note that individuals with such a score would struggle to cope with basic educational programmes, which seems comparable to the definition of ID stated in this chapter, though it should be noted that the prevalence of poor intellectual performance was much higher (~10%). Given that those already diagnosed with ID may be less likely to be invited for conscription the score of ≤2 may not represent the whole spectrum of ID. Increased odds of poor intellectual performance were found for moderate (1-9 cigarettes per day) and heavy (10+ cigarettes per day) smoking during pregnancy before and after adjustment for maternal age, height and BMI, maternal and paternal education and socioeconomic category, family situation, birth order and age at conscription. However, in an analysis of just under 13,000 full sibling pairs from the same cohort, Lundberg et al. compared the odds ratios of combinations of smoking across sibling pregnancies to not smoking in either pregnancy separately for the older and younger sibling. The analysis is an example of a negative control outcome comparison [8, 100, 101] where you would expect to find a stronger effect for smoking in the first pregnancy only for first born children and a stronger effect for smoking in the second pregnancy only for second born children if a causal effect exists. The results showed increased odds of poor intellectual performance for both siblings if the mother smoked only during the first pregnancy and a null association for both siblings if the mother smoked only during the second pregnancy. This finding was therefore not consistent with a causal effect and instead may suggest the association is due to familial confounding. Lundberg et al. also compared within-family and between-family estimates [102, 103] (described in further detail in Chapter 2) of the association between maternal smoking in pregnancy and offspring intellectual performance in the sibling sub-cohort. In this analysis they found that mothers who smoked during a larger proportion of their pregnancies had children with lower average intellectual performance after adjustment for confounders (i.e. a negative between-family estimate). They found that individuals who were exposed to maternal smoking during pregnancy were not likely to have different intellectual performance than their unexposed siblings (i.e. a null within-family estimate). This finding supports the results of the first sibling analysis, that the association between maternal smoking during pregnancy and offspring ID reflects a relationship between

factors related to smoking in pregnancy and offspring ID, but not necessarily a causal effect of smoking on ID itself.

Most recently, Hirvonen et al. [104] used a sample of over one million individuals from Finnish national registry data to assess the risk of offspring ID associated with various gestational lengths. They provided stratified estimates of the association between smoking in pregnancy and ID at each gestational length investigated. Associations between smoking in pregnancy and intellectual disability were found in term and late preterm pregnancies but not very/moderate preterm pregnancies. This may suggest that late preterm and term gestation are more sensitive periods, potentially due to time specific windows of fetal brain development [105, 106]. Smoking could therefore have an impact depending on the timing of the exposure with a greater impact later in pregnancy than earlier on. No association was found for individuals born post term however which is inconsistent with this interpretation, as these children would also have been exposed during the potential critical window. Issues exist with the analyses and presentation of results in this study. Notably, there was no control for socioeconomic status which is likely to be an important confounding factor and may have biased effect estimates. Further the results suffer from "the Table 2 Fallacy" [107] in which presenting all covariate coefficients adjusted for each other leads to difficult and often mistaken interpretation of effect estimates. Here the effects of smoking are adjusted for factors such as birth weight and pregnancy complications which are potential mediators of the association between smoking in pregnancy and offspring ID. Such an adjustment strategy can lead to underestimation of the size of a causal effect [108].

1.3.2 – Evidence of confounding factors

All observational research is susceptible to confounding bias which can lead to misleading conclusions. Confounder bias describes where there exists a common cause of exposure and outcome (discussed in detail in Chapter 2). The prior literature has suggested that the association between maternal smoking and offspring ID is likely to be at least partly attributable to confounding. Below, I describe the evidence for the presence of common causes (also referred to as confounders) of maternal smoking during pregnancy and offspring neurodevelopmental outcomes generally.

Several aspects of smoking behaviours have been suggested to be socially patterned. In their review of the social determinants of smoking, Hiscock et al. [109] highlight that smoking is more prevalent among those with lower socioeconomic position in most studies of developed countries using measures such as education, income, manual occupation, neighbourhood deprivation, single parenting and car and home ownership. Similar rates of quit attempts have

been found between social classes but those of higher SES are more likely to be successful in their quit attempts. A review of the literature on smoking and cessation in pregnancy [110] suggests that these socioeconomic patterns are relevant to the prenatal period.

Abnormal neurodevelopment is also socially patterned; associations have been found between childhood poverty, brain structure and academic performance [111] and between parental education and IQ with offspring IQ at age 5 [112]. Lower early life socioeconomic position has been suggested to reduce performance on measures of executive function, such as working memory and planning, taken early in life and persisting into middle childhood [113]. Socioeconomic status is therefore a likely common cause of both smoking in pregnancy and offspring neurodevelopment.

The socioeconomic confounding structure may be changing, however. The prevalence of smoking in pregnancy has been decreasing and the socioeconomic patterning has been getting stronger over time. Sellers et al. [114] compared smoking behaviour in pregnancy and its socioeconomic correlates between The National Child Development Study (NCDS), a UK based birth cohort beginning in 1958 and The Millennium Cohort Study (MCS), another UK based birth cohort beginning in the year 2000. The comparison showed that, in the UK, smoking in pregnancy has become less common in the 40 years between the cohorts but also that those who smoke are more likely to experience socioeconomic adversity as measured by home ownership, marital status, maternal education, manual occupation and age at birth. Given that identification and diagnoses of conditions such as ADHD, ASD and pervasive developmental disorders (PDD) have been increasing over time as well [115-117], this may make year of birth an important confounding factor.

Sellers et al. [114] exploited these changes over time to suggest an influence of confounding on the association between maternal smoking in pregnancy and offspring neurodevelopment and cognition. Using low birth weight (<2500g) as a positive control outcome (described further in Chapter 2), they compared the strength of associations in the NCDS with the MCS for the effects of smoking in pregnancy on offspring low birth weight and neurodevelopmental and cognitive outcomes including symptoms of conduct disorder and hyperactivity and scores for maths and reading. They found consistent effects of maternal smoking in pregnancy on low birth weight across the two cohorts born approximately 40 years apart but found greater effects of smoking on the cognitive measures (specifically conduct disorder and hyperactivity symptoms and reading scores) in the later cohort. This comparison suggests that the association between maternal smoking in pregnancy and offspring low birth weight is not influenced by changes in the determinants of smoking while the association between maternal smoking in pregnancy and offspring neurodevelopment is. The association with neurodevelopment is

therefore likely to be at least partly explained by socioeconomic confounders which have had increasing influence over time.

Factors other than socioeconomic status may also play a role as confounders of the association between maternal smoking during pregnancy and offspring cognition. Parental psychiatric disorders, immigration status, age at birth and parity are all associated with both smoking behaviours during pregnancy and offspring neurodevelopment.

Smoking behaviour has been found to be correlated with psychiatric disorders generally, both in the population as a whole [118] and among pregnant women [119]. It is unclear as to which of these factors would cause the other, or whether they both have a common cause. This is important for determining whether psychiatric disorders lie on a confounding or mediatory pathway of the association between maternal smoking and offspring neurodevelopment. There is some evidence from longitudinal studies that one's own smoking is associated with worse mental health [120], however if a causal effect of own smoking on mental health does exist, it is likely to happen over a prolonged period of time, before the pregnancy period. Conversely, psychiatric disorders may causally influence rates of smoking specific to the pregnancy period as those with a mental illness have been found to have lower rates of smoking cessation than those without [121] and therefore may find it harder to quit smoking when they get pregnant. It therefore seems sensible to treat maternal psychiatric disorders as a confounder and not a mediator.

Maternal psychiatric history may also be associated with offspring neurodevelopment as a result of shared genetics between parent and child [122] or as a result of an adverse early environment for the child who's parent suffers from mental illness [123]. If those who have psychiatric disorders are more likely to smoke during pregnancy and are also more likely to have children with a psychiatric disorder or ID then this would bias effect estimates to show an increased risk of ID even if no causal effect exists.

A growing body of literature has started to show that children of immigrant mothers have a greater risk of being diagnosed with several neurodevelopmental conditions compared to the offspring on non-immigrant mothers [124]. In two European based studies, smoking rates of pregnant immigrant mothers tended to differ from that of the native population [125, 126]. The socioeconomic patterning of smoking also differed between immigrant and non-immigrant mothers. Maternal migration status may therefore bias associations between maternal smoking during pregnancy and offspring neurodevelopment in a direction that is dependent on whether a population sample contains migrants who smoke more or less than the non-migrant population. For example, in a cohort in which migrant mothers were less likely to smoke but more likely to have offspring with a neurodevelopmental disorder, the association would be biased so that smoking would appear protective if migrant status was not accounted for (under the scenario

that there was no causal relationship between smoking during pregnancy and offspring neurodevelopmental disorders).

Younger mothers have been shown to be more likely to smoke, to smoke more often and more heavily across a number of cohorts [110]. Very young mothers (those under the age of 20) are much less likely to quit smoking than mothers of other ages. The association between parental age at birth and offspring neurodevelopment is complex with younger maternal age and older paternal age associated with increased risk of several offspring neurodevelopmental disorders including ID and behavioural disorders [127, 128]. Maternal and paternal age are strongly correlated, however, meaning that it is hard to conclude with confidence that the disorders are related specifically to maternal or paternal age or to a large difference between the two. Further, maternal and paternal age at birth may indicate socioeconomic status as younger parenthood tends to reflect lower socioeconomic backgrounds. Parental age may therefore proxy for the influence of socioeconomic effects while also having an influence itself on associations between smoking in pregnancy and offspring neurodevelopment.

Pregnant women who are smokers are less likely to quit smoking with each subsequent pregnancy [110]. Increasing parity also has a long-established association with lower intelligence scores that is independent of family size and is consistent across socioeconomic position [129, 130]. The association between parity and intelligence may not be causal and instead could reflect the influence of social rank within a family, as evidenced by a study in which no difference was found in the IQs of first born children as compared to second born children whose elder siblings had died. The role of parity as a confounding factor is therefore unclear. Complex relationships exist between the availability of resources and parity, however. As the number of children in a family increases, the available familial resources need to be distributed among a greater number of family members. Conversely, greater parity is correlated with greater parental age which is in turn associated with increasing wages [131]. Parity may act as a proxy variable for resource availability during pregnancy and therefore capture some of the confounding influence of socioeconomic position.

Finally, it is difficult to separate the possible effect of smoking in pregnancy from the effect of alcohol consumption during pregnancy. Alcohol and tobacco consumption are commonly comorbid in pregnancy [132]. Given that heavy alcohol use in pregnancy can lead to fetal alcohol syndrome, [133] which has several neurodevelopmental consequences including intellectual disability [134], an association between smoking during pregnancy with neurodevelopmental outcomes may simply reflect an effect of alcohol consumption. Alcohol use in pregnancy should therefore be accounted for when modelling the influence of smoking.

In summary, the association between maternal smoking during pregnancy and offspring neurodevelopment is likely to be confounded by several factors, including parental

socioeconomic status, age at birth, alcohol use in pregnancy, parity, psychiatric history, immigration status and year of birth. Any attempts to assess the causal effect of smoking in pregnancy on offspring neurodevelopment will need to account for these variables. As demonstrated by the sibling studies in the previous section, unmeasured cofounding may still be occurring in the association and therefore this list of confounders may well be incomplete.

1.4 – Investigation aims and thesis outline

This thesis sets out to test the hypothesis that maternal smoking during pregnancy is associated with increased risk of offspring ID. I aim to provide evidence of the causal nature of the association through triangulation of different causal inference methods. Part 1, consisting of Chapters 2, 3 and 4, sets out and develops methodological considerations of the investigation. In Chapter 2 I describe the different methods that were used in this investigation and, using directed acyclic graphs (DAGs), explain the biases that each method is likely to suffer from. In Chapter 3 I use DAGs supported by simulations to explore how assortative mating between parents may bias the negative control design, one of the causal inference methods implemented. In Chapter 4 I explore how the proportion of missing data affects bias and efficiency of estimates obtained from multiple imputation analyses.

Part 2 of this thesis, comprising of Chapters 5, 6 and 7, describes empirical investigations of the association between maternal smoking during pregnancy and offspring ID. In Chapter 5 sibling comparisons are used to hold fixed shared familial factors in a cohort derived from Danish registry data. Chapter 6 also employs sibling comparisons, this time using Swedish registry data, to examine the influence of smoking and snus (a moist smokeless form of tobacco) use in pregnancy on offspring risk of ID. The comparison between smoking and snus allows for insight to be gained into the mechanism by which smoking may act upon offspring development. Another comparison between associations with ID versus with fetal growth restriction is used as a positive control analysis to test the ability of the methods used to detect a causal effect where one is expected. In Chapter 7, data from the Avon Longitudinal Study of Parents and Children is used to present the results of observational and causal inference methods that include the negative control design and Mendelian Randomisation.

I conclude this thesis in Chapter 8, where I collate the evidence across the different studies, highlighting what this evidence shows in terms of the causal nature of the association between maternal smoking in pregnancy and offspring intellectual disability. I identify the gaps that remain in the literature following this work and attempt to suggest what further work could be realistically performed to address these.

PART 1: METHODOLOGICAL CONSIDERATIONS

Chapter 2 Establishing causation through triangulation of evidence: an overview of methodological approaches

Although the best way to establish causality is to use an experimental design, the ethics and feasibility of conducting a randomised control trial that investigates the influence of maternal smoking during pregnancy on offspring intellectual disability would be prohibitive. A trial that randomised pregnant women to a smoking arm when smoking is a known teratogen would not be morally acceptable. Randomising pregnant women to a smoking cessation intervention would be more palatable, but the follow up of a sufficient sample size over a period extending well into the offspring's adolescence would be prohibitively expensive. Observational analyses using data from established prospective cohorts are the most appropriate alternative. However, such analyses are susceptible to biases from several sources that can lead to erroneous conclusions. Triangulation of evidence has therefore been recommended [135].

Triangulation in aetiological epidemiology describes "the practice of strengthening causal inferences by integrating results from several different approaches, where each approach has different (and assumed to be largely unrelated) key sources of potential bias" [135]. Each methodological approach implemented to investigate a research question relies on assumptions which may lead to specific biases. By carefully selecting approaches that have complementary strengths and limitations, our confidence in results and conclusions can be improved. Consistent findings across approaches provides stronger evidence to suggest that an association is causal and not the result of any single approach-specific bias. In contrast, inconsistent findings can provide insight as to the broader relationship between variables in a network and inform future research as to what the biases are and how they may be accounted for.

To investigate the association between maternal smoking in pregnancy and offspring intellectual disability I will use traditional observational analyses and causal inference techniques including the negative control design, Mendelian randomisation, sibling comparison designs and cross-context comparisons. The assumptions and biases of each approach need to first be identified to assess whether the weaknesses of each study are addressed by the strengths of the others. It is then possible to identify which weaknesses remain and thereby whether the conclusions formed from the triangulation approach are still susceptible to biases. In this chapter I describe how

causality can be established from observational data and provide a description of each causal inference approach used, the key assumptions made, the key sources of bias and, where possible, the expected direction of that bias.

2.1 – Establishing causality from an observational study

Before the causal inference methods can be assessed I will set out how causality can be established from observational data and describe the different forms of bias that need to be addressed. This description has been guided by the unfinished (at the time of writing) work of Hernán and Robins [136] and framed in the context of investigating my specific research question of whether maternal smoking during pregnancy (the exposure, for this section a binary variable denoted by X) causally influences the offspring's risk of developing intellectual disability (ID; the outcome denoted by the binary variable Y).

2.1.1 – Causal effects

To determine with certainty if a causal effect exists we would require counterfactual information [136], recording for each individual their counterfactual outcome under the scenario in which they were exposed to smoking during pregnancy ($Y^{x=1}$) and comparing this to their counterfactual outcome under the scenario in which they were not exposed ($Y^{x=0}$). An effect can be obtained for each individual, i, by checking whether their outcome under the exposed scenario is not equal to their outcome under the unexposed scenario (i.e. $Y_i^{x=1} \neq Y_i^{x=0}$). A population-averaged causal effect could also be obtained across all individuals, for example as a ratio of the probability of developing ID in the scenario in which all individuals were exposed, relative to the probability of developing ID in the scenario in which no individuals were exposed,

$$\frac{Pr(Y^{x=1}=1)}{Pr(Y^{x=0}=1)},$$

also known as a risk ratio. An odds ratio could also be calculated as,

$$\frac{Pr(Y^{x=1}=1)/Pr(Y^{x=1}=0)}{Pr(Y^{x=0}=1)/Pr(Y^{x=0}=0)}.$$

If either of these measures was not equal to 1, this would indicate a causal effect for the population. An assessment of counterfactuals must obviously be left to the realm of science

fiction given that we cannot know an individual's outcome under both the exposed and unexposed scenarios.

More feasibly a clinical trial could be performed in which pregnant women were randomised either to smoke or to not smoke during pregnancy and their offspring followed up to determine if they developed intellectual disability. Here we would only have information on the observed, randomly assigned exposure value and the corresponding observed outcome. As a result, individual-level causal effects cannot be calculated, and the population-averaged effect measures now no longer assess the causal effect and instead provide a measure of association based on conditional probabilities. The risk ratio now becomes the ratio of the probability of developing ID, only among those who were exposed to smoking during pregnancy, relative to the probability of developing ID among those who were not exposed to smoking during pregnancy,

$$\frac{Pr(Y = 1|X = 1)}{Pr(Y = 1|X = 0)}.$$

Hernán and Robins describe explicitly the difference between a measure of causation versus association. They state that causation is defined as "a different risk in the same population under two different treatment values" whereas association is defined as "a different risk in two disjoint subsets of the population determined by the individuals' actual treatment" [136]. Based on these definitions it can be seen why the risk ratio for the counterfactual data is causal whereas the risk ratio for the randomised control trial is associational.

2.1.1.1 – *Exchangeability*

Some associational effect measures, including the risk ratio, are equivalent to causal effect measures provided a set of assumptions hold. The first of these assumptions is exchangeability. Exchangeability states that the counterfactual outcome is independent of the observed exposure ($Y^x \perp$ X for all x). The exchangeability assumption means that the risk of developing ID in the exposed group is the same as the risk of developing ID in the unexposed group if the unexposed group had in fact been exposed. In other words, the exposed and unexposed groups must be comparable in terms of their risk of developing ID that is not attributable to the exposure itself. Randomisation of women to smoke or not to smoke during pregnancy would ensure that the exchangeability assumption held and enable measures of association to reflect causal effects (provided several further conditions that are not discussed in detail here are met, including that randomisation was followed, the sample was sufficiently large, adequate concealment of allocation and blinding has been used and that there is no differential loss to follow up).

In some trials the randomisation procedure can be influenced by factors that also influence the outcome of interest. Hernán and Robins give a hypothetical example of patients who are in critical condition being randomised to receive a heart transplant with a greater probability than patients who are not in critical condition. The outcome of interest in their example, death, is also influenced by whether the patient is in critical condition. Exchangeability does not hold in this conditionally randomised trial as the untreated group, had they been treated, would not have the same risk of the outcome as the treated group because the treated group are more likely to be in critical condition. Exchangeability may hold, however, within levels of a factor, L (i.e. $Y^x \perp X \mid L$). This is known as conditional exchangeability. In the example provided by Hernán and Robins, exchangeability holds among those who are in critical condition due to the randomisation procedure; the same is true among those who are not in critical condition. By conditioning on this factor, we can ensure that measures of association, such as the risk ratio shown below, are still equivalent to measures of causal effect within subpopulations:

$$\frac{Pr(Y = 1|X = 1,\ L = l)}{Pr(Y = 1|X = 0,\ L = l)}.$$

In the context of my research question I may hypothetically decide, perhaps due to amoral tendencies, to randomise socioeconomically disadvantaged mothers to smoke with greater probability than mothers who are not disadvantaged. As socioeconomic disadvantage is related to the risk of the offspring developing ID, unconditional exchangeability does not hold. However, within levels of disadvantage (i.e. $l = 1$ for the disadvantaged group and $l = 0$ for the non-disadvantaged group) those who are treated will have the same risk of ID as those who were not treated, under the scenario they had been treated. Conditional exchangeability, therefore, holds within each level of socioeconomic disadvantage. An estimate of the population-averaged causal effect (i.e. the unconditional or marginal risk ratio) can be obtained by taking a weighted average of the causal effect estimates in the two subpopulations.

2.1.1.2 – *Positivity*

The next condition required for estimation of a causal effect from a randomised control trial is positivity. The positivity condition states that there is a probability greater than zero for each individual of experiencing each level of the exposure group. For a conditionally randomised experiment the probability of being assigned to each exposure group must be greater than zero within each stratum defined by the variable conditioned upon ($\Pr(X = x|L = l) > 0$). Further, positivity is required for all variables that need to be conditioned upon for exchangeability to hold.

2.1.1.3 – *Consistency*

Next is the consistency assumption which states that the observed outcome for each exposed individual is equal to their counterfactual outcome under the scenario that they had been exposed (i.e. $Y^x = Y$ for every individual with $X = x$). The opposite for the unexposed group is also true. This assumption ties the theoretical counterfactual outcome to the realisable observed outcome and relies on a precise definition of the counterfactual outcome which is achieved by creating a detailed specification of the levels x of exposure X. This is also known as the "no-multiple-versions-of-treatment" component of the Stable Unit Treatment Value Assumption (SUTVA) framework of causality developed by Rubin [137].

2.1.1.4 – *No interference*

The final condition states that the counterfactual outcome of an individual is not influenced by the treatment status of others [138]. This condition is known as the "no interference" assumption and is the second component of the SUTVA framework. It is required as the causal effect of treatment on outcome is not well defined in the presence of interference.

2.1.1.5 – *Non-collapsibility*

Even when all conditions have been met, a causal effect cannot be estimated using all measures of association. The odds ratio is susceptible to an issue known as noncollapsibility that can prevent a causal effect being estimated. Noncollapsibility occurs when an unconditional population-averaged measure cannot be estimated as the weighted average of the conditional measures. For risk ratios the unconditional population-averaged effect will always lie between the values of the conditional measures and is therefore collapsible. In contrast, the unconditional population-averaged odds ratio may lie outside the bounds of the values of the conditional measures (i.e. the unconditional effect may be larger than the largest value of the conditional effect or may be smaller than the smallest value of the conditional effect). This property means that, even when conditional exchangeability holds, the population-averaged odds ratio cannot easily be estimated as a combination of the conditional odds ratios and therefore a causal effect estimate cannot be obtained (see the example by Greenland et al. in Table 1 [139]). Provided that the outcome is rare, however, it can be shown that the odds ratio approximates the risk ratio (proof given by Greenland in [140]), hence the influence of noncollapsibility is also reduced under this scenario. It has been highlighted that the outcome must be rare within all levels of L and not simply rare in the population as a whole [136].

2.1.2 – Relating observational analyses to a randomised control trial

As highlighted in the opening of this chapter, a randomised control trial would not be ethical to conduct. Provided the conditions of exchangeability, positivity, consistency and a lack of interference hold, then analysis of observational data can conceptually be thought of as being equivalent to a randomised control trial and a causal effect measure can be identified. The definitions of these conditions are the same for observational studies as they are for randomised control trials, however, the feasibility of these conditions in an observational study are more difficult due to the lack of experimental control. Exchangeability in observational studies, and its relationship to types of systematic bias is explored in detail below in Section 2.2

In a randomised control trial positivity occurs by design, in part by making those individuals for whom positivity would be untenable, ineligible for inclusion. Positivity is harder to ensure in an observational study. In an observational setting, the conditioning set L required for conditional exchangeability, is likely to contain several factors. This can lead to many strata, some of which may contain a set of individuals with a zero probability of experiencing the exposure. This can be checked empirically, however, by testing whether there are exposed and unexposed individuals at every level of L (i.e. across all combinations of adjusted variables).

In an observational setting the exposure is not applied uniformly to all those who experience it. As a result, the definition of the levels of exposure as a binary measure may not be precise enough for consistency to hold. The current epidemiological literature presented in Chapter 1 suggested that timing of exposure and the quantity of cigarettes smoked may both influence the risk of developing ID. It may be possible that the offspring of a mother who smoked early in pregnancy but then quit would not develop ID, while that same offspring would develop ID if their mother smoked throughout pregnancy. The consistency condition would not hold here as in both cases the mother smoked during pregnancy, but the observed outcome is not the same (Y^x does not always equal the same value of Y despite $x = 1$ in both scenarios). A similar argument could be made regarding the quantity of cigarettes smoked during pregnancy. Factors such as timing and dosage of exposure therefore need to be accounted for when defining levels of binary exposure variables for consistency to hold and causal effects to be estimated.

I have now set out a general framework by which measures, including the risk ratio and odds ratio, under very specific conditions, can be used to estimate a causal effect of an exposure on an outcome in a study of observational data.

2.2 – Bias in observational studies

In this section I will describe the scenarios under which bias is likely to occur in observational studies. Bias under the null is of particular importance here. It describes the estimation of an association between the exposure and outcome even when no causal effect exists. Bias of any kind can be detrimental as it can lead to the over, or underestimation of the true influence of an exposure in the risk of disease which can have important consequences for policy decisions. The biases to be described are known as confounding bias, selection bias, information bias and missing data bias. As will be seen, the concepts of confounding and selection bias are closely related to the concept of exchangeability which has already been covered.

2.2.1 – A brief introduction to directed acyclic graphs

To aid in this description I will make use of directed acyclic graphs (DAG) [141]. The explanation of how DAGs function will make use of the following:

DAG (i) - $A \rightarrow B \rightarrow C$

DAG (ii) - $A \leftarrow B \rightarrow C$

DAG (iii) - $A \rightarrow B \leftarrow C$

DAG (iv) - A $\rightarrow \boxed{B} \rightarrow C$

DAG (v) - A $\leftarrow \boxed{B} \rightarrow C$

DAG (vi) - A $\rightarrow \boxed{B} \leftarrow C$

For those unfamiliar with DAGs, they are directed, meaning that all edges between nodes indicate the direction of the causal relationship between those variables, and acyclic, meaning you cannot follow the path extending from a node back to that node (no variable can cause itself). The absence of an edge between nodes means that it is assumed that there is an absence of a causal relationship between those nodes. Known common causes of any two variables must be included within the graph. DAGs can be used to establish whether any two variables that do not have a direct connection between them are associated (also referred to as d-connected) or are independent (d-separated). Three rules guide d-connection and d-separation [141]:

Rule 1. Unconditional separation – A and C are d-connected if there is an unblocked path between them.

Here a path is any set of consecutively connected nodes, irrespective of the direction of the edges connecting those nodes. The path is unblocked if there are no collider nodes (i.e. a node with two arrows both directed towards it). For example, A and C are d-connected in the DAGs (i) and (ii), but are d-separated in DAG (iii) because B is a collider node/variable.

Rule 2. Blocking by conditioning – A and C are d-connected, conditioned on a set of nodes Z if there is a collider-free path between A and C that traverses no member of Z. If no such path exists, then A and C are d-separated or "blocked" by Z.

DAGs use a square around a variable to indicate that the variable has been conditioned upon. Using Rule 2 it can be seen that A and C are d-separated in DAGs (iv) and (v) as the path is blocked by the set Z which consists of a single variable B.

Rule 3. Conditioning on a collider – If a collider is a member of the conditioning set Z, or has a descendant in Z, then it no longer blocks any path that traces this collider.

Rule 3 means that A and C are d-connected in DAG (vi) as B is a collider variable that has been conditioned upon. Similarly, if there was a variable D that was caused by B, and we conditioned upon D, then A and C would also be d-connected in this situation.

2.2.2 – Counterfactuals and directed acyclic graphs

The concepts of exchangeability and DAGs were developed under two separate frameworks of causal inference: the counterfactual and causal-graph frameworks respectively. The two frameworks do not overlap perfectly, however an attempt to identify and justify commonality between them has been attempted by Flanders and Eldridge [142]. Their work highlighted that the following claims hold:

Claim 1. The presence of exchangeability and faithfulness result in the absence of a biasing path.

Here faithfulness is used to define whether the relationships between variables in a DAG (i.e. d-connectedness and d-separation) are reflected in the statistical dependencies between those variables. A biasing path is defined as a path between exposure and outcome that is unblocked by controlled collider variables or uncontrolled non-collider variables. This path must be undirected, in that the arrows on the path are in both directions (such as is the case in DAG (ii) described in Section 2.2.1). This is in contrast to a directed path from exposure to outcome, in which all arrows flow in the same direction (such as DAG (i)). Claim 1 means that, provided a DAG truly reflects the data observed in the real world, if exchangeability holds, we should not be able to find a biasing path from exposure to outcome.

Claim 2. Absence of a biasing path implies exchangeability.

Claim 2 is, in essence, the reverse of Claim 1, though Flanders and Eldridge provide separate mathematical proofs for both claims and the proof for claim 2 does not require some of the subtler details of faithfulness, related to setting specific values for parameters, that I have not attempted to describe here. Claim 2 therefore implies that we can use DAGs to infer exchangeability.

Claim 3. Absence of a biasing path and presence of consistency implies no bias.

The final claim presented here (several more are presented by Flanders and Eldridge) states that the combination of the absence of a biasing path, and therefore exchangeability, and presence of consistency will lead to no bias of the exposure-outcome association measure, which therefore reflects the causal effect measure. This claim seems incomplete however based on our understanding of required identifiability assumptions laid out in Section 2.1 . It stands to reason that positivity and lack of interference should also hold within the DAG for causal effects to be identified.

2.2.3 – Confounding bias

Confounding bias occurs when there is a set of determinants, L, of the exposure that also influence the risk of the outcome. This results in the counterfactual risk of developing the outcome not being independent of the actual exposure: L influences the probability of the exposure and the outcome meaning that those who were exposed do not have the same risk of developing the outcome as those who were not exposed in the counterfactual scenario that they were exposed. Conditioning on the full set L would ensure conditional exchangeability (i.e. $Y^{x} \perp X \mid L$), however, in observational data not all confounding variables will have been identified or measured. The full set L contains the subsets C, measured confounding variables, and U, unmeasured confounding variables. A DAG of this scenario is presented in Figure 2-1. In this DAG both variables, C and U are determinants of the exposure and outcome (there is an arrow from the confounding variable to the exposure and to the outcome) and therefore create biasing pathways along $X \leftarrow C \rightarrow Y$ and $X \leftarrow U \rightarrow Y$. The biasing pathway $X \leftarrow C \rightarrow Y$ can be blocked by conditioning on C (i.e. $X \leftarrow \boxed{C} \rightarrow Y$), however the path $X \leftarrow U \rightarrow Y$ cannot be blocked as there is no recorded information on U. This can prevent our use of conditional exchangeability in obtaining an unbiased estimate of the causal effect (i.e. our associational measure does not reflect the causal effect measure).

The size of the bias in the associational effect measure from the causal effect measure is a function of the size and direction of the associations of the confounder with each of the exposure and the outcome. Inaccurate measurement of confounding variables can limit the amount of confounding bias that is accounted for by adjustment.

Confounding bias need not strictly be the result of a single determinant of the exposure and outcome. For example a third variable A may be related to C such that a biasing path could be created along either $X \leftarrow C \leftarrow A \rightarrow Y$ or alternatively $X \leftarrow A \rightarrow C \rightarrow Y$. Both pathways provide examples of confounding bias. If A is observed then controlling for either C or A would block the biasing pathway and provide conditional exchangeability. Where A is not observed then only C can be conditioned upon.

In Chapter 1 I have described how the literature suggests that the association between maternal smoking in pregnancy and offspring ID is likely to be confounded. Further evidence suggests that previous studies have been unable to control for all confounding and that residual confounding remains, possibly due to potential confounders not having been identified. Other methods than simple conditioning (also known as adjustment) must therefore be used to account for this confounding.

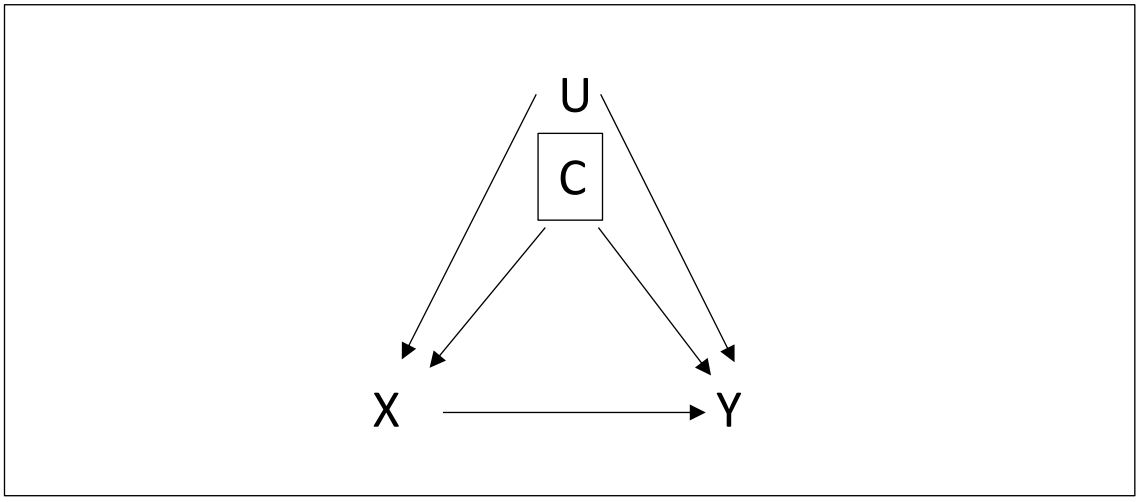

Figure 2-1: DAG of the relationship between exposure (X) and outcome (Y), both of which are influenced by confounders that are measured (C) or unmeasured (U). The box around C, means that we condition, or adjust, for this variable in models..

2.2.4 – Selection bias

Selection bias occurs when conditioning on a variable, or set of variables, S, leads to a lack of exchangeability. It can be seen in Figure 2-2a that X and Y are both determinants of S. If S is not conditioned on, then it blocks the path $X \rightarrow S \leftarrow Y$, as it is a collider variable (see Rule 2 of DAGs in Section 2.2.1) and only the path $X \rightarrow Y$ remains unblocked. Conditioning on S opens a biasing pathway along $X \rightarrow \boxed{S} \leftarrow Y$. Here unconditional exchangeability holds but conditional exchangeability does not. Where S represents selection into a study it is conditioned on by

default as we are looking at the effect of exposure on outcome only among those who are in our study population. More complicated examples can also occur such as that shown in Figure 2-2b where selection is influenced by both X and a second variable A that influences the outcome, thereby leading to the open biasing pathway $X \rightarrow \boxed{S} \leftarrow A \rightarrow Y$. Conditioning on A would close this biasing pathway, however, if we are unaware of A's relationship with selection into a study or haven't measured A, then we are unable to do so.

When creating a cohort of individuals to study, ensuring that the selected cohort is representative of the population of interest aids in reducing selection bias. In DAG terms, a representative sample of the population will have as few variables as possible that point to S, selection, and therefore the chance of a biasing pathway existing as a result of conditioning on S is reduced. In a counterfactual framework, if the sample is not representative of the total population, then even if exchangeability held for the total population, it would likely not among the selected sample.

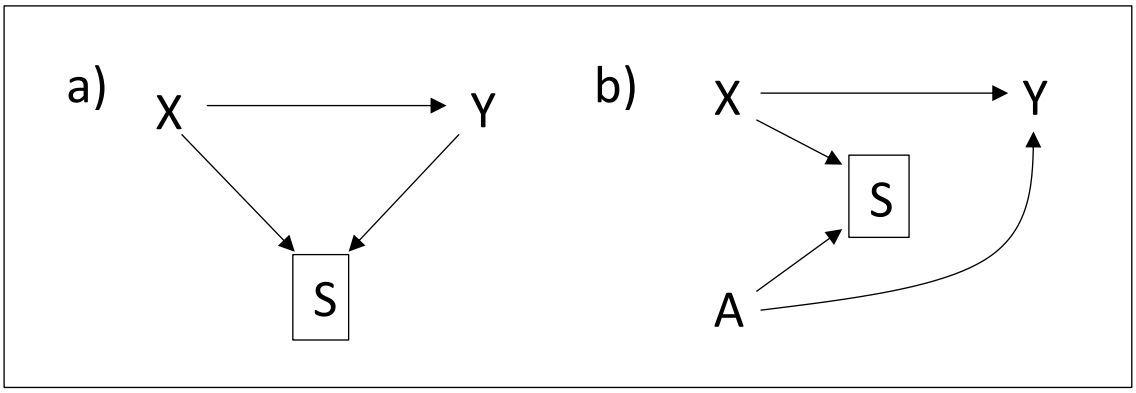

Figure 2-2: DAGs of the relationship between exposure (X) and outcome (Y), conditioned upon the variable S which is either a common effect of the exposure and outcome (as in part a) or is a common effect of the exposure/outcome and an ancestor variable for the outcome/exposure (as in part b).

2.2.5 – Information bias

Variables are frequently measured with error that can lead to bias in effect estimates under specific scenarios. Error terms are often ignored in DAGs but can be represented easily based on the works of Hernán and Cole [143] and Shahar [144]. In Figure 2-3a the values X^* and Y^* are used to represent the value of the exposure, X, and outcome, Y, that are observed with error, e_X and e_Y. Analysis is based on assessing the association between the observable X^* and Y^*, which represents the underlying association of the true values of the exposure and outcome, X and Y. In this example the error structure is independent as the error terms do not have a common cause and non-differential as the error terms are not influenced by other variables in the

network. This error structure can lead to information bias, but will not lead to bias under the null.

Information bias under the null will occur when the error terms are either dependent, or the error structure is differential. Where error terms are produced by the same mechanism (they have a common cause on a DAG) they will be dependent (see Figure 2-3b). Such would be the case if the same person reported information on the exposure and the outcome. As a result there is an open biasing pathway from X^* to Y^* along $X^* \leftarrow e_X \leftarrow e_{XY} \rightarrow e_Y \rightarrow Y^*$, where e_{XY} is the common cause of e_X and e_Y. This type of information bias can be reduced by using different sources for the ascertainment of exposure and outcome.

Differential error can occur in the exposure or the outcome (see Figure 2-3c for an example of independent differential exposure error and Figure 2-3d for an example of independent differential outcome error). Differential exposure error, where the probability of exposure misclassification is influenced by the outcome, will lead to an open biasing pathway from X^* to Y^* along $X^* \leftarrow e_X \leftarrow Y \rightarrow Y^*$. Even if no association between exposure and outcome exists (i.e. no arrow from X to Y), an association between X^* and Y^* may still be estimated due to bias from this pathway. If $X \rightarrow Y$ does exist, then the point estimate of the association between X^* and Y^* may be biased away from the value of the point estimate for the association between X and Y due to the same biasing path. Differential outcome error occurs when the probability of outcome misclassification is influenced by the exposure. The corresponding biasing pathway is opened along the path $Y^* \leftarrow e_Y \leftarrow X \rightarrow X^*$.

Differential exposure error is unlikely in the context of my research question provided that I use prospectively collected information on maternal smoking during pregnancy. Failure to do so could lead to mothers of children who have developed ID systematically reporting their smoking during pregnancy with greater (or perhaps lesser) accuracy than the mothers who do not have a child who developed ID. This is commonly referred to as recall bias. Similarly, differential outcome error seems unlikely as the temporal distance between exposure to smoking and diagnosis of ID mean that smoking in pregnancy is unlikely to directly influence measurement error in the outcome.

More complex scenarios could exist where an external factor creates an association between exposure/outcome and error. Indirect paths from Y to e_X or from X to e_Y could be created if determinants or descendants of the exposure/outcome are also determinants of measurement error in the outcome/exposure. Several examples can be thought of such as (i) a genetic propensity for intellectual disability in the mother could influence the accuracy of her self-report of smoking during pregnancy, (ii) lower socioeconomic status, which predicts increased smoking, may influence access to professional assessments of intellectual disability which could

in turn lead to misclassification of the outcome and (iii) potential consequences of exposure to smoking such as reduced lung function may increase contact with medical services which may increase the likelihood of an accurate assessment of intellectual disability. As with confounder bias, if the cause of the error is unknown or unmeasured then we will be unable to adjust for it in analyses, which would block the biasing pathway. Methods beyond simple adjustment strategies will need to be used to account for information bias from differential error structures that involve indirect pathways.

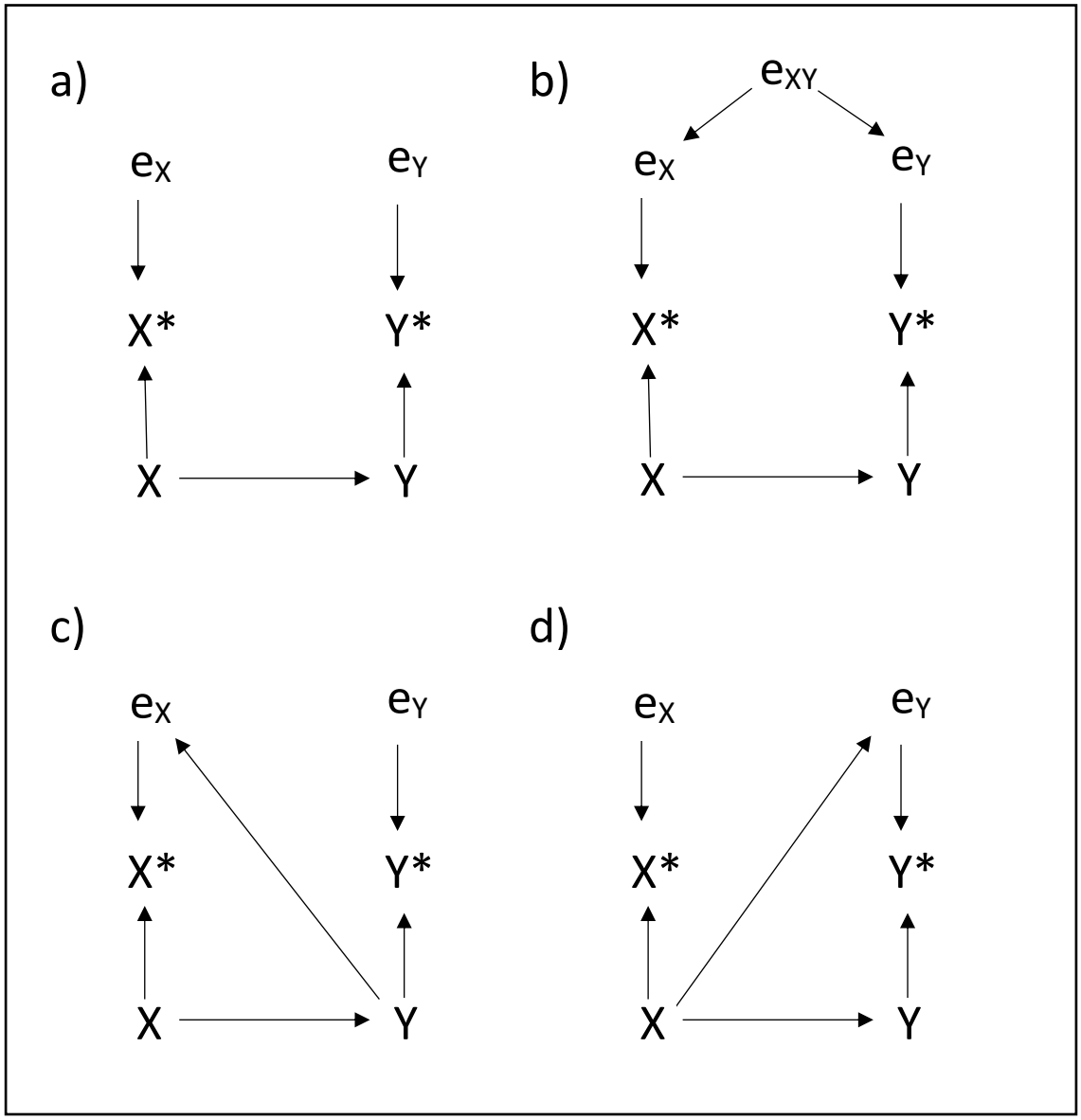

Figure 2-3:DAGs of the relationship between exposure X and outcome Y, measured with error e_X and e_Y respectively to give the observed values X and Y*. The relationship between each of these variables provide error structures that are a) independent non-differential, b) dependent non-differential or c) and d) independent differential (see the main text for a description of these terms. Dependent differential error (not shown) is also possible.*

2.2.6 – Missing data bias

Missing data can lead to selection bias under specific circumstances as a result of conditioning upon whether the data has been observed. Bias arising in this way is almost inevitable in epidemiological studies and is therefore worth specific mention. Using Rubin's terminology [145], reasons for missing data (also referred to as missingness) are classified as: missing completely at random (MCAR) where the probability of missingness does not depend on either observed or missing data, missing at random (MAR) where conditional on the observed data the probability of missingness is independent of unobserved data and missing not at random (MNAR) where the probability of missingness is dependent on unobserved data even after conditioning on observed data.

It is commonly believed that analysis using only data with non-missing values for all variables (complete case analysis; CCA) is unbiased only in the case where the missing data mechanism [146] is MCAR. This has been shown not to be the case; CCA can be unbiased in instances where data are MAR or MNAR [147-149]. Recent work by Hughes et al. [150] showed that for linear regression the exposure coefficient is biased when the probability of missingness in any variable depends on the outcome variable, and for logistic regression, when the probability of missingness in any variable is dependent on the outcome and the exposure.

Where CCA is biased, other methods such as multiple imputation (MI) [151-154] or inverse probability weighting [155] can be used. MI requires that the data be MCAR or MAR, though has the benefit that the probability of missingness can be explained by auxiliary data that is not included in the analysis model [156]. This is useful in cases where the MAR assumption is fulfilled by mediator variables for the exposure outcome association, inclusion of which in the analysis model would remove the portion of the effect of exposure on outcome that acts via the mediator. It is often impossible to verify that the data are MAR, however, and expert knowledge of the patterns of missing data is required. My recent work (reproduced in Chapter 4) displayed that MI analyses in which the data are not MAR show increasing amounts of bias with increasing proportions of missing data [157]. While we argued that the proportion of missing data should not be used to guide decisions on MI, we also concluded that sufficient auxiliary information must be available to fulfil the MAR assumption and avoid bias using this method.

Previous work in the Avon Longitudinal Study of Parents and Children (ALSPAC) has shown that increased probability of drop out from the cohort study is predicted by low socioeconomic status, maternal smoking in pregnancy and by having a neurodevelopmental disorder [158]. Missing data are therefore dependent on both the exposure and outcome and CCA is likely to be biased. If missing data in the outcome variable is more likely among those who have the outcome, for example if having ID made attendance at study clinics prohibitive, then the data

are MNAR and MI will also be biased. Methods other than MI are likely to be needed to address the biases that arise from missing data.

2.3 – Assessment of causal inference methods

I have described in Section 2.2 how different forms of bias arise in epidemiological studies. Some, but not all, of these biases can be addressed by careful choice of which variables to condition upon (adjust for in regression models) and when and how each variable is measured. The presence of variables in the causal network that are unmeasured or are unknown causes of the exposure, outcome, or error of the measurement of either/both can lead to bias that cannot be addressed by conditioning on measured variables. In this section I will describe methods that attempt to address such biases, including the negative control design, Mendelian randomisations, sibling comparison designs and cross context comparisons. Other methods for causal inference such as marginal structural models and propensity scores could be used (described here [159]) but were not implemented as part of this thesis.

2.3.1 – Negative control design

In biological research the negative control design is implemented to test whether factors other than the treatment of interest may have led to a causal interpretation of experimental results (see the work by Lipsitch et al. [100] for examples). The design compares the magnitude of an estimate of a treatment-outcome association against the estimate of another association (the negative control association) in which either the treatment, or the outcome, has been replaced with a variable such that the new association is not plausibly causal via the hypothesized mechanism.

The negative control design has been adapted for use in epidemiological research [8, 100, 101, 160]. In this setting the negative control design concedes that bias in the estimate of the causal effect is inevitable, however, it exploits the fact that different associations may be biased to the same extent. Provided the association of interest (AOI) and the negative control association (NCA) share similar biasing pathways the two estimates can be compared. The bias away from the causal effect estimates of both associations should be influenced to the same extent by these biasing pathways. If the effect estimate of the AOI is substantially more extreme than that of the NCA then this provides evidence in favour of the association being causal in nature. Interpretation of the association estimates of the AOI and NCA themselves is therefore not the primary purpose of this design, only the comparison of the two. It is left to the researcher to

subjectively interpret whether the size of the difference in estimates is clinically meaningful. Bootstrapping can be used to create a confidence interval to allow for statistical testing of this difference.

Negative control designs can be used to assess whether prenatal exposures are causally related to outcomes via an in-utero pathway. Here the association of maternal exposure with an outcome (the AOI) is compared to the association of the paternal exposure with the same outcome (the NCA). Early applications of the design assessed the association of maternal smoking in pregnancy on offspring low birthweight (see the commentary by Keyes et al. [161] for a brief history) while more recent examples using ALSPAC are provided by Taylor et al. [162] and Richmond et al. [163]. These studies respectively assessed whether maternal smoking in pregnancy is associated with offspring depression and whether maternal body mass index (BMI) is associated with methylation of the offspring HIF3A gene. The paternal exposure may have an in-utero effect, such as would be the case in the first example where passive smoking or changes in sperm quality as a result of smoking may influence offspring depression. It has been argued by Davey Smith, however, that associations arising by such different mechanisms are unlikely to be equal in magnitude to that of the association with the maternal exposure [8].

2.3.1.1 – *Assumptions of the negative control design*

A DAG of the negative control design assessing the causal nature of a prenatal exposure is presented in Figure 2-4. The DAG is split into two parts to reflect biasing pathways from confounding bias and from information bias. It should be noted the DAG additionally includes bidirectional arrows, ↔, to show that two variables have common causes that for simplicity are not shown in the DAG, but the result of which is an open path between the two variables. In context of my research question, the variables M and P are the true values of maternal and paternal smoking during pregnancy, Y is offspring intellectual disability, and C and U are measured and unmeasured confounding respectively. The variables M^*, P^* and Y^* are the observed values of M, P and Y measured with error e_M, e_P and e_Y respectively. In the design, we are comparing the sum of all paths from M to Y to the sum of all paths from P to Y. For this comparison to be useful in informing whether the association between M and Y is causal, the following assumptions are made:

Asumption 1. The bias from path $M \leftarrow U \rightarrow Y$ should be approximately equivalent to P $\leftarrow U \rightarrow Y$ (see Figure 2-4a).

Asumption 2. The bias from $M^* \leftarrow e_M \leftrightarrow Y \rightarrow Y^*$ should be approximately equivalent to $P^* \leftarrow e_P \leftrightarrow Y \rightarrow Y^*$ (see Figure 2-4b).

Asumption 3. The bias from $Y^* \leftarrow e_Y \leftrightarrow M$ should be approximately equivalent to $Y^* \leftarrow e_Y \leftrightarrow P$ (see Figure 2-4b).

Assumption 1, describing the equivalence of biasing paths arising from unmeasured confounding, may not be reasonable in the context of smoking during pregnancy. Societal norms mean that a mother who smokes in pregnancy may be more likely to suffer judgement from her peers than if the father smokes during the same pregnancy. It is therefore possible that the smoking behaviour of mothers are on average more likely to change than fathers in response to becoming pregnant. Though the confounding structures of maternal and paternal smoking may usually be similar, they may not be so during the period of pregnancy.

Assumptions 2 and 3 describe biasing pathways arising from information bias. In Section 2.2.5 I described how information bias via direct routes can be minimised, and provided these approaches are taken for both the maternal and paternal associations assumption 2 and 3 should hold. In the same section I also described three scenarios by which information bias could arise via indirect routes. Each of these scenarios would likely affect both the maternal and paternal associations equally and hence assumption 2 and 3 would likely hold.

2.3.1.2 – *Further considerations*

Measurement error in the exposure or negative exposure has been shown to bias the estimate of the difference in maternal and paternal association sizes even in the absence of a biasing pathway. Sanderson et al. [164] used simulations to show that bias in the estimate of the difference will arise if there is greater error in the measurement of the exposure than the negative exposure (or vice versa). If there is no effect of the exposure or negative exposure on the outcome, then the estimate of the difference will be unbiased if there is equivalent error in both measures. However, when there is an effect of the exposure on the outcome, but no effect of the negative exposure, equivalent rates of error in both measures will lead to an underestimate of the difference in association sizes. This finding is important in guiding the choice of measure for maternal and paternal smoking during pregnancy. Paternal smoking during pregnancy is often recorded either via self-report or via maternal report. Where smoking is reported by a second individual, as in the case of maternal report of paternal smoking, this may increase the rate of error in paternal smoking measures while error in maternal smoking remains unchanged. The work by Sanderson et al. shows that this may bias the estimate of the difference in association sizes. I am therefore encouraged to make use of paternal self-report, though it is important to note that increased perceived stigma of maternal than paternal smoking during pregnancy may lead to greater error in maternal report of maternal smoking than the error in paternal report of paternal smoking.

The design requires two further conditions for the comparison of maternal to paternal smoking in pregnancy to be valid. First, the exposure and negative exposure should be measured on the same scale and second, the timing of the exposure and the negative exposure must be equivalent so that the comparison is relevant.

In summary the negative control design makes no attempt to control for biases in the maternal and paternal associations that are the result of unknown or unmeasured variables. Instead the design assumes the two associations to be equivalently biased and looks to assess the difference in size between them. Assortative mating has been highlighted as a potential issue for the negative control design and a method for resolving this has been proposed [101]. This issue and its proposed solution will be explored in depth in Chapter 3 using DAGs and a simulation study.

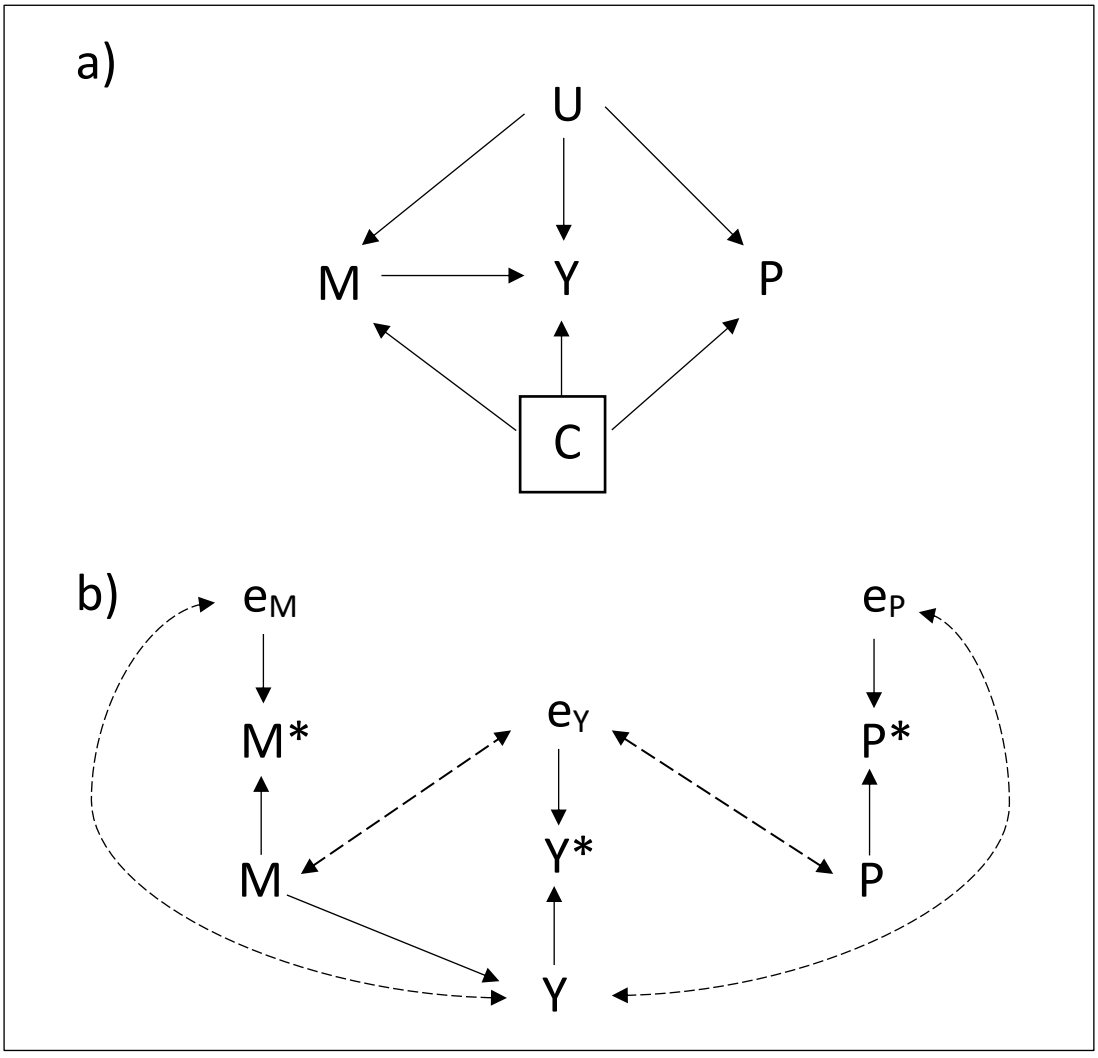

Figure 2-4: DAGs of the negative control design which compares the association between maternal exposure (M) and outcome (Y) to paternal exposure (P) and outcome. The associations can be influenced by confounding from measured (C) and unmeasured confounders (as in part a) and by error in the exposure and outcome (as in part b).

2.3.2 – Mendelian randomisation

Where an exposure-outcome association is biased by confounding factors, variables that strongly predict the exposure but are not associated with the confounding factors can be used as proxies for the exposure to mitigate the influence of the confounding and allow a causal effect to be estimated. Such an analysis is known as an instrumental variable analysis and the variables used to proxy the exposure are known as instruments. Mendelian randomisation (MR) [165, 166] describes a form of instrumental variable analysis in which the instruments are genetic variants, otherwise known as single nucleotide polymorphisms (SNPs). Mendel's laws of segregation and independent assortment mean that MR can be thought of as emulating the randomisation process in clinical trials.

2.3.2.1 – *MR assumptions*

MR relies on three key assumptions (see Figure 2-5 for a visual guide):

Asumption 1. The instrument, Z, is strongly associated with the exposure, X

Asumption 2. The instrument does not have a common cause with the outcome (i.e. is not associated with confounder variables, U)

Asumption 3. The instrument influences the outcome, Y, only via its influence on the exposure

Provided these assumptions hold, a test of the null hypothesis, that there is no causal effect of exposure on outcome, can be performed. Verification of this set of assumptions is not straightforward, however.

Fulfilment of Assumption 1 has largely relied on the strength of association of SNPs with the exposure obtained from genome wide association studies (GWAS) where a genome wide significance threshold of $p < 10^{-8}$ has been employed as a cut off. Smoking has been suggested to be strongly influenced by genetic variation [167] and a recent GWAS of smoking behaviours in 1.2 million individuals (the GSCAN study [168]) has revealed a number of variants associated with smoking initiation and number of cigarettes smoked per day at the required genome-wide significance threshold. Relying on the strength of association in GWAS studies is naïve however, as recent work by Morris et al. has identified that these associations may be biased as a result of population stratification, dynastic effects and assortative mating [169]. Population stratification, describing systematic differences in allele frequencies between subgroups of a population, is frequently accounted for in GWAS by restricting to a homogenous population (for example restricting only to members of one race) and adjusting for the first 10-20 principal components of genotype, which capture differences in allele frequency between subpopulations. It has been highlighted that genetic associations still reflect geographic

structure, and therefore potentially population stratification, even after the first 100 principal components have been adjusted for in models [170]. Dynastic effects, where offspring phenotype is influenced by inherited genetics as well as directly by the realised phenotype of the parents, and assortative mating, where parents select each other based upon phenotype, can also lead to biased estimates of genotype-phenotype associations. Morris et al. suggest that GWAS should use family-based designs instead of focusing on samples of unrelated individuals in order to account for these effects. Until such GWAS have been performed we must use GWAS data with potentially biased associations, keeping in mind the effects this bias may have on MR analyses.

The biasing effect of population stratification can lead to violation of Assumption 2, the independence assumption. When the subgroups of the population with differing allele frequency are defined by a determinant of the outcome, the relationship between the instrumental variable and the outcome is confounded. Assessment of whether the independence assumption is plausible can be performed by investigating the relationship between the genetic variants and the determinants of the outcome [171].

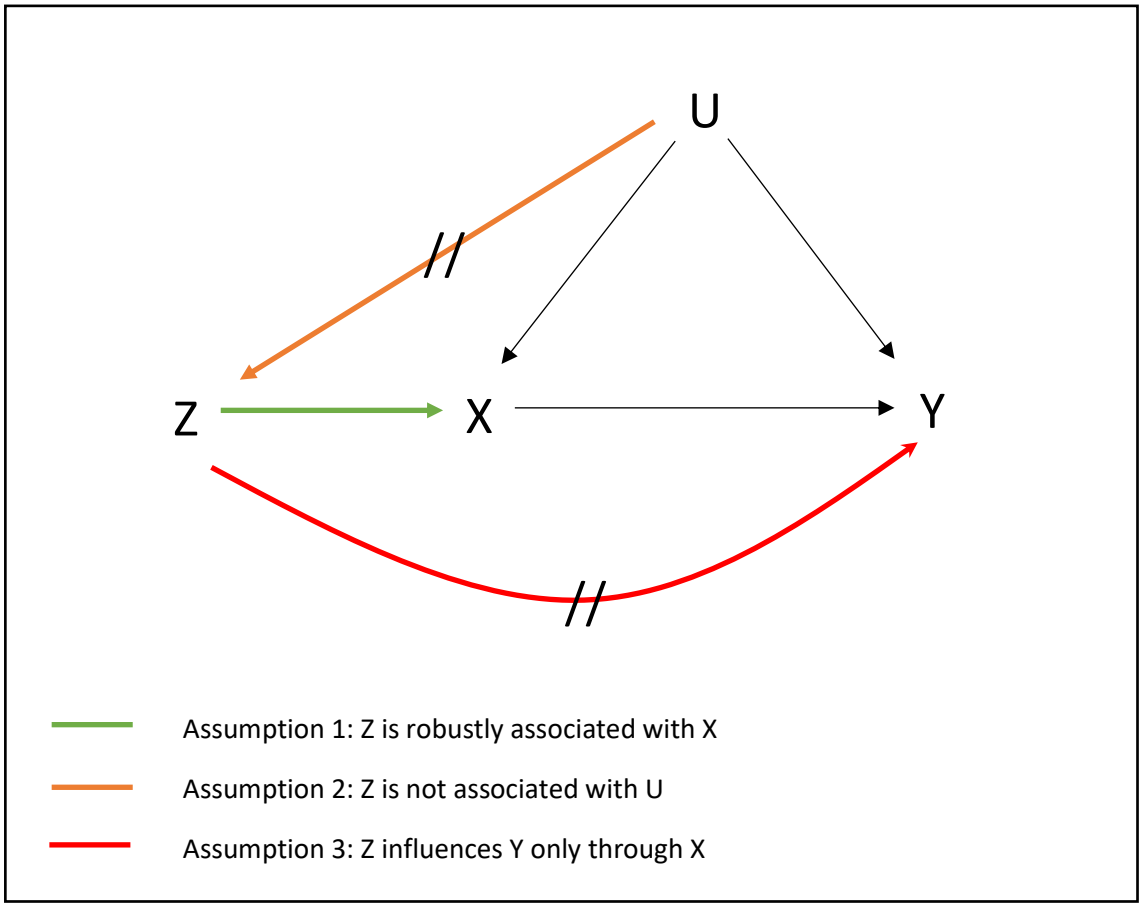

Figure 2-5: DAG of the assumptions made in a Mendelian randomisation analysis in which a genetic variant (Z) is used as a proxy for the exposure (X). Z is not caused by confounding factors or causes (U) of the outcome (Y) and only influences the outcome via the exposure. Here // has been used to indicate that the path between variables should not exist.

Assumption 3, the exclusion restriction criteria, can be violated as a result of horizontal pleiotropy [171], where genetic variants result in multiple processes that can cause the outcome, at least one of which does not act via the exposure of interest. For example genetic variants for educational attainment also predict smoking behaviour [172] and therefore may show up as hits in a GWAS of smoking behaviours. These maternal SNPs for educational attainment could influence the risk of offspring ID independently of the mother's smoking behaviour thereby violating Assumption 3.

2.3.2.2 – *MR implementation*

MR can be implemented using summary statistics from two independent GWAS, one for the exposure and one for the outcome, in what is known as two-sample MR [173, 174]. Here the causal effect of exposure on outcome can be obtained for each SNP contained in both GWAS by using a ratio of the beta coefficient for the exposure relative to the beta coefficient for the outcome. Where both exposure and outcome are binary, this is a ratio of the mean change in log-odds of the outcome with each allele for a given SNP, relative to the mean change in log-odds of the exposure with each allele for the SNP. An average effect across all available SNPs can then be obtained using the inverse variance weighted (IVW) method [173].

Sensitivity analyses, such as fitting MR Egger [175, 176], weighted median [177, 178] and weighted mode [179] estimators can be used to explore whether the IVW estimate is biased by violations to the MR assumptions. Each of these estimators makes different modifications to the MR assumptions and can therefore be used collectively to provide information about potential bias [135].

MR-Egger relaxes Assumption 3, allowing and correcting for the presence of horizontal pleiotropy (i.e. the genetic variant influences the outcome via a path other than through the exposure). Using multiple genetic variants as instruments, MR-Egger fits a weighted linear regression model of the SNP-outcome coefficients, obtained from GWAS, on the SNP-exposure coefficients with an intercept term. If the intercept were constrained to zero, this model would produce a coefficient for the SNP-exposure term that is equal to the IVW estimate. MR-Egger does not constrain the intercept to zero. If there is no horizontal pleiotropy then the intercept term would be equal to zero as when the SNP-exposure association is zero there should also be no change in the outcome. A non-zero intercept therefore estimates the average influence of all SNPs on the outcome via pathways other than via the exposure. This means that MR-Egger produces a causal effect estimate of the influence of exposure on outcome (the MR-Egger beta coefficient) adjusted for horizontal pleiotropy. This estimate is consistent even if all SNPs violate the exclusion restriction criteria.

The MR-Egger method assumes that violations of the exclusion restriction criteria are as a result of a direct effect of the SNP on the outcome and not via confounders [135]. This is known as the InSIDE assumption (instrument strength independent of direct effect). The presence of confounding would lead to correlation between the SNP-exposure association and the quantity of the SNP-outcome association not mediated by the exposure which would prevent a causal estimate from being obtained. Further, inclusion of the intercept term comes at the cost of statistical power to detect an effect. This is why MR-Egger is used as a sensitivity analysis and not a primary analysis.

The weighted median estimator takes the 50th percentile of the ratio estimates, weighted by the inverse of the variance [135]. The method assumes that at least 50% of the SNPs are valid instruments (they meet Assumption 1-3) and that no single SNP contributes more than 50% of the weight. The weighted mode estimator takes the most common causal effect estimate as the causal effect of exposure on outcome.

In summary, MR is another powerful method for dealing with confounding by making use of the natural randomisation of genetic variants at conception. Exchangeability holds for the association between the genetic predictors of the exposure and the outcome, provided the set of MR assumptions hold. Verification of all assumptions is not possible, however methods to assess their plausibility have been referred to.

2.3.3 – Sibling comparison designs

Siblings provide the setting for a natural experiment of the influence of a prenatal exposure. They share 50% of their genetics on average and in most cases share a common environment during development. The sibling comparison is therefore useful for causal inference as it provides a method for controlling for unmeasured genetic and environmental confounding that is shared between the siblings without actively adjusting for the confounding variables in models.

2.3.3.1 – *Within-family and between-family effects*

Early work by Begg and Parides [102] and by Carlin and colleagues [103] decomposed population-averaged effects into within-family and between-family effects by taking the average exposure value between siblings (the family-averaged exposure) and then including coefficients for both the individual-level exposure (within-family effect) and family-averaged exposure (between-family effect) in regression models as

$$g(\mathrm{E}[Y_j|X_j, \bar{X}]) = \beta_0 + \beta_W X_j + \beta_B \bar{X},$$

where j denotes the sibling contained within a family, β_W is the within-family effect, β_B is the between-family effect, $g(.)$ is the link function and $E[.]$ is the expectation operator.

The regression models can take the form of generalised estimating equations (GEE) or random effects models in order to account for the correlation within families. Begg and Parides [102] note that the coefficient of the individual-level exposure in a regression model (β_W) will be the same regardless of whether the exposure is centred on the family averaged exposure or not (i.e. β_W will be the same if the term $\beta_W(X_j - \bar{X})$ is used in place of the term $\beta_W X_j$). In contrast the coefficient of the family-averaged exposure (β_B) will change based on these two model formulations. Thinking of the within-family effect as the effect on the outcome of a unit deviation of the individual exposure value from the family-averaged exposure value can be useful in conceptualising how the model works.

If for example we were looking at the effect of birthweight on ID, the within-between model attempts to answer the question, what is the effect of a unit increase in birthweight on ID, given that the mother tends to have offspring of a certain weight on average. Certain factors that influence the risk of ID may also influence the mother's tendency to have larger or smaller children on average, thereby confounding the association between the two. By looking at the deviation of the individual birthweight from the family-averaged birthweight we can assess the influence of birthweight on ID, unconfounded by the factors that influence the family-averaged birthweight.

In context of the current research question, in which we assess the influence of maternal smoking during pregnancy on offspring ID, the family-averaged exposure can be thought of as a maternal propensity to smoke during pregnancy. Here, the within-between model asks the question, what is the influence of being exposed to maternal smoking during pregnancy given the mother's propensity to smoke during pregnancy (which is influenced by shared familial confounding factors, while the individual deviation from this propensity is not). Carlin and colleagues [103] briefly argue that, when the exposure is binary as is the case here, and only 2 siblings are used in each family, the family-averaged exposure can only take the values 0, ½ or 1 and therefore is not meaningful. It may therefore make more sense to use families of greater size to allow the family-averaged exposure to take a greater range of values and therefore more accurately reflect a propensity for smoking during pregnancy.

Sjölander and colleagues later established a more complete mathematical framework for estimating a causal effect from the sibling design, formalising the key component of the within-between model that, by including the between-family effect in a model, the within-family effect becomes adjusted for all shared confounders [180]. In their work they highlighted that for binary exposures it would be more accurate to interpret the within-family effect (in the absence

of non-shared confounders between siblings) as a causal effect within the exposure-discordant subpopulation. This is because only members of the exposure-discordant subpopulation contribute any information to the within-family effect coefficient. Families in which the mother smoked in all pregnancies, or in no pregnancies, will have no variation in the deviation from the family-averaged exposure, and therefore cannot contribute to the causal effect. For ease of terminology, I will continue to refer to the causal effect within the exposure-discordant subpopulation as the "within-family effect", however it should be kept in mind that this is not strictly accurate.

2.3.3.2 – *Considerations of within-between models*

Sjölander's paper further showed, using a series of equations, that within-between models need to adjust for the non-shared confounders of both the index child and the sibling in order to produce an unbiased estimate of the within-family effect [180]. This is because adjustment of the individual-level exposure for the family-averaged exposure, $\bar{X}$, is equivalent to conditioning the index child's exposure on the sibling's exposure. Using DAG notation, and denoting subscript $j = 1$ as the index child, subscript $j = 2$ as the sibling of the index child, U as an unobserved shared confounder and C as an observed non-shared confounder, conditioning on X_2 opens a path from the index child's exposure to the sibling's non-shared confounder along $X_1 \leftarrow U \rightarrow \boxed{X_2} \leftarrow C_2$, and from the index child's outcome to sibling's non-shared confounder along $Y_1 \leftarrow U \rightarrow \boxed{X_2} \leftarrow C_2$ (see Figure 2-6a). The sibling's non-shared confounder, C_2, therefore becomes a confounder of the association between the index child's exposure and outcome, X_1 and Y_1.

Using simulations, Sjölander and colleagues went on to show that adjusting for the non-shared confounder of the index child but not the non-shared confounder of the sibling lead to a constant level of bias. In comparison, adjusting for no non-shared confounders lead to bias that was proportional to the strength of the association between the non-shared confounder and the outcome. This means that where the effect of the non-shared confounder on the outcome is weak, adjusting for the non-shared confounders of the index child only, will lead to more bias than not adjusting for any non-shared confounders. Adjusting for both siblings non-shared confounder variables resulted in no bias, irrespective of the strength of the association between the non-shared confounder and the outcome.

The same research group, this time lead by Frisell, later showed that familial correlation between siblings in the exposure variables and in the confounder variables can influence the bias arising from only conditioning on the non-shared confounder of the index child [181]. They note that in a within-between model there is selection on siblings that are discordant for the

exposure. Because the exposure values for the siblings are different, this implies that the non-shared causal variables of the exposure (i.e. the non-shared confounding variables) are more likely to take different values, therefore leading to an increase in the association between the exposure and the confounder than in the general population.

Using the DAG in Figure 2-6b, in which F_C, F_X and F_Y are familial factors that respectively influence the values of the confounders, exposure and outcome in siblings within a family, Frisell and colleagues state that conditioning on X_1, which occurs by default in the within-between regression model, leads to negative correlation between C_1 and X_2 as a result of the path $C_1 \rightarrow \boxed{X_1} \leftarrow F_X \rightarrow X_2$. If X_2 is equal to 0 then F_X is also likely to take value 0 (i.e. if the second sibling was not exposed then the mother's propensity to smoke is likely to be lower). This means that X_1 (index child) has an increased probability of also taking the value 0 (not exposed), but if in spite of this X_1 takes the value 1 (exposed), as is restricted to in the exposure discordant population, it means that C_1 has an overriding influence on the value of X_1.

In contrast a positive correlation is induced between C_1 and X_2 along the path $C_1 \leftarrow F_C \rightarrow C_2 \rightarrow X_2$. If X_2 is equal to 0, this decreases the probability that C_2, F_C and ultimately C_1 are equal to 1 (and vice versa). Restricting to the exposure discordant sample (where $X_1 \neq X_2$), this decreases the influence of confounding on exposure (i.e. a smaller association between C_1 and X_1).

There are therefore opposing effects of quantity of confounding of the within-family effect estimate that are created by the size of the sibling correlation in exposure values (represented by the strength of association of F_X with X_1 and X_2) and the sibling correlation in non-shared confounders (represented by the strength of association of F_C with C_1 and C_2). Using simulations, Frisell and colleagues displayed that if the confounders are perfectly correlated then there is no bias (as the non-shared confounders would instead be shared confounders). If the correlation in the confounders is greater than the correlation in exposures (strength of influence of F_C is greater than influence of F_X), then bias in within-family effect estimate is less than the bias in the crude effect estimate of a standard regression model. Finally, if the correlation in confounders is less than the correlation in exposures (strength of influence of F_C is less than influence of F_X), then the bias in the within-family effect estimate, adjusting for the index child's non-shared confounders only, will be greater than the bias in a crude effect estimate.

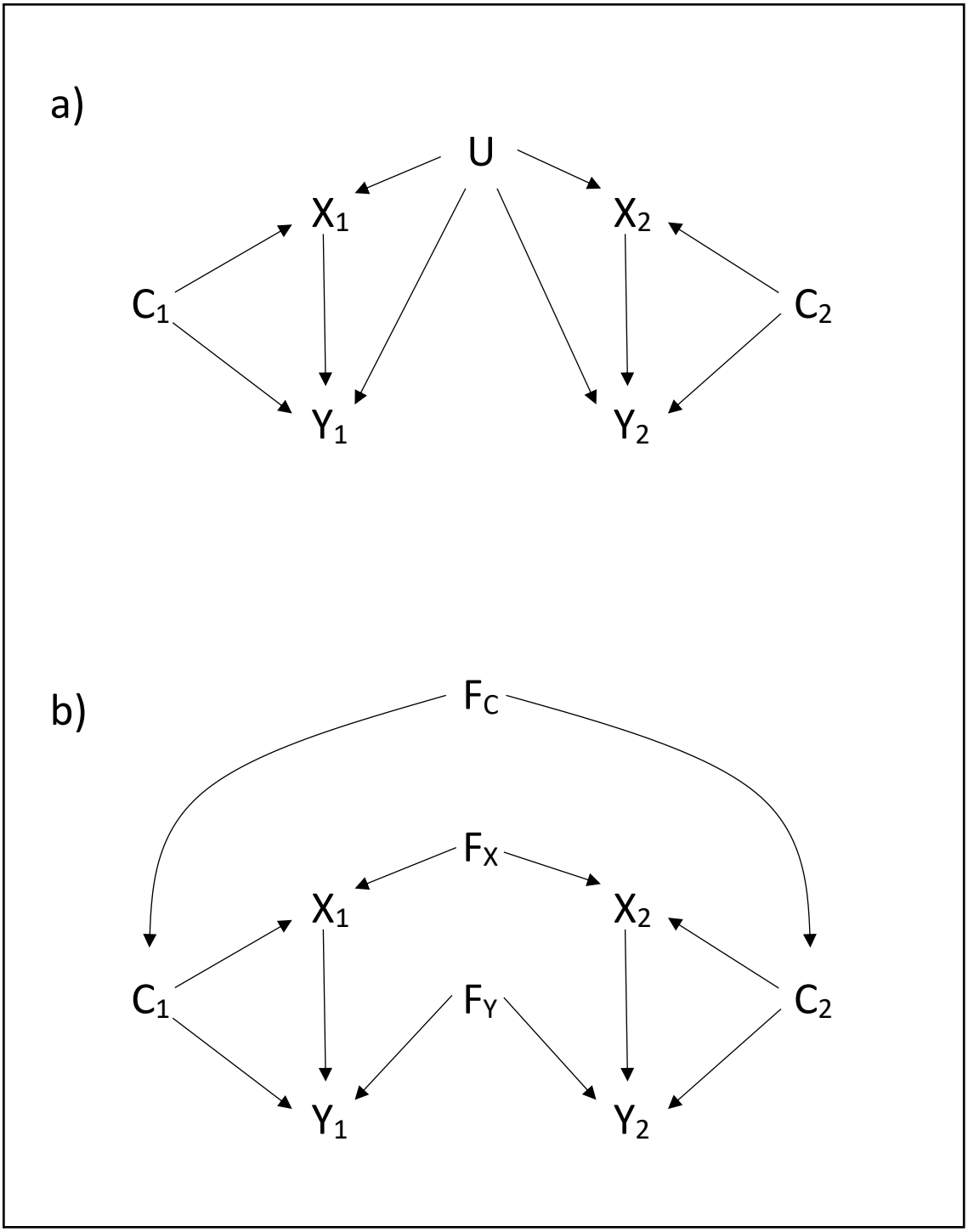

Figure 2-6: DAGs of the relationship between variables in the sibling design. X_j, Y_j and C_j are the exposure, outcome and non-shared confounders of child j respectively. U are shared confounders. F_C, F_X and F_Y are familial factors influencing the non-shared confounders, exposure and outcome respectively of both children in the family.

A final consideration of the sibling design is its potential to be biased by carry-over effects. Carry-over effects occur when the exposure or outcome of one sibling has a causal influence on the exposure or outcome of another sibling [182]. The current research question may be particularly susceptible to carry-over effects. For example, exposure of the first sibling to smoking during pregnancy may influence the second sibling's risk of developing ID if smoking has long term reproductive health consequences for the mother that influence the offspring's risk of ID. Conversely, exposure of the second sibling to smoking during pregnancy may increase the risk of ID in the first sibling if exposure to second-hand smoke early in life increases the risk of ID. A final plausible scenario is that if the first sibling develops ID, and the mother thinks this may be related to her smoking behaviour in pregnancy, she may be less likely to smoke during a second pregnancy. Sjölander and colleagues highlight that several logical tests can be performed to rule out whether carry-over effects are biasing effect estimates, but, in their words, cannot "rule-in" if carry effects are present [182].

In summary sibling designs are a useful tool for addressing bias from unmeasured shared confounders between siblings. These designs are susceptible, however, to increased bias as a result of incorrect conditioning of non-shared confounders in regression models and from carry-over effects that need to be accounted for when designing analyses and interpreting results.

2.3.4 – Cross-context comparison

Cross-context comparisons may be useful for informing about the nature of a null result or conversely investigating the mechanisms by which a causal effect could occur. In this section I will describe how a positive control comparison and comparison with a second method of nicotine delivery will be used in my investigation.

The association between smoking in pregnancy and offspring fetal growth restriction has strong evidence of being causal in nature from complementary causal inference designs [8, 9, 183]. If using the same causal inference methods, an association is found for fetal growth restriction but not ID then this will support the interpretation that observational associations with ID are the result of residual confounding. This type of comparison will be referred to as a positive control analysis.

If a causal effect of smoking in pregnancy on ID does exist, then a cross-context comparison between the associations of snus use and smoking in pregnancy with offspring ID can be used to investigate whether effects are the result of nicotine or the combustible components of cigarette smoke. Snus is a moist, smokeless tobacco that is increasingly being used as a smoking cessation aid in Sweden [184, 185], with some suggestion that it is more successful as an aid to stop smoking than nicotine patches or gum [186, 187]. Snus delivers nicotine in quantities that

are comparable to cigarette smoke though with slower absorption and higher plasma nicotine concentration over an extended period [184, 188].

2.4 – Chapter summary

In this chapter I have established how causal effects can be obtained from counterfactuals, clinical trials and observational studies, highlighting the importance of the underlying assumptions. Violation of these assumptions, particularly the exchangeability assumption, can lead to several forms of bias, including confounding, selection, information and missing data bias, which have been described using DAGs. In observational studies these biases can be harder to prevent due to the lack of experimental control.

Several causal inference methods have also been described which aim to account for bias, with a strong focus on accounting for confounding. The negative control design compares two associations that suffer equivalently from confounding, one of which is not expected to reflect a causal relationship. Mendelian randomisation uses the natural randomisation of genetic variants at conception to mimic the randomisation process used in clinical trial designs, thereby improving the balance of the distribution of confounding factors between the exposed and unexposed groups. Finally, the sibling design holds fixed all confounders that are shared between siblings and therefore provides a way of accounting for a portion of unmeasured confounding bias. Each of these methods suffer from their own potential biases, described in detail in the sections above. These analysis specific biases mean that the conclusions of each analysis cannot be taken in isolation. Triangulation of consistent evidence across each study design is required to establish whether there is a causal effect of maternal smoking in pregnancy on offspring risk of developing ID.

Chapter 3 Assessing the influence of assortative mating in the negative control design

The contents of this chapter have been published in the following peer reviewed journal article:

- *Madley-Dowd P., Rai D., Zammit S., and Heron J., Simulations and directed acyclic graphs explained why assortative mating biases the prenatal negative control design. Journal of clinical epidemiology, 2020. 118: p. 9-17. DOI:* https://doi.org/10.1016/j.jclinepi.2019.10.008

In Chapter 2 I highlighted that the results of the negative control design may be influenced by assortative mating. Positive assortative mating describes the tendency for individuals to mate with a partner who has the same value of a given characteristic as themselves while negative assortative mating describes preference for mates with a differing value on the characteristic to one's own value [189].

Smoking [190-193], alcohol use [190, 191, 194-199], caffeine use [191] and body mass index (BMI) [200-202] have all been suggested to be characteristics correlated within pairs as a result of positive assortative mating; these characteristics are also commonly examined as in-utero exposures in negative control designs. Exposure characteristics may be similar within a pair due to (i) mate selection based on the characteristic itself (e.g. a non-smoking individual may limit their selection of partner to non-smokers as they do not want to be exposed to smoke) or (ii) selection based on determinants of the characteristic (e.g. age, education, and psychiatric and personality traits influence smoking behaviours and may also be selected upon by individuals choosing a partner [201, 203-209]). Evaluation of the evidence of the nature of exposure characteristics being similar between parents is beyond the scope of this chapter, however, I believe this work will show that the impact of scenarios (i) and (ii) are similar in the context of negative control designs.

The most common approach in a negative control design of a prenatal exposure is to run three models (irrespective of additional models adjusting for potential confounders). Model 1 assesses the association between maternal exposure and outcome. Model 2 assesses the association between paternal exposure and outcome. Model 3 mutually adjusts both maternal and paternal exposure for each other. The maternal and paternal effect estimates are then compared against each other between Model 1 and 2 and also within Model 3.

The value of comparing estimates within Model 3 over comparing estimates between Models 1 and 2 is mentioned briefly in the appendices of Lipsitch et al.'s early description of the negative control design's use in epidemiology [100] and in Davey Smith's letter to the editor regarding this paper [101]. Why this is the case has not been adequately demonstrated or discussed in the literature so far. In this chapter I aim to explain the importance of interpreting the difference in effect sizes obtained from the mutually adjusted model (Model 3) where exposure and negative exposure are influenced by assortative mating, using directed acyclic graphs (DAGs) and a simulation study to guide the explanation.

3.1 – Directed acyclic graphs

I motivate the remainder of this chapter using an example that is closely aligned to the research question of the overall thesis. The example compares the influence of maternal smoking during pregnancy (the exposure) to the influence of paternal smoking during pregnancy (the negative exposure) on offspring intelligence quotient (IQ) score (the outcome). Figure 3-1 shows DAGs of the relationship between variables. Here M is maternal smoking during pregnancy, P is paternal smoking during pregnancy and Y is the offspring outcome. C_M and C_P are sets of confounding variables for the maternal and paternal associations with the outcome. Mate selection influenced by the exposure variable is represented by S_{exp} while selection influenced by confounding variables is represented by S_C. As any of several possible mates could have been selected, S_{exp} and S_C can be considered to be random variables. When a couple has a child together then mate selection has occurred, the couple have selected each other, and this variable can be considered as having been controlled upon (represented by the box drawn around the variable). S_{exp} and S_C are collider variables, therefore controlling for them will lead to correlation between maternal and paternal exposure variables and maternal and paternal confounder variables. For simplicity in this DAG I have assumed that paternal smoking during pregnancy is not causally associated with offspring outcome. The DAGs presented here are an adaptation of the initial DAGs of the negative control design created by Lipsitch et al. [100] (presented in Chapter 2 as Figure 2-4a) with the change that confounding shared between M and P is ignored and non-shared confounding between the two is included. Figure 3-1(i) shows a simplified example where the exposure behaviour is selected on. Confounding is ignored in this example. Only one variable, M, directly connects to the outcome. A single backdoor pathway exists that connects P to Y (along P $\rightarrow \boxed{S_{exp}} \leftarrow M \rightarrow Y$). The paternal coefficient of the paternal only model (Model 2) will be biased by the backdoor path. Mutual adjustment for M and P in a single model (as in Model 3) will close this backdoor path and eliminate the bias for the paternal

coefficient. No backdoor paths exist for M. As a result, the maternal coefficient of both the maternal only model (Model 1) and the mutually adjusted model (Model 3) will be unbiased.

Figure 3-1(ii) provides an example in which the correlation between maternal and paternal exposure behaviour is caused by mate selection based on determinants of these exposures, confounders for the association of maternal and paternal smoking with offspring outcome. The association of C_M with M and with Y is assumed to be equivalent to the association of C_P with P and with Y. In this example maternal and paternal smoking during pregnancy share some but not all backdoor paths to the outcome. Three variables directly connect to the outcome: M, C_M and C_P. Backdoor paths along $C_M \rightarrow Y$ and $C_P \rightarrow Y$ exist for both M and P (e.g. $M \leftarrow C_M \rightarrow Y$; $M \leftarrow C_M \rightarrow \boxed{S_C} \leftarrow C_P \rightarrow Y$; $P \leftarrow C_P \rightarrow Y$; $P \leftarrow C_P \rightarrow \boxed{S_C} \leftarrow C_M \rightarrow Y$). An additional backdoor path exists for P that does not exist for M, via confounding variables ($P \leftarrow C_P \rightarrow \boxed{S_C} \leftarrow C_M \rightarrow M \rightarrow Y$). As a result, there will be additional bias for the negative control association (NCA) in Model 2 that will not occur for the association of interest (AOI) in Model 1. By mutually adjusting for M and P the additional backdoor paths for P will be closed. M and P will then have the same backdoor paths again ensuring that the biasing of the AOI and NCA are once again equivalent in Model 3.

Figure 3-1(iii) combines the examples shown in Figure 3-1(i) and Figure 3-1(ii), showing the situation in which correlation in exposure behaviours is due to selective mating based on both the exposure and the confounder variables. The backdoor paths that exist for P but not M now include both $P \leftarrow C_P \rightarrow \boxed{S_C} \leftarrow C_M \rightarrow M \rightarrow Y$ and $P \rightarrow \boxed{S_{exp}} \leftarrow M \rightarrow Y$ which will lead to greater bias for the NCA than the AOI. Mutual adjustment for both M and P will close both of these backdoor paths leading to equivalent bias of the AOI and the NCA thereby making them comparable for the purpose of interpreting whether a causal effect may exist.

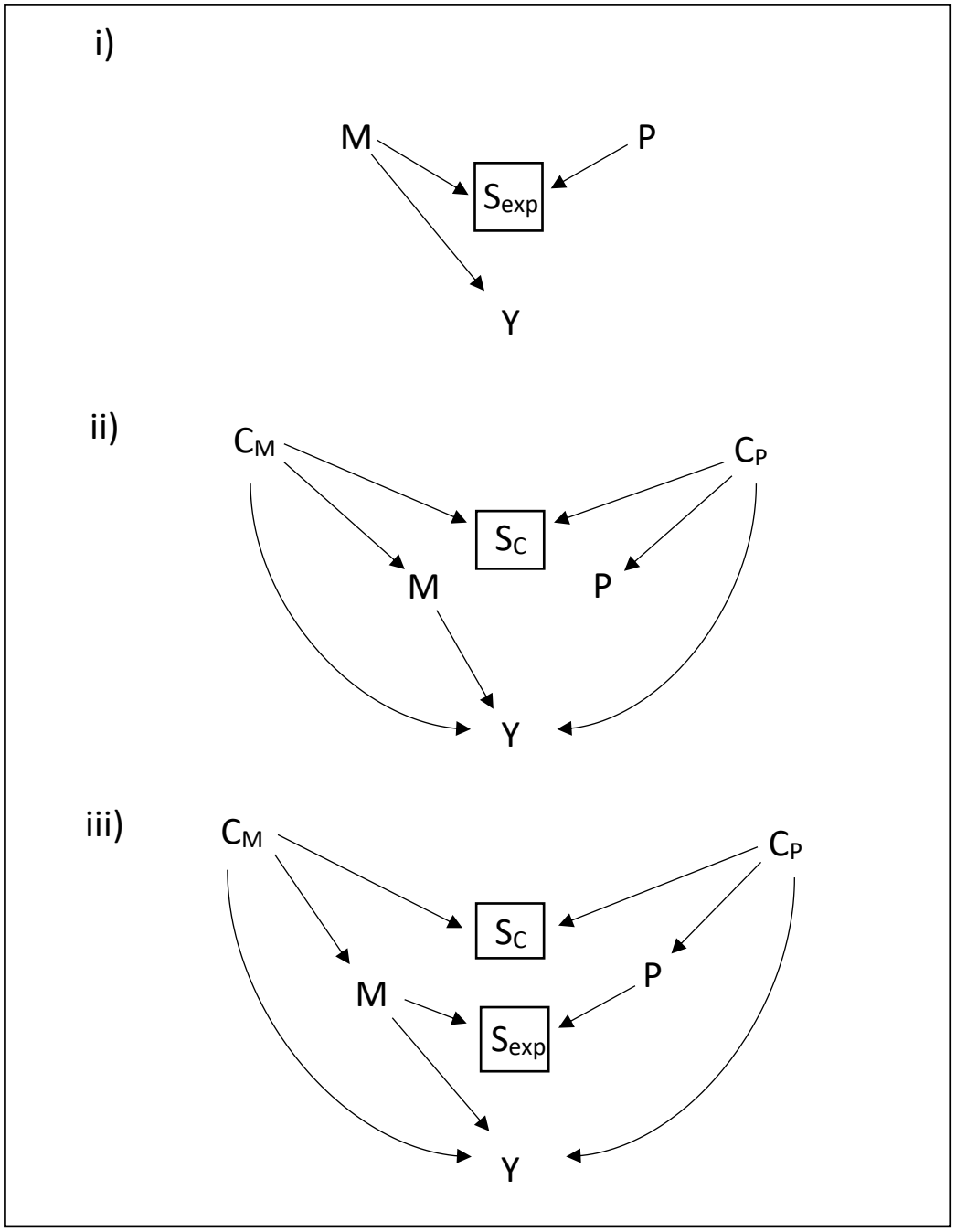

Figure 3-1: Directed acyclic graphs (DAG) of the associations between variables in a negative control design with assortative behaviours. Refer to the text for descriptions of what (i), (ii) and (iii) represent. M is maternal smoking during pregnancy, P is paternal smoking during pregnancy, Y is the offspring outcome. C_M and C_P are maternal and paternal specific confounders respectively. S_C and S_{exp} are variables indicating mate selection based upon confounding variables and upon the exposure variable. S_C and S_{exp} are collider variables that when controlled for (such as when a couple have a child) induce correlation between the maternal and paternal confounders/exposures.

3.2 – Simulation study

3.2.1 – Methods

3.2.1.1 – *Part 1: Simulation study of the influence of assortative mating on conclusions from the negative control design*

I empirically tested how assortative mating can influence the results and conclusions from a negative control design using a simulation study. The study was motivated by the same example as for the DAG in Figure 3-1i such that shared backdoor paths along confounder variables were ignored so that only the $P \rightarrow \boxed{S_{exp}} \leftarrow M \rightarrow Y$ backdoor path exists.

I first simulated a binary exposure (maternal smoking in pregnancy, M), a binary negative exposure (paternal smoking in pregnancy, P) and assortative mating between the two. I simulated each exposure-pair to fall within one of the four categories of maternal and paternal smoking combinations. I fixed the prevalence of maternal smoking during pregnancy at 24% (to mimic the prevalence observed in ALSPAC) and allowed the prevalence of paternal smoking to vary across settings as I varied the extent to which smoking was assortative. Assortative mating was quantified using the pair sexual isolation index (I_{PSI}, see Appendix A.1 for formula) [210, 211], a commonly used measure in evolutionary biology literature that ranges from -1 to 1. Values closer to 0 indicate no assortative mating while values closer to 1 indicate a mating pair are more likely to be similar on the chosen characteristic. I investigated I_{PSI} values between 0 and 0.8, derived from the frequency in each smoking combination category (see Table 3-1). I did not consider negative assortative mating ($I_{PSI}<0$) as most assortative traits that are used as prenatal exposures in negative control designs display positive, not negative, patterns of assortative mating.

Table 3-1: Frequency of observations falling into each category of maternal and paternal smoking, and the quantity of assortative mating, measured using the I_{PSI}.

Frequency in category (%)				I_{PSI} value
1) No parent smokes	2) Mother only smokes	3) Father only smokes	3) Both parents smoke	(quantity of assortative mating)
38.0	12.0	38.0	12.0	0.0
45.6	9.6	30.4	14.4	0.2
53.2	7.2	22.8	16.8	0.4
60.8	4.8	15.2	19.2	0.6
68.4	2.4	7.6	21.6	0.8

I then simulated a continuous outcome, which in the context of the research question was labelled "IQ score". I simulated a normal distribution with mean 0 and variance 1 for the children who were unexposed to maternal smoking in pregnancy and a normal distribution with mean $\mu_{\mathrm{M\,true}}$ and variance 1 for those exposed to maternal smoking in pregnancy. The value of $\mu_{\mathrm{M\,true}}$ was varied between -5 and 5 in increments of 1. There was no effect of paternal smoking for all simulation settings.

Three linear regression models were fitted to the simulated data:

Model 1 - $\mathrm{g}(\mathrm{E}[Y|M]) = \beta_0 + \beta_M M$

Model 2 - $\mathrm{g}(E[Y|P]) = \beta_0 + \beta_P \mathrm{P}$

Model 3 - $g(E[Y|M,P]) = \beta_0 + \beta_M M + \beta_P P$

Model 1, the maternal only model, regressed the outcome on maternal smoking only. Model 2, the paternal only model, regressed the outcome on paternal smoking only. Finally Model 3, the mutually adjusted model, regressed the outcome on both maternal and paternal smoking. I calculated the difference between β_{M} and β_{P}, the coefficients for maternal and paternal smoking, between Model 1 and 2 and again within Model 3. Confidence intervals for these differences were produced using bootstrapping with 1000 replications.

Across 1000 simulations I investigated sample sizes of 100, 1000 and 10,000. I measured the average bias of β_{M} and β_{P} and their Monte-Carlo standard error across simulations using the *simsum* command in Stata [212]. I calculated the average difference between β_{M} and β_{P}, as well as the average lower and upper bound of the confidence interval of the difference, across simulations.

I repeated the simulation study using a binary outcome. The findings did not differ substantially from those for a continuous outcome and are presented in Appendix A.2.

3.2.1.2 – *Part 2: Simulation study of a negative control design with assortative mating where the negative exposure influences the outcome independently of the exposure*

In part 1 of the simulation study I have assumed that the negative exposure has no influence on the outcome. For some exposures the negative exposure may have an independent effect on the outcome. For example, paternal smoking may influence offspring neurodevelopment through a prenatal effect (reduced sperm quality), antenatal effect (exposing the mother to smoke) or a post-natal effect (exposing the offspring to smoke). I therefore investigated how this scenario would influence the estimates of each model in the presence of assortative mating.

I repeated the simulation study, this time including an association between paternal exposure to smoking and the outcome. The outcome for this analysis was generated by simulating normal

distributions (all with variance 1) with mean 0 for children who were unexposed to maternal or paternal smoking in pregnancy, mean $\mu_{\text{m true}}$ for those exposed to maternal but not paternal smoking in pregnancy, mean 2 for those exposed to paternal but not maternal smoking in pregnancy and mean $\mu_{\text{m true}}$ + 2 for those exposed to maternal and paternal smoking in pregnancy. Paternal smoking increased the outcome score by a value of 2 for all simulation settings and, as before, the value of $\mu_{\text{m true}}$ was varied between -5 and 5 in increments of 1.

3.2.2 – Results

3.2.2.1 – *Part 1: Simulation study of the influence of assortative mating on conclusions from the negative control design*

The bias of coefficient estimates against I_{PSI} is displayed in Figure 3-2. Part (i) of the figure shows that the maternal coefficient is unbiased in both Model 1 (maternal only model) and Model 3 (mutually adjusted model) for all quantities of assortative mating. This is true for positive and negative $\mu_{\text{M true}}$ values. Part (ii) of the figure shows there is no bias for the paternal coefficient in Model 3 but there is increasing absolute bias for Model 2 (paternal only model) with increasing assortative mating. No bias is observed at an I_{PSI} of 0. This represents the case where S_{exp} does not exist and so there is no backdoor path along $P \rightarrow \boxed{S_{exp}} \leftarrow M \rightarrow Y$.

A designed increase in the outcome in response to maternal smoking led to positive bias of the paternal coefficient in Model 2 while a designed decrease in the outcome in response to maternal smoking led to negative bias of the paternal coefficient. As a result, the modelled difference between the maternal and paternal coefficients from Model 1 and 2 would be smaller than the true difference when assortative mating occurs. I show this empirically in Figure 3-3 where I display the mean difference across simulations (and corresponding mean 95% CI for this difference) between the maternal and paternal coefficient against the I_{PSI} for different sample and effect sizes. As the quantity of assortative mating increased the difference in coefficients between Model 1 and 2 tended towards 0. The difference in coefficients within Model 3 were unaffected by assortative mating and accurately estimated the true difference.

As the quantity of assortative mating increases the collinearity between the maternal and paternal coefficient within Model 3 increases also. This can be problematic, particularly when the sample size is small. In Figure 3-3 the width of the confidence interval for the difference between coefficients within Model 3 becomes larger with increasing assortative mating. For small effect sizes this could lead to the conclusion of a null difference when one in fact does exist (see row 2, column 1 of the figure which shows a sample size of 100 and true difference of 1).

3.2.2.2 – *Part 2: Simulation study of a negative control design with assortative mating where the negative exposure influences the outcome independently of the exposure*

In part 2 of the simulation study I consider a scenario in which there is an independent effect of paternal smoking in pregnancy on the outcome. Again, this example is like that displayed in Figure 3-1i, but with an additional arrow from P to Y. The AOI and NCA now have the same backdoor paths, but where before only the NCA was biased by the effect size of the AOI, now both the AOI and NCA will be biased by each other where there is assortative mating in the exposure/negative exposure. Mutual adjustment for M and P can again eliminate this bias (but not bias by unadjusted confounding structures).

In the simulations I show that the introduction of a paternal effect leads to bias in the maternal coefficient in the presence of assortative mating for models that do not employ mutual adjustment (see Figure 3-4). The bias increases with increasing assortative mating. Bias in maternal β is the same for true maternal effect size of -5 as it is for +5 while the bias in the paternal β appears unchanged compared to that of the data where there is no paternal effect. This suggests that the size and direction of bias for each coefficient is dependent on the size and direction of the effect size of the other coefficient and not on the coefficient's own effect size. Models with mutual adjustment display no bias for either estimate in any setting.

Despite the introduction of bias to the maternal coefficient for the maternal only model (data in part 2) compared to data where there is no paternal effect on the outcome (data in part 1), there was little change in the pattern of results for the difference in coefficients between the data in part 1 and 2. Supplementary Figure A.3-1 (see Appendix A.3) shows the mean difference across simulations between the maternal and paternal coefficient against the I_{PSI} for the data in part 2. Comparison with Figure 3-3 shows very similar findings. This suggests that conclusions drawn from the maternal only and paternal only models will be influenced similarly by assortative mating in data where there is a paternal effect (NCA present) and where there is no paternal effect (null NCA).

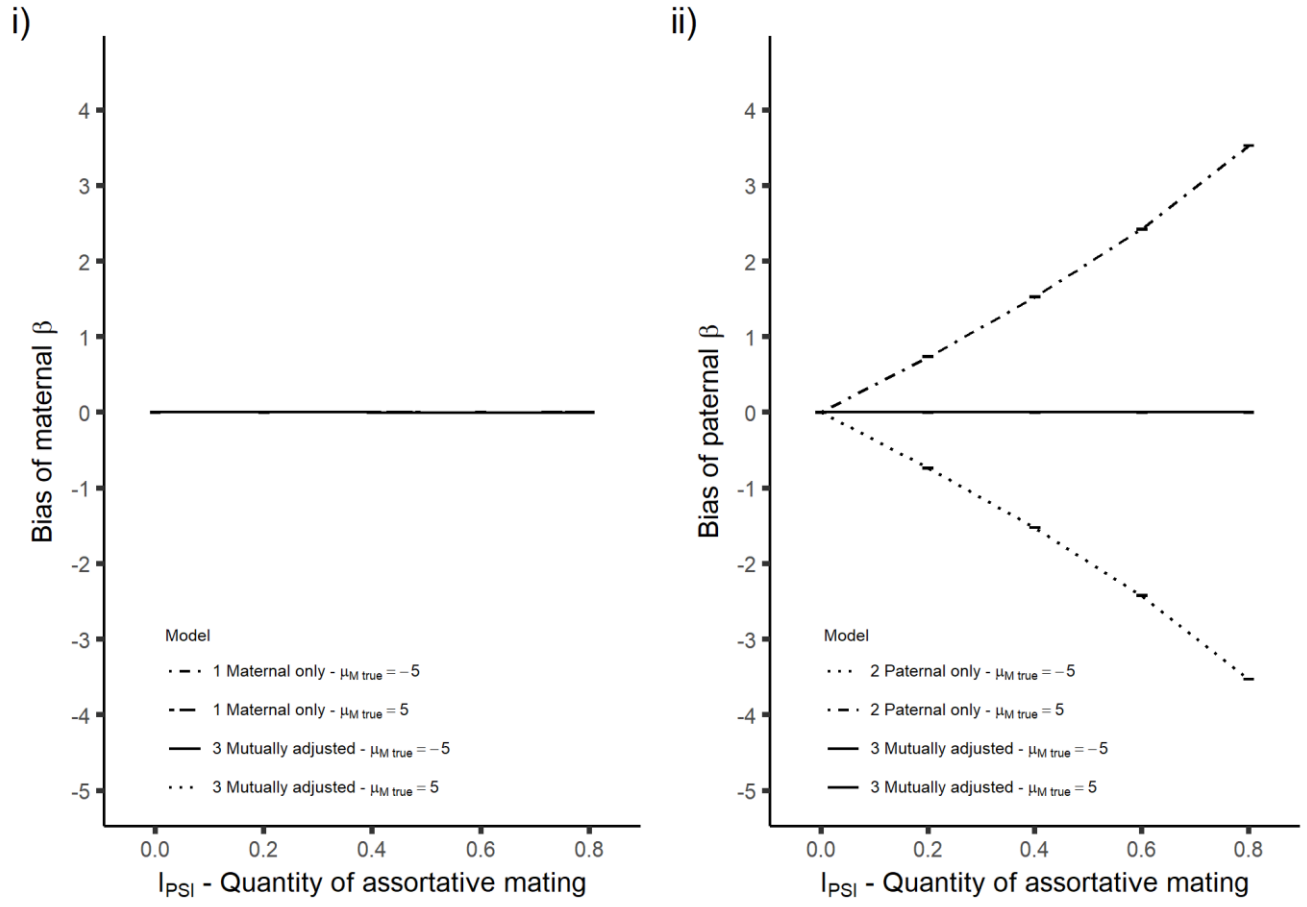

Figure 3-2: Plots of bias against quantity of assortative behaviour for continuous outcome data for a) the maternal coefficient and b) the paternal coefficient. Error bars are 95% Monte Carlo confidence intervals across simulations. Sample size for data shown is 10,000.

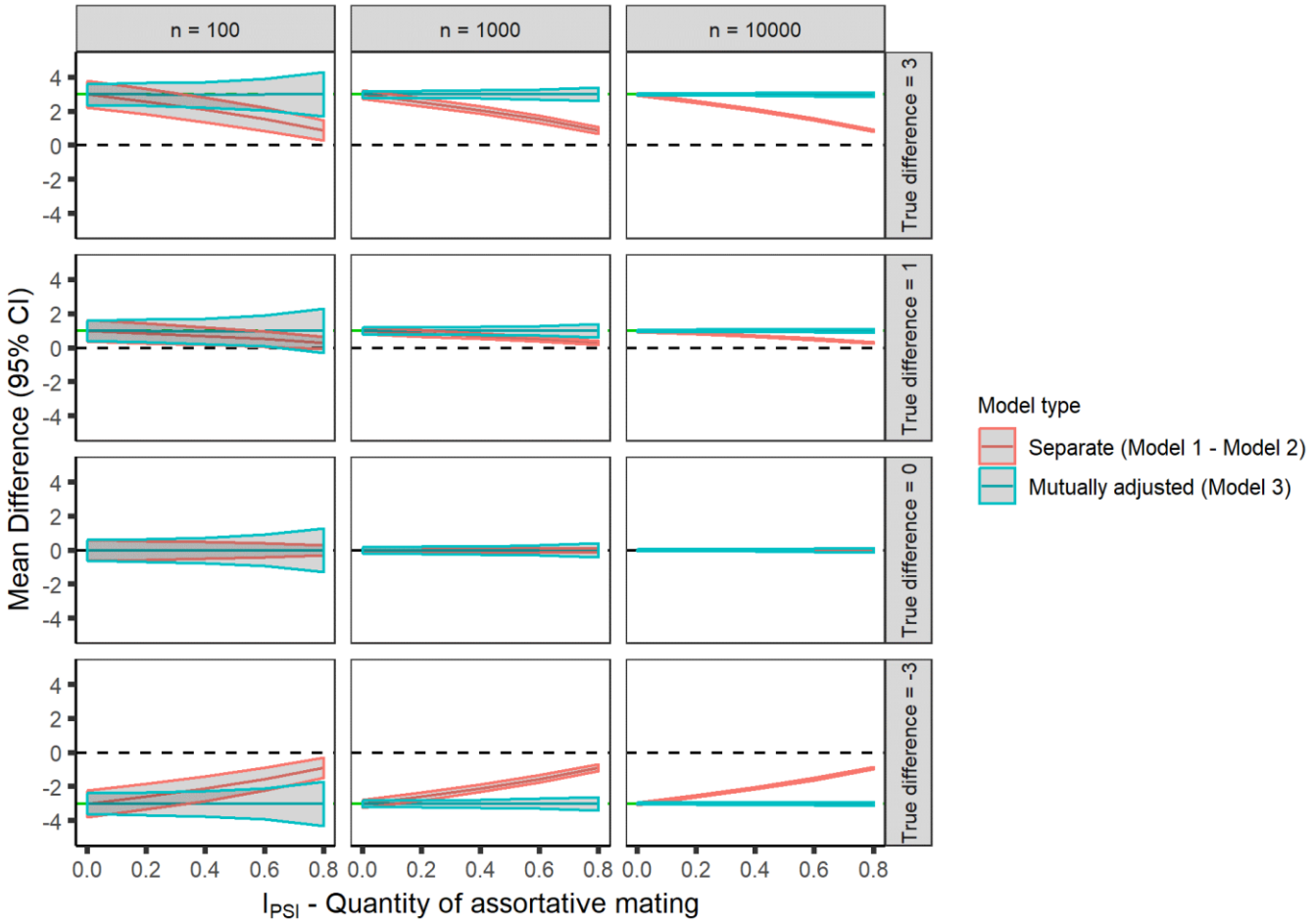

Figure 3-3: Plot of the mean difference across simulations of maternal and paternal β coefficients against the quantity of assortative mating. 95% confidence bands are the mean lower and upper CI for the difference, produced using bootstrapping. We present the difference between the coefficients of the maternal and paternal only models (red band) and the mutually adjusted model (blue band) for sample sizes of 100, 1 000 and 10 000.

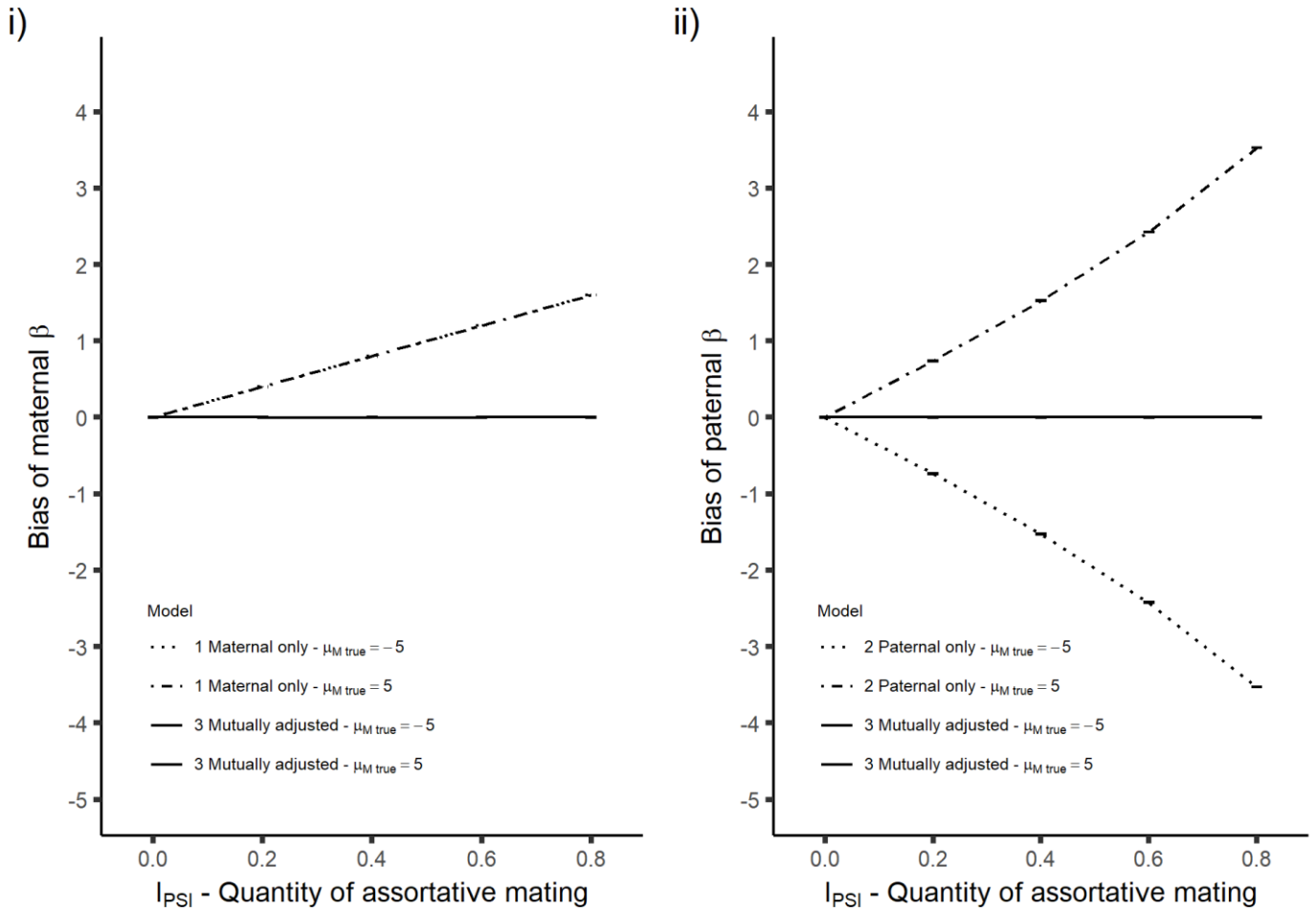

Figure 3-4: Plots of bias against quantity of assortative mating for continuous outcome data with a maternal and paternal effect for a) the maternal coefficient and b) the paternal coefficient. Error bars are 95% Monte Carlo confidence intervals across simulations. Sample size for data shown is 10,000.

3.3 – Discussion

In the negative control design, correlation between the exposure and negative exposure as a result of assortative mating leads to biased effect estimates where the two exposures have not been mutually adjusted for one another. The effect estimate of one exposure is biased by the "other" exposure (i.e. the effect size of the AOI leads to bias in the effect estimate of the NCA and vice versa). Assortative mating can therefore lead to more similar effect estimates between the exposure and negative exposure. This may lead to the erroneous conclusion that there is no causal effect when one may exist. Mutual adjustment resolves this by blocking the backdoor pathway that exists via the "other" exposure. However, when the quantity of assortative mating is high the strong correlation between exposure and negative exposure leads to large standard errors for mutually adjusted model coefficients as a result of high collinearity between the two variables. This is particularly true when the sample size is small. This makes the size of the difference between the AOI and NCA more ambiguous by enlarging the confidence interval.

An important assumption of the negative control design is that the confounding structure of the AOI is equivalent or shared with that of the NCA. It is also important to remember that confounders that have not been accounted for in models or that have not been well measured will still lead to bias. An alternative approach to dealing with assortative mating has been suggested in which the father's association is modelled only in families where the mother does not smoke [213]. A possible pitfall of this approach is that it may change the distribution of confounding factors across levels of maternal and paternal smoking behaviour in the dataset used for analysis, leading to bias even after mutual adjustment. As the AOI and NCA would no longer share the same confounding structure the two associations would be biased to different extents by confounders and so comparison of the two may not be useful. I would argue that mutual adjustment in a dataset that includes all families is a better strategy as it maintains equivalent confounding structures while blocking backdoor paths resulting from assortative mating.

Mutual adjustment is not able to resolve non-linear combinatory effects of exposure and negative exposure. There is evidence for differences in the smoking behaviours between couples who are concordant and discordant for smoking during pregnancy [193]. Concordant couples are likely to smoke when their partner is present while the smoking partner in a discordant couple is likely to smoke more cigarettes per day than in concordant couples. This is not something I have assessed in this study as I have only used a binary measure of smoking which would not have the ability to capture quantity of smoking. The influence on risk of outcome when using such a binary variable may therefore not be accurately represented by a model using a simple linear combination of maternal and paternal effect, as is done in the mutually adjusted model. It may be better to use categories of smoking concordance between parents (equivalent

to using an interaction term between exposure and negative exposure) to account for non-linear combinatory effects. However, if these categories are different to one another in underlying confounding structure then the negative control design may not be appropriate for this research question.

For simplicity I did not include confounding variables in the simulations. Where confounding variables are selected upon in assortative mating or are strongly correlated with a variable that is selected upon, the maternal and partner values are likely to correlate highly. The value of the maternal confounding variable may then be used as a proxy for the partner variable. Where there is low correlation between maternal and partner confounding variables and both are important predictors of the outcome, both may need to be included in the set of adjustment variables. To my knowledge the influence of this adjustment strategy on the bias of maternal and paternal association estimates has not been tested empirically. It is possible that adjustment for maternal but not paternal confounding variables may result in the NCA containing more bias than the AOI. Inclusion of confounding variables into the simulation study would have provided the opportunity to investigate whether imbalanced adjustment of maternal and paternal non-shared confounding variables influences the bias of point estimates. Recent debate on using paternal exposure as a negative control has highlighted the importance of including non-shared confounders between parents in models [214-216] and expanding Lipsitch et al.'s initial DAG of the negative control design [100] to reflect both shared and non-shared confounder variables. Brew and colleagues have suggested based on these expanded models that mutual adjustment may open biasing paths for the AOI and NCA as a result of the exposure/negative exposure being collider variables for shared and non-shared confounder variables [217]. These backdoor paths could be closed by controlling for the non-shared confounders of both parents provided they are measured.

Contamination effects, where one partners' behaviour influences the other were also not considered in the simulation study (for example, shared meals between partners may lead to correlation in BMI and changes in BMI [200] or one partner may convince the other partner to give up smoking). Brew et al. mention that mutual adjustment is commonly used to address the issue of contamination of exposure between partners, however, they make no attempt to investigate the appropriateness of the strategy in this situation [217]. This may be a more complicated scenario than that presented in this study, potentially requiring a different solution to resolve, and requires further investigation.

I did not consider the influence of measurement error in the DAGs or simulation study. This is a pertinent issue as in some cohorts the mother provides information on both her own and her partners exposure; the latter may suffer more from measurement error. Sanderson et al. [164] have shown that measurement error in the exposure or negative exposure will lead to biased

effect estimates (see Section 2.3.1 in chapter 2 for further detail). In Appendix A.4 I explored how the introduction of measurement error to the negative exposure variable can influence the conclusions of a negative control study in the context of an exposure affected by assortative mating. Briefly, error in a binary negative exposure can lead to bias by artificially increasing or decreasing the correlation between the exposure and negative exposure.

3.4 – Conclusions and chapter summary

When performing a negative control study in the presence of assortative mating, the estimates used for interpretation should be those of the mutually adjusted model, though this will not resolve all issues of the negative control design. I would suggest that future studies using paternal exposure as a negative control should perform a literature review to assess whether the exposure, negative exposure and relevant determinant variables may be involved in mate selection.

Chapter 4 Accounting for bias from missing data when the proportion of missing data is large

The contents of this chapter have been published in the following peer reviewed journal article:

- *Madley-Dowd P., Hughes R., Tilling K., and Heron J., The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of clinical epidemiology, 2019. 110: p. 63-73. DOI:* https://doi.org/10.1016/j.jclinepi.2019.02.016

The work in this chapter has been published in the Journal of Clinical Epidemiology [218] and is reproduced here. The work was a mini-project conducted at the start of the PhD with the intention of dispelling canards held by many reviewers in social science fields that multiple imputation should not be used where there is a substantial proportion of missing data. The Avon Longitudinal Study of Parents and Children, used in an empirical example in this study and also later used in Chapter 7 of the thesis, contains a large proportion of missing data that could bias complete case analyses. The current chapter is included in the thesis as a justification for applying multiple imputation methods to a dataset in which there is substantial missing data in order to reduce bias in effect estimates.

4.1 – Background

Missing data is a common problem in epidemiology, and participant drop out can substantially reduce the sample size available for analysis even in initially large cohorts. Missing data (also referred to as missingness) may cause bias and will always cause a reduction in efficiency. Analyses that account for missing data must consider the reasons for missingness (known as a missingness mechanism). Using Rubin's terminology [145], reasons for missing data are classified as: missing completely at random (MCAR) where the probability of missingness does not depend on either observed or missing data, missing at random (MAR) where conditional on the observed data the probability of missingness is independent of unobserved data and missing not at random (MNAR) where the probability of missingness is dependent on unobserved data even after conditioning on observed data. Readers may wish to refer to [219] or [220] for intuitive explanations of these terms.

A common approach [221] (and the default in most statistical packages) for dealing with missing data is complete case analysis (CCA), which restricts the analysis to individuals with complete data. An alternative to CCA, is multiple imputation (MI) [222, 223] which creates m copies of the dataset, replacing the missing values in each dataset with independent random draws from the predictive distribution of the missing values under a specific model (the imputation model). The analysis model is then fitted to each imputed dataset and the multiple results are combined into one inference using Rubin's rules [222]. The imputation model should contain all variables in the analysis model [224-226] as well as any interactions between variables [227]. The imputation model can additionally include variables not included in the analysis model, which are known as auxiliary variables. These are included to make the MAR assumption (required in the standard implementation of MI to produce unbiased estimates) more plausible, and to provide information about the missing values [156].

Researchers in a variety of fields often ask what proportion of missing data warrants the use of MI [151-154]. Varying guidance exists; in the literature 5% missingness has been suggested as a lower threshold below which MI provides negligible benefit [228]. In contrast one online tutorial has stated that 5% missing data is the maximum upper threshold for large datasets [229]. Statistical guidance papers have stated that bias is likely in analyses with more than 10% missingness and that if more than 40% data is missing in important variables then results should only be considered as hypothesis generating [230, 231].

The above suggested cut off points, with respect to specified proportions of missing data, have a limited evidence base to support them. A small number of studies have investigated bias and efficiency in datasets with increasing proportions of missing data. This has commonly been done with a maximum of 50% missing data in studies that showed increasing variability of effect estimates with increased missingness [232-234]; mixed results were found for bias. Where more than 50% missingness has been investigated, the use of auxiliary variables has often not been examined [235, 236]. Evidence of how varying quantities of missing data and auxiliary information jointly affect estimates obtained from MI is lacking in the literature as a result. The influence of the proportion of missing data on bias and efficiency (measured jointly using mean squared error) was shown to depend on the type of missingness (MCAR, MAR or MNAR) [235] and which variable (outcome, exposure or confounder) is missing [236]. Where both more than 50% missingness and auxiliary variables have been used the study sample size was very small ($N \leq 200$) thus limiting the applicability of results to larger epidemiological studies [237].

The proportion of missing data is a common measure of how much information has been lost because of missing values in a dataset. However, it does not reflect the information retained by auxiliary variables. Alternative measures such as the fraction of missing information (FMI) may

be more useful as a tool for determining potential efficiency gains from MI. The FMI is a parameter specific measure that is able to quantify the loss of information due to missingness, while accounting for the amount of information retained by other variables within a dataset [156, 238]. The FMI, derived from MI theory [222, 239], can be interpreted as the fraction of the total variance (including both between and within imputation variance, see Appendix B, section B.1) of a parameter, such as a regression coefficient, that is attributable to between imputation variance, for large numbers of imputations m. Values of FMI range between 0 and 1. A large FMI (close to 1) indicates high variability between imputed datasets; that is, the observed data in the imputation model does not provide much information about the missing values.

In this chapter, I have conducted a simulation study to show 1) that MI can be used to provide unbiased estimates with improved efficiency compared to CCA at any proportion of missing data, and 2) the utility of the FMI as a guide to the likely efficiency gains from using MI. I then use an applied example to show the influence of auxiliary information on the FMI, examining the association between maternal smoking during pregnancy and offspring intelligence quotient (IQ) score at age 15 using the Avon Longitudinal Study of Parents and Children (ALSPAC). Finally, I present a discussion of the findings and conclusions.

4.2 – Simulation study

4.2.1 – Methods

Via simulations, I compare FMI and the proportion of missing data to measure gain in information from MI compared to CCA, in scenarios with different available auxiliary information and amounts of missing data. The simulated datasets are motivated by a prospective cohort study where all baseline data are available, but some follow up data are missing.

4.2.1.1 - *Data model*

I simulated data from a multivariate normal distribution where all variables had a mean of 0 and a standard deviation of 1. Each simulated dataset contained 1000 observations on continuous variables outcome Y, exposure X and auxiliary variables $Z_1 - Z_{11}$. All variables were correlated with Y and all variables except Y had zero correlation with each other. The correlation between Y and X was 0.6, Y and $Z_1 - Z_2$ was 0.4, Y and $Z_3 - Z_7$ was 0.2 and finally between Y and $Z_8 - Z_{11}$ was 0.1.

Missingness was simulated under an MCAR mechanism to examine the benefit of MI to improve efficiency in the absence of bias, and an MAR mechanism to further examine bias reduction. The MCAR missingness mechanism removed the first p observations such that $\frac{p}{n}$ gives the required proportion of missing data. MAR missingness was simulated under a logistic regression model using

$$\text{logit}(\lambda_i) = \alpha + \text{Z}_{1i} + X_i.$$

The value of α was manipulated for the different simulation settings to provide the required proportion of missing data on average across datasets.

4.2.1.2 – *Analysis model*

For each simulation setting and imputation model the following linear regression analysis model was used:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where β_0 (true value equal to 0) and β_1 (true value equal to 0.6) are the intercept and exposure coefficient respectively, and ε_i are independently and identically distributed random errors with distribution $N(0, \sigma^2)$.

Each simulated dataset was analysed using CCA and MI. Where data were simulated as MCAR both MI and CCA are valid models [149]. For MAR data, with missingness dependent on X and Z_1, CCA is biased unless both X and Z_1 are included in the analysis model. For MAR data, MI is valid provided both X and Z_1 are included in the imputation model. MI was performed using the Stata [240] package *mi impute*. The analysis model, and the combination across imputed datasets using Rubin's rules, was implemented via Stata's *mi estimate.*

4.2.1.3 – *Imputation models*

Five imputation models were considered for both MCAR and MAR data (see Table 4-1). All models contained the variables included in the analysis model and used linear regression to impute the missing outcome. Model 1 contained no auxiliary information. Models 2-5 contained increasing quantities of auxiliary information, achieved by increasing the number of Z variables included in the imputation model. The squared coefficient of multiple correlation with the outcome variable, ${R_Y}^2$, was used as a measure of the quantity of auxiliary information. This reflects a sum of the independent contributions of each auxiliary variable to the imputation model.

For each imputation model 1000 imputations were run. FMI is a highly variable estimate at low numbers of imputations [241] hence the need for a large number of imputations. See Appendix B.2 for why I chose 1000 imputations.

Table 4-1. Description of the Imputation Models Used for Both MCAR and MAR Data.

Imputation model	Variables included	$R_Y{}^2$ [a]
1 (least auxiliary information)	*Y, X*	0.36
2	*Y, X, Z_3*	0.40
3	*Y, X, Z_1*	0.52
4	*Y, X, Z_{1-4}*	0.76
5 (most auxiliary information)	*Y, X, Z_{1-11}*	0.92

[a] $R_Y{}^2$, the total coefficient of multiple correlation with the outcome Y for all variables included in the imputation model, is displayed as a measure of the strength of the auxiliary information in each imputation model

4.2.1.4 – *Comparisons*

I repeated the simulation study for 1%, 5%, 10%, 20%, 40% 60%, 80% and 90% missing data. For all scenarios, I generated 1000 independent simulated datasets. Separately for the exposure coefficient and the constant coefficient, I compared the complete case analysis and multiple imputation analyses with respect to the bias, empirical standard error and FMI of the coefficient estimates. Bias and empirical standard error were estimated using the *simsum* command in Stata [212], and FMI was calculated using Stata's *mi estimate*. I report the median value and interquartile range of the FMI across simulations. Further measures are described and presented in Appendix B.3 along with formulae for all performance statistics.

4.2.2 – Results

Figure 4-1 displays the empirical SE of the MI exposure coefficient against the FMI, according to proportions of missing data (see Appendix B.4 for presentation of the data separated by panels of percentage missing data, which demonstrates that for any given proportion of missing data, the empirical SE increases as the FMI increases – with this association being most noticeable at high proportions of missing data). For every value of the proportion of missing data, the FMI for models with no auxiliary information was approximately equal to the

proportion of missing data. The FMI decreased with increasing quantities of auxiliary information. For different proportions of missing data but similar FMI values the empirical SE of MI coefficient estimates was approximately the same. For example, compare model 2 for 40% missing data (FMI=0.38, empirical SE=0.032) with model 4 for 60% missing data (FMI=0.37, empirical SE=0.031) and model 5 for 80% missing data (FMI=0.35, empirical SE=0.030). A second example is given by the comparison of model 1 for 60% missing data (FMI=0.60, empirical SE=0.039), model 4 of 80% missing data (FMI=0.63, empirical SE=0.041) and model 5 of 90% missing data (FMI=0.56, empirical SE=0.039) while a third example is given by model 2 for 80% missing data (FMI=0.79, empirical SE=0.055) and model 4 for 90% missing data (FMI=0.78, empirical SE=0.054). This indicates that the FMI is a good measure of estimate precision while the proportion of missing data is not.

Table 4-2 displays the percentage reduction in empirical SE compared to CCA for each MI model. Increasing auxiliary information in the imputation model led to increasing gains in efficiency (greater reduction in empirical SE) with greater effects seen at larger proportions of missing data. For low proportions of missing data there was little efficiency gain from MI even for the model with the largest quantity of added auxiliary information.

Figure 4-2 shows that for CCA there are increasing levels of bias in estimating the exposure coefficient with increasing proportions of missing data. A single exception to this occurs at 90% missing data which may be due to increased variability of the estimate. For MI, no bias was observed at any proportion of missing data provided the imputation model included all variables related to missingness (models 3-5). These findings provide an example of valid estimates from properly specified MI at much larger proportions of missing data than current guidance [231] advises. When the imputation model did not include these variables (models 1-2) then the magnitude of bias was similar to that of CCA.

All performance statistics for the exposure coefficient across simulations of MCAR and MAR data are presented in Appendix B.5. With respect to FMI and efficiency of the MI estimates, the results for the MAR scenario followed the same patterns as noted for the MCAR scenario. The results of FMI and efficiency gains were similar when missingness depended on the auxiliary variable and when missingness did not depend on the auxiliary variable (presented in Appendix B.6).

Table 4-2. Percentage Reduction in Empirical SE and Bias Compared to CCA for MCAR and MAR Results of the Exposure Coefficient in the Simulation Study.

% Missing	Imputation model [a b]	% Reduction in SE compared to CCA [c] MCAR data	MAR data	% Reduction in bias compared to CCA [d] MAR data
1	1: R^2=0.36 (No aux info)	0.00%	-0.01%	1.46%
	2: R^2=0.40	0.16%	0.24%	1.91%
	3: R^2=0.52	0.24%	0.11%	79.03%
	4: R^2=0.76	0.55%	0.41%	79.54%
	5: R^2=0.92	0.52%	0.58%	81.42%
5	1: R^2=0.36 (No aux info)	0.02%	-0.03%	0.16%
	2: R^2=0.40	0.19%	0.03%	-1.26%
	3: R^2=0.52	1.04%	0.93%	97.92%
	4: R^2=0.76	1.99%	2.63%	94.91%
	5: R^2=0.92	1.57%	3.64%	93.74%
10	1: R^2=0.36 (No aux info)	-0.05%	-0.06%	0.40%
	2: R^2=0.40	0.37%	0.75%	-0.35%
	3: R^2=0.52	0.58%	1.12%	97.38%
	4: R^2=0.76	2.59%	4.61%	96.73%
	5: R^2=0.92	2.89%	6.76%	96.41%
20	1: R^2=0.36 (No aux info)	0.03%	-0.05%	-0.19%
	2: R^2=0.40	1.08%	1.03%	-0.65%
	3: R^2=0.52	2.59%	3.42%	97.94%
	4: R^2=0.76	8.28%	7.94%	97.33%
	5: R^2=0.92	10.53%	10.26%	97.29%
40	1: R^2=0.36 (No aux info)	0.05%	-0.06%	-0.21%
	2: R^2=0.40	2.00%	1.25%	0.10%
	3: R^2=0.52	5.37%	5.06%	97.84%
	4: R^2=0.76	15.56%	14.11%	98.56%
	5: R^2=0.92	21.10%	22.86%	98.64%
60	1: R^2=0.36 (No aux info)	-0.04%	-0.02%	0.21%
	2: R^2=0.40	2.55%	1.68%	0.02%
	3: R^2=0.52	5.48%	6.74%	99.77%
	4: R^2=0.76	21.02%	18.45%	99.43%
	5: R^2=0.92	31.59%	31.96%	98.22%
80	1: R^2=0.36 (No aux info)	-0.03%	-0.14%	0.00%
	2: R^2=0.40	2.16%	1.57%	1.34%
	3: R^2=0.52	8.18%	9.86%	96.47%
	4: R^2=0.76	27.56%	28.21%	99.62%
	5: R^2=0.92	45.88%	44.66%	98.77%
90	1: R^2=0.36 (No aux info)	0.03%	0.11%	0.04%
	2: R^2=0.40	1.40%	2.18%	0.89%
	3: R^2=0.52	12.44%	8.86%	99.97%
	4: R^2=0.76	34.82%	33.76%	95.78%
	5: R^2=0.92	53.09%	52.96%	98.73%

CCA – Complete case analysis; MAR – Missing at random; MCAR – Missing completely at random; SE – Standard error

[a] R^2 refers to the squared coefficient of multiple correlation which is used as a measure of auxiliary information

[b] Models 1 and 2 do not include all variables in the missingness mechanism and so are biased (as expected) for the MAR data. Models 3-5 do include all variables in the missingness mechanism and so are unbiased (as expected).

[c] Calculated using $100 \times (se_{CCA} - se_{MI})/se_{CCA}$ where se_{CCA} and se_{MI} are the empirical standard error of the CCA model and the MI model respectively

[d] Calculated using $100 \times (abs(bias_{CCA})\text{-}abs(bias_{MI}))/abs(bias_{CCA})$ where abs(.) is a function giving the absolute value and $bias_{CCA}$ and $bias_{MI}$ are the bias of the CCA model and the MI model respectively.

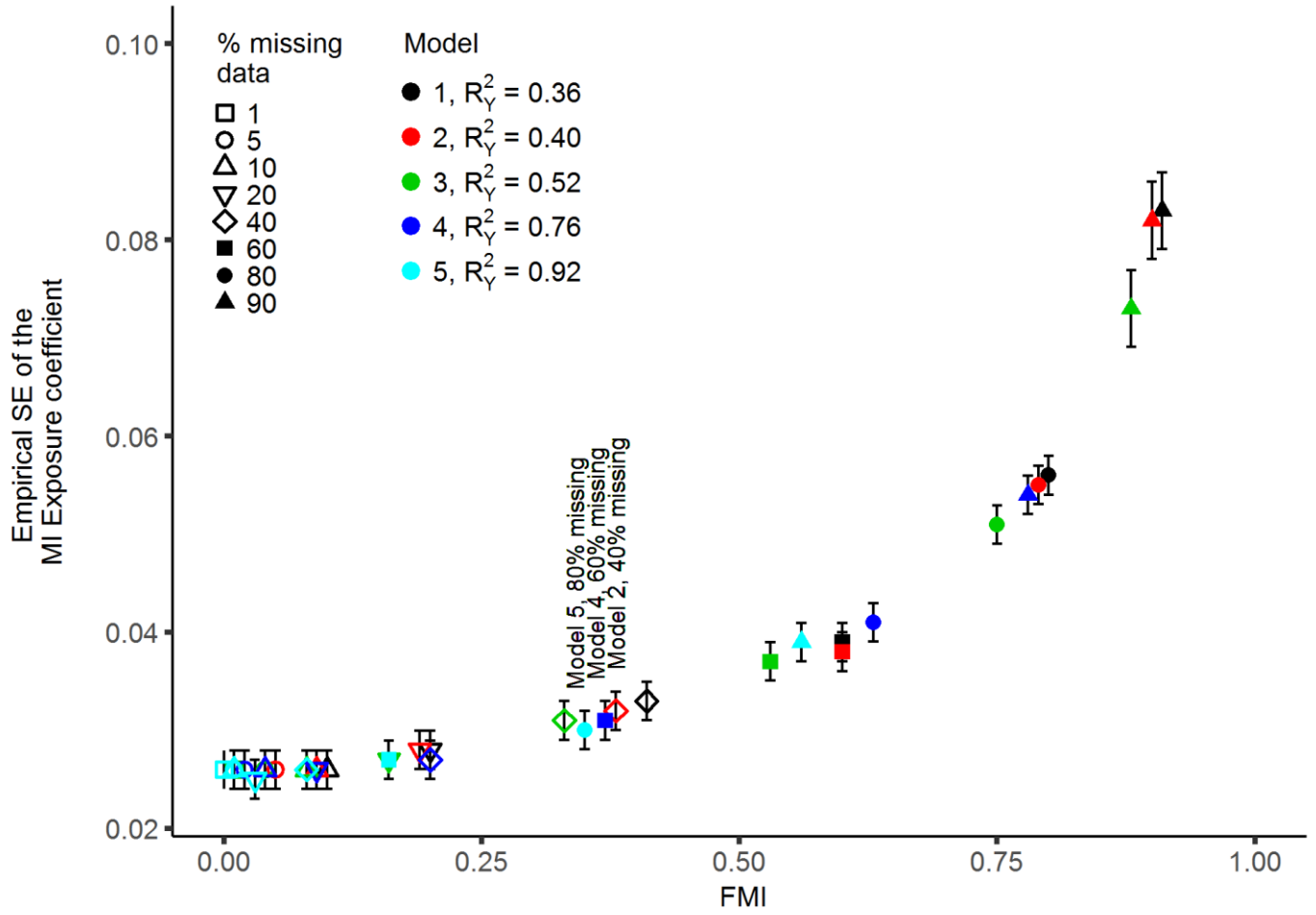

Figure 4-1: Empirical SE of the MI exposure coefficient plotted against FMI for simulated MCAR data. Error bars are 95% confidence intervals based on Monte Carlo standard errors across simulations. FMI = fraction of missing information; SE = standard error.

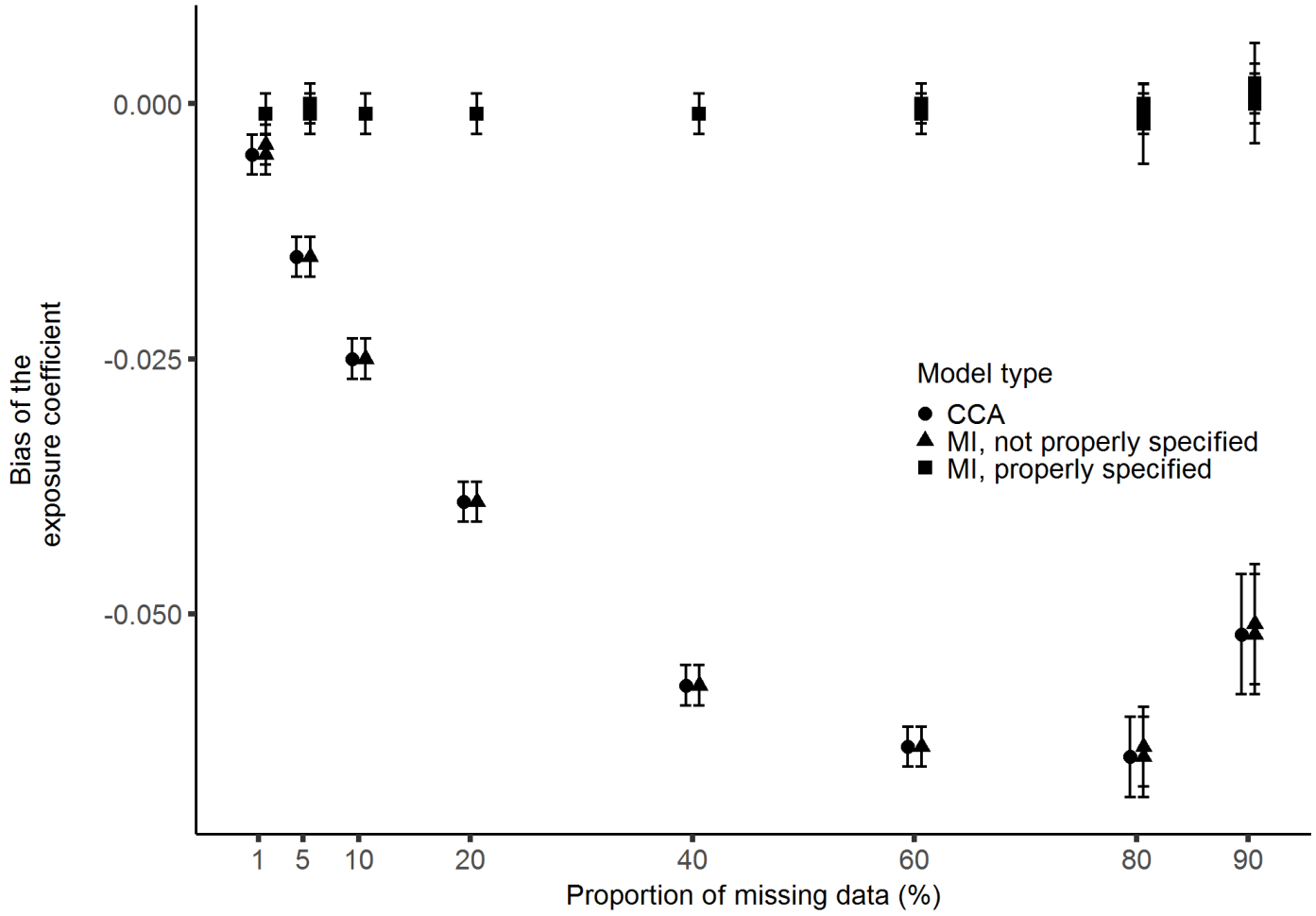

Figure 4-2: Bias of the CCA and MI exposure coefficient plotted against the proportion of missing data for simulated MAR data. Error bars are 95% confidence intervals based on Monte Carlo standard errors across simulations. MI = multiple imputation; FMI = fraction of missing information; SE = standard error

4.3 – Applied example

4.3.1 – Ethical approval

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees - http://www.bristol.ac.uk/alspac/researchers/research-ethics/.

4.3.2 – Methods

Data were taken from ALSPAC [242, 243] which recruited 14,541 pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992. Of these pregnancies, there were 13,988 children who were alive at 1 year of age. Please note the study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bristol.ac.uk/alspac/researchers/our-data/).

I investigated the relationship between a binary measure of maternal smoking during pregnancy, self-reported at 18 weeks gestation and offspring IQ measured using the Wechsler Abbreviated Scale of Intelligence at age 15 years [244]. The substantive analysis was a linear regression of offspring IQ at age 15 on maternal smoking in pregnancy. I shall refer to this as the "unadjusted" analysis. I also considered an "adjusted" analysis which controlled for the possible confounders maternal age, parity and education and offspring sex.

In order to simplify this illustrative example, observations were removed if they had missing data for any of the confounders. The justification for this decision is that these variables were measured at the start of the study and if they were missing then the participant was likely to be missing data in most other variables. Appendix B.7 shows excluded participants with missing values in the confounders were more likely to have a larger number of missing variables for the outcome, exposure and auxiliary variables. This exclusion criteria left a total sample size of $n = 11911$. Among the included participants the exposure was fully observed. See Appendix B.8 for the patterns of missing data for the outcome and auxiliary variables.

The auxiliary variables used in imputation models were IQ at age 8 measured using the Wechsler Intelligence Scale for Children - III [245], intelligibility and fluency at age 9 measured using the Children's Communication Checklist [246], a binary indicator of ever having learning difficulties, and, measured in school year 6, the child's teacher-reported maths and literacy streaming groups as well as the score from a maths assessment.

I performed chained equations imputation [247] using Stata's *mi impute chained* command with 1000 imputations. I used this large number of imputations to ensure that a reliable estimate of

the FMI was obtained. Twelve imputation models with differing amounts of auxiliary information were investigated. A description of the variables included in each model is displayed in Table 4-3. Model A contains only the confounders in the adjusted model, model B-E include one auxiliary variable each. Model F includes one variable each for the maths and literacy streaming groups. Models G-L include differing combinations of auxiliary variables.

The same imputation models were used for the unadjusted and adjusted analyses. For a given analysis model, an imputation model was defined as containing auxiliary variables if it included variables that were not in the analysis model. So, for the unadjusted analysis every imputation model contained auxiliary variables, whereas for the adjusted analysis, the simplest imputation model contained no auxiliary variables.

Table 4-3. Imputation Models for the Applied Example, Bristol, United Kingdom, 1991-2007.

Model	Variables included [a]	% missing data
A	No extra variables	62.47%
B	IQ at age 8	66.64%
C	Intelligibility and fluency at age 9	66.68%
D	Maths assessment score	76.59%
E	Learning difficulties	78.84%
F	Streaming for maths and English	81.75%
G	IQ at age 8 and intelligibility	69.34%
H	IQ at age 8 and maths assessment	79.11%
I	IQ at age 8, intelligibility and maths assessment	80.62%
J	IQ at age 8, intelligibility, maths assessment and LD	84.17%
K	IQ at age 8, intelligibility, maths assessment and streaming groups	86.42%
L	IQ at age 8, intelligibility, maths assessment, LD and streaming groups	86.51%

[a] All models additionally contained IQ at age 15, a binary measure of maternal smoking in pregnancy and the set of all confounders. Continuous variables (IQ at age 8 and 15, intelligibility and maths assessment score) were imputed using a linear regression model, binary variables (sex and learning difficulties) were imputed using logistic regression, and ordinal variables (maternal age and education, parity and maths and literacy streaming group) were imputed using ordinal logistic regression.

4.3.3 – Results

Table 4-4 shows that the proportion of missing data in the outcome variable was 62%, with all auxiliary variables having a lower proportion of missing data. IQ at age 8 and maths assessment score explained the most variance in the outcome. Intelligibility and ever having learning difficulties were the weakest predictors. The exposure and all confounder and auxiliary variables were associated with the likelihood of missingness in the outcome variable.

The results for the estimate, standard error, FMI and percentage reduction in SE compared to CCA for the exposure coefficient of the adjusted linear regression are presented in Figure 4-3. The estimated association between maternal smoking and IQ is further from the null when the imputation model includes more variables. The estimates provided by the CCA model would lead to different conclusions to those provided by MI model H-L.

Figure 4-3 shows that for the exposure coefficient, the MI standard errors for most imputation models were smaller than that of CCA; models A, C and E are exceptions displaying slight increases, likely due to these models containing low levels of auxiliary information. No model led to larger FMI than that of model A which included no auxiliary information.

Including more than one auxiliary variable in the imputation model had inconsistent influence on FMI and SE for the exposure coefficient. For example, the addition of intelligibility to model B (see model G) lead to increased FMI and a reduced gain in efficiency versus CCA, as measured by percentage reduction in SE. The addition of the maths assessment score to model B (see model H) lead to the greatest estimate precision and lowest FMI. Once intelligibility had been added to model H (see model I-L) further addition of variables to the model could not achieve the efficiency gains observed in model H. It is possible that this is because missing information in intelligibility led to increased variability that could not be counteracted by introducing further information about missing outcomes via the inclusion of more auxiliary variables. The confidence intervals of the exposure coefficient estimates overlap for all imputation models investigated.

Comparison of Figure 4-3 with the figure in Appendix B.9 shows that greater reductions in efficiency, relative to CCA, were made when the analysis model was an unadjusted model. This is because confounders are likely to explain some of the covariation between the exposure and outcome as well as some of the missingness in the outcome. The remaining unexplained variation that is available to be accounted for by auxiliary variables is therefore less in the adjusted models.

Table 4-4: Variable Description, Including the Proportion of Missing Data and Relationship with Observed and Missing Values in the Outcome Variable for the Applied Example, Bristol, United Kingdom, 1991-2007.

Variable	Type	% missing data	R^2 with outcome [a]	OR for missing data in outcome [b]	95% CI [b]
IQ at age 15	Continuous	62.47			
Maternal smoking in pregnancy	Binary	0.00	0.01	2.18	1..98, 2.39
Maternal age	Categorical	0.00	0.04		
	- ≤ 24 years			Reference	Reference
	- 25-29 years			0.57	0.51, 0.64
	- 30-34 years			0.42	0.38, 0.47
	- ≥ 35 years			0.41	0.35, 0.47
Parity	Categorical	0.00	0.01		
	- 0			Reference	Reference
	- 1			1.18	1.09, 1.29
	- 2			1.46	1.30, 1.64
	- ≥ 3			2.06	1.72, 2.48
Sex	Binary	0.00	<0.01		
	- female			Reference	Reference
	- male			1.27	1.18, 1.37
Maternal education	Categorical	0.00	0.11		
	- Vocational			Reference	Reference
	- CSE/O level			0.91	0.80, 1.05
	- A level/Degree			0.45	0.39, 0.52
IQ at age 8	Continuous	44.49	0.37	0.98	0.98, 0.98
Intelligibility and fluency at age 9	Continuous	37.96	0.01	0.95	0.93, 0.97
Maths assessment score	Continuous	44.39	0.24	0.15	0.12, 0.19
Ever had learning difficulties	Binary	48.57	0.08	2.02	1.75, 2.33
Maths streaming group	Ordinal	52.76	0.20		
	- lowest			Reference	Reference
	- middle			0.58	0.50, 0.69
	- highest			0.42	0.36, 0.49
Literacy streaming group	Ordinal	55.03	0.16		
	- lowest			Reference	Reference
	- middle			0.59	0.50, 0.69
	- highest			0.39	0.33, 0.45

CCA – complete case analysis; CI – Confidence interval; IQ – Intelligence quotient; OR – Odds ratio; R^2 – variance explained in the outcome

[a] Regressed IQ at age 15, on each variable with no adjustment for other variables. CCA analysis was used in all models.

[b] Using logistic regression, the odds of having a missing value for the outcome were regressed on each variable with no adjustment for other variables. CCA analysis was used in all models.

Model	Exposure Coefficient	SE	FMI	% reduction in SE
CCA	-0.58	0.497		
A: No auxiliary information	-0.58	0.504	0.724	-1.39
B: IQ at age 8	-0.89	0.459	0.667	7.65
C: Intelligibility at age 9	-0.53	0.505	0.724	-1.62
D: Maths assessment score	-0.84	0.472	0.683	5.06
E: Learning Disabilities	-0.80	0.501	0.716	-0.80
F: Streaming for Maths and English	-0.86	0.476	0.686	4.34
G: IQ at age 8 and intelligibility at age 9	-0.87	0.470	0.682	5.39
H: IQ at age 8 and maths assessment score	-1.04	0.453	0.656	8.80
I: IQ, intel. and maths assessment	-1.04	0.458	0.663	7.78
J: IQ, intel., maths assessment and LD	-1.14	0.456	0.656	8.19
K: IQ, intel., maths assessment and streaming groups	-1.10	0.457	0.658	8.17
L: IQ, intel., maths assessment, LD and streaming groups	-1.13	0.461	0.664	7.17

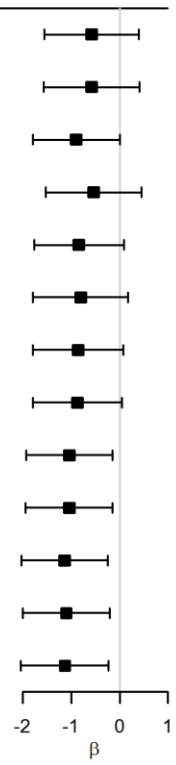

Figure 4-3: Estimate, standard error and FMI for the exposure coefficient in the applied example adjusted analysis model. Reduction in SE is relative to CCA. CCA=complete case analysis; FMI= fraction of missing information; SE = standard error.

4.4 – Discussion

This study showed that at all proportions of missingness in the outcome there is benefit to using MI in terms of reducing bias and improving efficiency, and that FMI can be used as a better guide to the efficiency gains to be made from MI than the proportion of missing data. I found that, compared to CCA, MI with auxiliary information improved efficiency of effect estimates at any proportion of missing data. Provided the imputation model was correctly specified and included all variables related to missingness then MI eliminated bias when data were MAR regardless of the amount of missing data. CCA was always biased because the analysis model did not include all variables related to missingness [149, 223, 248]. The simulations (both MCAR and MAR) revealed that similar FMI values can result from datasets with differing proportions of missing data if they have differing amounts of auxiliary information. In models with the same FMI, the empirical SE was approximately equal despite the different proportions of missing data. The biggest factor affecting the gain in precision of effect estimates from using multiple imputation is therefore not the proportion of missing data but instead the FMI.

The results of the applied example show that auxiliary information influences the standard error and FMI of effect estimates in a real world dataset. The example also demonstrates that the introduction of extra variables to the imputation model, without reducing the FMI, can be harmful to the precision of model estimates. This can likely be explained by the additional missing data in the auxiliary variable leading to a loss in estimate precision. Out of all models tested I would recommend the use of model J because it had the lowest FMI and included more variables that predicted missingness than model H which had an equivalent FMI. Model L additionally included the streaming group variables, which also predicted missingness, but there was very little difference in the coefficient estimate compared to model J while its FMI was greater than model J.

An inclusive strategy of auxiliary variables has been suggested as preferable to a restrictive strategy to try to include all variables that may be associated with the missingness mechanism [249]. Using too many auxiliary variables is harmful, however, when the sample size is small [250]. This leads to a ratio of observed values to model parameters that is close to unity which in turn leads to poor model fit. Where the sample size is large an inclusive strategy of auxiliary variables is acceptable – however the results show that the FMI should be checked to see whether missing data in auxiliary variables decreased efficiency (as in the applied example). Those variables which make the MAR assumption plausible should always be included in the imputation model.

The simulation study was limited by its single sample size, simple analysis model and that I considered missingness in only one variable. In real world datasets auxiliary variables are often

correlated, which will reduce the independent contribution of each variable to the imputation model but may aid in prediction of missing values in an auxiliary variable itself. Missingness often occurs in several variables within a dataset, although this should not bias the estimate of the effect of exposure on outcome, provided missingness is not related to the outcome (for CCA) or that all variables are MAR (for MI) [149]. Sample size has been shown to influence efficiency gains obtained via MI for binary outcomes [237] with smaller sample sizes associated with smaller gains at equivalent proportions of missing data. It is possible that greater efficiency gains could be achieved at the smaller proportions of missing data than was observed in this study if a greater sample size was used. Bias reduction has also been found to be greater with increasing sample size for longitudinal data [234]. Finally, I have only investigated correctly specified MI – if the imputation model is incorrectly specified, the bias may not be completely removed, or could even be larger than in the CCA [226, 227, 251]. In practice, the variables related to missingness are seldom known with certainty.

Further work needs to investigate the applicability of the results to models with binary and time to event outcomes. Logistic regression sometimes differs to linear regression with regard to missing data; for example, logistic regression is more robust to bias in the presence of missing data [147]. In Appendix B.10 I display a simple example of the simulation study for a binary outcome. For MI of a logistic regression analysis model, the simulation results show that the FMI is reduced with increasing auxiliary information, which was also shown by the results of the simulation study for the linear regression model. More thorough investigation is warranted.

This study is the first to investigate the influence of increasing auxiliary information on bias and efficiency of MI analyses at proportions of missing data greater than 50% missingness. Studies that have looked at large proportions of missing data, in the absence of auxiliary information, have also shown MI to reduce bias and improve efficiency over CCA [235, 236]. These studies highlighted the importance of a properly specified imputation model to reducing bias.

For MI to be valid, the data must be MAR (given the variables in the imputation model), and both analysis and imputation models must be correctly specified. This may be harder to investigate as the number of participants with complete data (rather than the proportion of the sample with missing data) decreases. For example, investigating whether interactions or non-linearities need to be included in the imputation model will be harder as the number of complete cases gets smaller. However, the complete case analysis also depends on the analysis model being correctly specified, and data being MAR given the variables in the analysis model. These assumptions will be similarly hard to investigate as the number of complete cases decreases. Thus, where conclusions are being drawn from a small number of complete cases, I recommend sensitivity analyses to explore a range of plausible analysis and imputation models, as well as the impact of deviations from MAR [226, 252].

These results have important implications for epidemiologists, and reviewers, for the conduct and reporting of analysis on incomplete data. The results imply that researchers should consider whether all the variables related to missingness can plausibly be included in the imputation model (to limit bias), and then whether there are auxiliary variables that can lower the FMI (to improve efficiency). I recommend that all papers reporting results of analyses with incomplete data show: a table of characteristics of those with complete data vs those with incomplete data (to assess factors associated with missingness) and a table showing variables associated with incomplete variables (to assess auxiliary information). The FMI of multiple imputation analyses should be reported, along with a discussion of whether it is plausible that all variables related to missingness have been included in the imputation models.

A key finding of this study is that the proportion of missing data should not be used as a guide to whether to use MI (or CCA) or not – I have shown that correctly specified MI can reduce bias and improve efficiency for analysis of MAR data at any proportion of missingness. If the imputation model cannot be correctly specified, then alternatives to MI such as inverse probability weighting [155] or study-specific sensitivity analysis may be preferable. This work shows that the FMI provides better insight into the amount of information retained using MI than does the proportion of missing data. It may be useful to check the FMI when adding auxiliary variables to an imputation model to see which variables are not adding information (e.g. due to the proportion of missing data in an auxiliary variable).

PART 2: EMPIRICAL INVESTIGATIONS

Chapter 5 Maternal smoking during pregnancy and offspring intellectual disability: sibling analysis in an intergenerational Danish cohort

The contents of this chapter have been published in the following peer reviewed journal article:

- *Madley-Dowd P., Kalkbrenner A.E., Heuvelman H., Heron J., Zammit S., Rai D., and Schendel D., Maternal smoking during pregnancy and offspring intellectual disability: sibling analysis in an intergenerational Danish cohort. Psychological medicine, 2020: p. 1-10. DOI: https://doi.org/10.1017/S0033291720003621*

In Chapter 1 I described that a systematic review has suggested that smoking during pregnancy is associated with a small increase in the risk of offspring intellectual disability (ID) [97], though the studies included did not adequately account for confounding or information bias. Two better quality studies not included in the review found an association between smoking in pregnancy and offspring risk of ID but both suggested that this may be the result of residual confounding [98, 99]. Further triangulation of evidence from different causal inference techniques is required to establish whether such an interpretation is likely [135, 160, 253]. In contrast, the association between smoking in pregnancy and offspring fetal growth restriction has strong evidence of being causal in nature from complementary causal inference designs [8, 9, 183]. This latter association can be used as a positive control for smoking in pregnancy and offspring ID. By this I mean that, using the same causal inference methods, if an association is found for fetal growth restriction but not ID then this will support the interpretation that observational associations with ID are the result of residual confounding.

Important questions are outstanding in the current literature, including whether the association between maternal smoking and offspring ID differs by offspring gender or with the presence of other comorbid disorders, and whether timing and dosage of exposure are associated with changes in the strength of association. In the first empirical chapter of this thesis I aimed to investigate the association between maternal smoking during pregnancy and risk of ID in offspring, and assess causality, using data from a large Danish population-based cohort with data available on parents and siblings. Secondary aims were to investigate the association among subgroups (separated by severity of ID, comorbid ASD and ADHD, gender) and the associations for different timings and dosages of exposure.

5.1 – Methods

5.1.1 – Ethics approval

This study was approved by the Danish Scientific Ethics Committee, the Danish Health Data Authority, the Danish Data Protection Agency and the Danish Neonatal Screening Biobank Steering Committee. Consent from individuals for this register-based study using anonymised data was not required.

5.1.2 – Cohort for analysis

The study cohort consisted of all individuals born in Denmark between January 1st 1995 and December 31st 2012 (n=1 337 491). After excluding children not born in Denmark, those who died or emigrated before the age of 1, those who had a missing link to a maternal or paternal identifier and those who had a known genetic or metabolic cause of intellectual disability (see Table C.1.1-1 in Appendix C), the remaining sample included 1 119 146 individuals (study flow chart in Figure 5-1). The cohort and analytic variables were defined using several registry datasets linked by a unique identification number [254]: the Danish Medical Birth Registry (MBR) [255], the Danish Psychiatric Registry (DPR) [256], the Danish National Patient Registry (NPR) [257] and Statistics Denmark registries of education [258] and income [259].

Most clinical contacts related to intellectual disability occurred in an outpatient setting. I therefore defined the start year of the cohort as 1995 when the DPR and NPR started recording outpatient contact in addition to inpatient admissions. I selected 2012 as the end year for inclusion in the cohort to allow a minimum of 4 years follow up until the latest date for available data, 10th of April 2017. The youngest and oldest members of the cohort were followed up until approximately 4.3 and 22.3 years of age respectively.

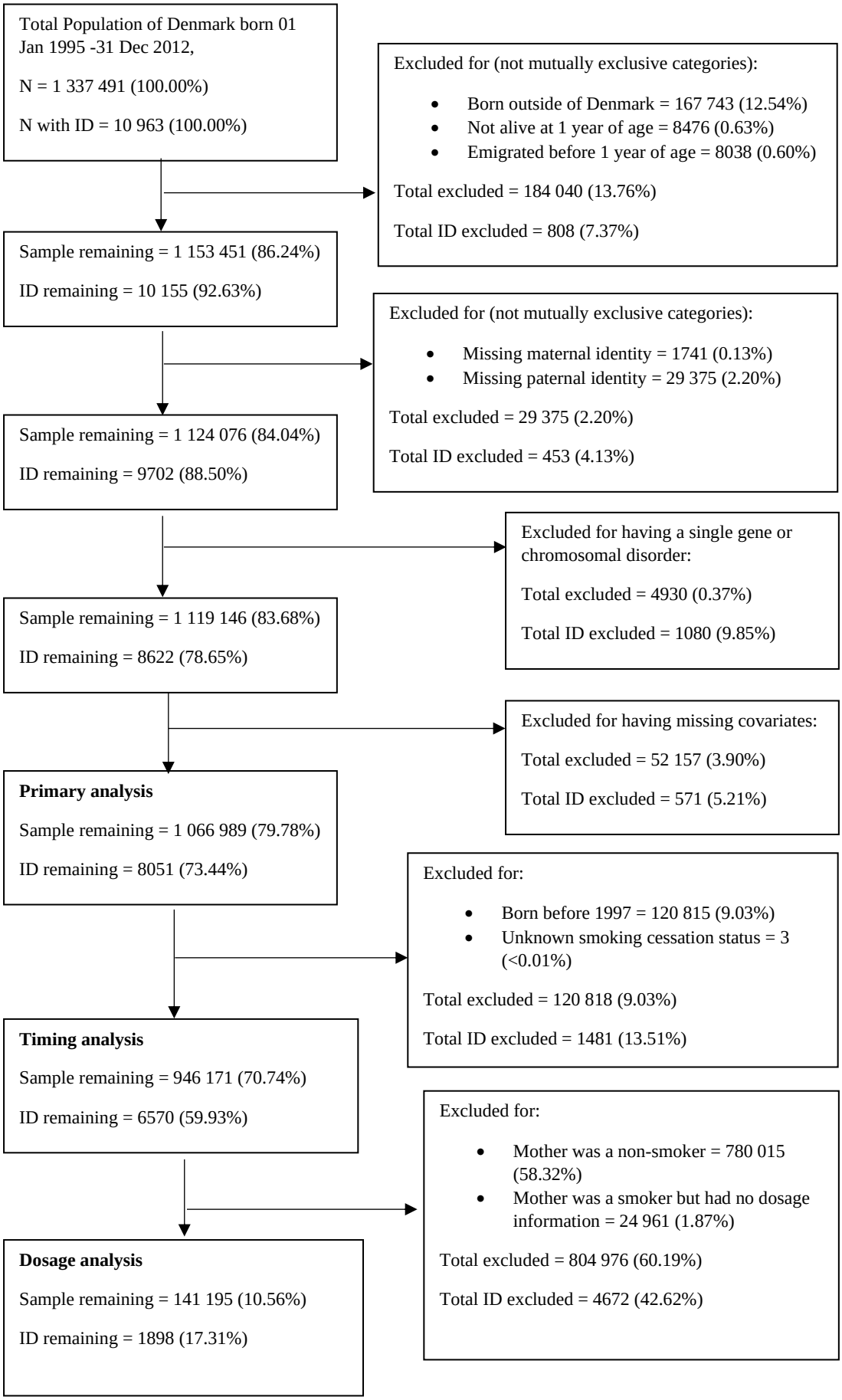

Figure 5-1: Flowchart of cohort derivation.

5.1.3 – Exposure definition: maternal smoking during pregnancy

Information about maternal smoking during pregnancy was obtained from the MBR, abstracted from midwife interviews at the first antenatal contact. A dichotomous smoking variable (yes/no) is available for births between 1995 and 1996. From 1997 additional information on duration (i.e. whether they stopped smoking and whether smoking cessation was before or after the first trimester) and number of cigarettes smoked per day (up to 5 cigarettes, 6-10 cigarettes, 11-20 cigarettes, >20 cigarettes) was added. Reporting of the additional duration and dosage data did not occur until late 1997 and took a few years to reach >95% completeness [255].

I created a dichotomous smoking variable that indicated smoking at any time during pregnancy. The validity of the MBR smoking measure is supported by correlations between the MBR maternal smoking data and biomarkers of smoking-related methylation in new-born offspring [260]. I created a categorical timing variable among those born from 1997 onwards with available data (n=946,171) that indicated whether mothers did not smoke during pregnancy, smoked but gave up before the end of the first trimester or smoked beyond the end of the first trimester. Finally, I created a continuous dosage variable, where data were available (n=141,195), that indicated the number of cigarettes smoked per day using the lower bound of the dosage groups (i.e. 1, 6, 11 or 21 cigarettes smoked per day).

I assessed the reliability of the dichotomous smoking measure by comparing smoking status during pregnancy in the MBR against the NPR. Where data were available in the NPR, over 99% of smokers identified in the NPR were also recorded as smokers in the MBR, while 0.01% of non-smokers in the NPR were classified as smokers in the MBR (see Table 5-1). Smokers, according to the MBR, were more likely to have missing data in the NPR than were non-smokers (OR=1.82; 95% CI=1.80-1.84).

Table 5-1: Concordance of exposure in Medical Birth Register against National Patient Register

	NPR value		
MBR value	Smokers, N(%)	Non-smokers, N(%)	Missing value, N(%)
Smoker	142737 (99.99)	969 (0.13)	54671 (26.68)
Non-smoker	14 (0.01)	718327 (99.87)	150271 (73.32)

5.1.4 – Outcome definition: intellectual disability

ID was defined as having an ICD-10 [12] code of F70-F79 (F70: mild ID, F71: moderate ID, F72: severe ID, F73: profound ID, F78 other ID, F79: unspecified ID), recorded as a primary or

secondary diagnosis in either the DPR or the NPR. After exclusions (see Figure 5-1), 8051 cases (0.75% of included persons during the included follow-up period) were identified.

5.1.5 – Comorbid autism spectrum disorder and attention deficit hyperactivity disorder definition

Matching the definition used by the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) consortium [261], individuals with the ICD-10 diagnosis codes F84.0, F84.1, F84.5, F84.8 and F84.9 were identified as having ASD. Individuals with the diagnosis codes F90.0 were identified as having ADHD. Where iPSYCH used diagnoses from only the DPR, I also used diagnoses from the NPR.

5.1.6 – Covariate and confounder definitions

The covariates and confounders adjusted for in statistical models were child sex, parity, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin. Full details of variable definitions can be found below.

Parental age at the time of the child's birth was derived from the parent's date of birth contained in the CRS. Highest educational attainment of either parent was obtained from Danish education registers and derived into a categorical variable separated into primary education (6-16 years old), general/vocational education (post 16 education) and higher education (university level of any duration). Parental country of origin was grouped according to the following locations: Denmark (including Greenland), Africa, Americas (North and South), Europe, Middle East, Oceania (Asia and Australia) and Scandinavia.

Parental income was obtained from the Statistics Denmark registry of income [259]. I used a variable for total personal income, excluding calculated rental value of own housing and before interest deduction [262]. From this variable I took the mean value of both parent's income in the year of the child's birth. If income data was missing for one parent, then the value was taken as the non-missing parent's income. I then divided the mean income by the number of children in the family at the pregnancy's conclusion (e.g. if 1 child was already in the family and this pregnancy was with twins then the mean income was divided by 3) to create a child adjusted mean income. For 82 children with unknown multiplicity I assumed the children to be singletons. I then created deciles of the child adjusted mean income among all families with non-missing covariate data (i.e. the final sample) for each year.

For parental psychiatric history I derived indicator variables for diagnoses of affective disorders, anxiety disorders, psychotic disorders and substance use disorders (excluding nicotine related disorders; ICD-10 code F17) in either the DPR or the NPR for each parent at any time before the child's birth. As the diagnostic system used in Denmark changed from ICD-8 to ICD-10 in 1994 I used the conversion table presented by Pederson et al. [263] to convert between the two classification systems. The diagnosis codes used to derive the indicator variables are presented in Table C.1.2-1 in Appendix C.

Data on maternal and paternal identity were obtained from the MBR to ensure that the biological mother is identified; using the maternal identifier from other sources may lead to adoptive parents being identified, for example, which would create error in the potentially confounding variables. For the paternal identifier, this is who the biological mother says is the father of the child and may not represent the biological father of the child. It may be another woman in the case of same sex couples. I do not believe that misidentification of the biological father will be an issue within this study. Instead, provided that the paternal identifier identifies the partner of the biological mother during her pregnancy, this should capture information relevant to the environment of the developing foetus, and hence the identified confounding structures.

5.1.7 – Assessment of missing data

There were little missing data (overall 3.9%, 52 157 individuals) due to missing exposure, confounder or covariate variables (n remaining=1,066,989 individuals from 658,335 families). I assessed whether any measured variable predicted whether a value would be excluded from the cohort using complete case logistic regression of a binary indicator for exclusion on each variable separately with no adjustment.

5.1.8 – Statistical analysis

All analyses were performed using R version 3.4.3 [264]. Following descriptive analyses, the primary analyses involved logistic regression of ID on maternal smoking in pregnancy. The family structure present within the cohort means that the data violates the assumption of independence between observations which can lead to underestimation of standard errors. I therefore used generalised estimating equations (GEE)[265], with an exchangeable correlation structure for mother and father combinations. This means the analyses accounted for correlations between full siblings, but half siblings, cousins and other relations were treated as independent. All models (including those referred to as unadjusted) were adjusted for child's

grouped year of birth (1995-1997, 1998-2000, 2001-2003, 2004-2006, 2007-2009, 2010-2012) to account for cohort effects and the differing length of follow up across birth years.

5.1.8.1 – *Primary analysis*

I ran the following four models (ignoring the term for child's grouped year of birth):

Model 1 - $g\left(\mathrm{E}\left[Y_j \middle| X_j\right]\right) = \beta_0 + \beta_P X_j,$

Model 2 - $g\left(\mathrm{E}\left[Y_j \middle| X_j, C_j\right]\right) = \beta_0 + \beta_P X_j + \beta_C C_j,$

Model 3 - $g\left(\mathrm{E}\left[Y_j \middle| X_j, \bar{X}\right]\right) = \beta_0 + \beta_W X_j + \beta_B \bar{X},$

Model 4 - $g\left(\mathrm{E}\left[Y_j \middle| X_j, \bar{X}, \mathrm{C_j}\right]\right) = \beta_0 + \beta_W X_j + \beta_B \bar{X} + \beta_C C_j,$

Model 1 did not adjust for any variables (other than grouped year of birth). Model 2 was adjusted for covariates and confounders. Model 3 adjusted for family-averaged smoking by including a term equal to the proportion of pregnancies in the family in which the mother was recorded as having smoked, thus making use of model formulation 2 suggested by Begg and Parides [102], but without other covariates. Model 4 adjusted for all covariates, confounders and the family-averaged smoking variable. Model 1 and 2 therefore provide population averaged effect estimates (β_P) while Model 3 and 4 provide both within-family (β_W) and between-family effect estimates (β_B; see Section 2.3.3 for an explanation of these terms).

5.1.8.2 – *Positive control analysis*

To test the validity of this approach I performed a positive control analysis in which I repeated the analyses using birthweight instead of ID, an outcome that is well established as having a causal relationship with maternal smoking in pregnancy. I repeated the four primary analysis models using low birthweight as the outcome, defined as a birthweight of less than 2500g (4.73% of included persons). Birthweight (mean value=3948g, SD=590g) was obtained from the MBR for 1,062,474 individuals (99.6% of the primary analysis sample).

5.1.8.3 – *Secondary analysis*

In secondary analyses I assessed the association between maternal smoking and offspring ID for different severities of ID and comorbidities of ID with ADHD and ASD. I also assessed differences in effect size based on sex, smoking timing and dosage.

To assess whether there was a pattern of increasing effect sizes across the severity of diagnosed ID, I performed an unadjusted multinomial logistic GEE model of a categorical variable of highest severity of ID recorded (6 levels: F70, F71, F72, F73, F78/F79, no ID) on maternal

smoking during pregnancy with no ID as the reference category. Adjusted analyses could not be performed due to model convergence issues.

To assess whether comorbidity of ID with ADHD and ASD influenced the association between smoking in pregnancy and ID, I performed unadjusted and confounder adjusted analyses using a multinomial logistic GEE model of a categorical variable of comorbidity of ID and ASD (4 levels: ID only, ASD only, ID and ASD, No ID or ASD) on maternal smoking during pregnancy, using no ID or ASD as the reference category. This was repeated for comorbidity of ID with ADHD.

To assess whether the associations between smoking in pregnancy and offspring ID were influenced by offspring sex I tested an interaction term between maternal smoking and offspring sex. I performed these analyses unadjusted and adjusted for covariates and confounders.

To assess the impact of timing of maternal smoking exposure, I performed unadjusted and confounder adjusted logistic GEE models of ID on the categorical smoking variable of whether the mother gave up smoking during the first trimester of pregnancy. To assess the impact of number of cigarettes (dose), I repeated the four primary analyses models among those who continued to smoke throughout pregnancy and had dosage data available using the number of cigarettes smoked per day as the independent variable. A family-averaged dosage variable was created by taking the mean number of cigarettes smoked across each pregnancy a mother had for which dosage data were available.

5.1.8.4 – *Sensitivity analysis*

In sensitivity analyses I assessed whether results were robust to: (i) measurement error in the outcome; (ii) collinearity between the individual and family-averaged exposures; (iii) differing lengths of follow up between cohort years; and (iv) potential biases arising from smoking patterns in the cohort.

I re-ran the primary analyses using a more stringent criterium for ID, defined as having at least two F70-F79 diagnoses recorded in either the DPR or the NPR. A total of 4,452 cases (prevalence = 0.42%) were identified using this definition. I did this in order to examine whether the results were robust to measurement error in the outcome.

I tested whether my conclusions were influenced by collinearity between the individual and family-averaged exposures. Strong polychoric correlation existed between the individual and family-averaged exposures ($\hat{\rho}$=0.996). I used an individual level exposure centred on the family-averaged exposure (corresponding to Begg and Parides' model 3 [102]) to induce orthogonality between the individual and family-averaged exposures ($\hat{\rho}$=-0.008). Models 3 and 4 were

repeated using this new exposure variable. This sensitivity analyses also serves the purpose of showing that the within-family effect estimate is the same whether the exposure is entered into the model uncentred (as X_j) or centred on the family-averaged exposure (as $X_j - \bar{X}$; see the description in Section 2.3.3)

I assessed whether differing lengths of follow up between cohort years influenced my conclusions as younger individuals at the end of follow up may have had less opportunity to be diagnosed with ID. I therefore repeated the four primary analysis models among sub-cohorts that were grouped by birthyear. I also repeated the four primary analysis models using Cox regression in the survival R package [266, 267]. The Cox models assessed the influence of maternal smoking in pregnancy (binary) on the time in days from birth to the first diagnosis of ID (defined as the first recorded start date of contact with services with a diagnostic code for ID). Individuals who died or emigrated were censored at the date of death or first emigration. Robust standard errors were used for family clusters.

To check for biases arising from smoking patterns in the cohort (see below) I performed additional sensitivity analyses in which I repeated the primary analyses using three restricted cohorts. These cohorts were (i) a cohort of single-child families only (n=175,043; n with ID=1,533), (ii) a cohort of multiple-child families only (n=891,946; n with ID=6,517) and (iii) a cohort of multiple-child families in which all children were born after the start of the cohort (n=801,109; n with ID=5,007).

5.2 – Results

5.2.1 – Description of the cohort

Characteristics of the study cohort are displayed in Table 5-2. Maternal smoking was reported in 18.6% of pregnancies and was associated with lower maternal and paternal age at pregnancy, lower parental education, being in a lower decile of income, and increased parity. All psychiatric disorders were more common in smokers and their partners compared to families in which the mother did not smoke during pregnancy. The prevalence of maternal smoking during pregnancy decreased over time.

Table 5-2: Characteristics of the sample by maternal smoking during pregnancy (exposure) status.

Characteristic	Smokers	Non-smokers	p-value [a]
Total, N (%)	198 377 (18.6)	868 612 (81.4)	
Maternal age, mean (SD)	28.7 (5.26)	30.1 (4.64)	<.001
Paternal age, mean (SD)	31.4 (6.16)	32.7 (5.58)	<.001
Highest parental education, N (%)			<.001
- Primary	46 456 (23.4)	60 817 (7.0)	
- General/vocational	110 361 (55.6)	351 798 (40.5)	
- Higher	41 560 (21.0)	455 997 (52.5)	
Income decile, median (IQR)	4 (1-6)	5 (2-7)	<.001
Maternal country of origin, N (%)			<.001
- Denmark	181 453 (91.5)	746 737 (86.0)	
- Africa	844 (0.4)	16 286 (1.9)	
- Americas	700 (0.4)	5313 (0.6)	
- Europe	8474 (4.3)	37 821 (4.4)	
- Middle East	2000 (1.0)	23 403 (2.7)	
- Oceana	2045 (1.0)	27 073 (3.1)	
- Scandinavia	2861 (1.4)	11 979 (1.4)	
Paternal country of origin, N (%)			<.001
- Denmark	180 019 (90.7)	751 091 (86.5)	
- Africa	1339 (0.7)	17 641 (2.0)	
- Americas	728 (0.4)	4898 (0.6)	
- Europe	10 165 (5.1)	39 870 (4.6)	
- Middle East	2989 (1.5)	27 359 (3.1)	
- Oceana	1162 (0.6)	18 884 (2.2)	
- Scandinavia	1975 (1.0)	8869 (1.0)	
Maternal Psychiatric history, N (%)			
- Affective disorder	5398 (2.7)	12 945 (1.5)	<.001
- Anxiety disorder	12 527 (6.3)	27 332 (3.1)	<.001
- Psychotic disorder	1953 (1.0)	3119 (0.4)	<.001
- Substance use disorder	8433 (4.3)	9587 (1.1)	<.001
Paternal Psychiatric history, N (%)			
- Affective disorder	1868 (0.9)	4840 (0.6)	<.001
- Anxiety disorder	5245 (2.6)	12 158 (1.4)	<.001
- Psychotic disorder	1440 (0.7)	3537 (0.4)	<.001
- Substance use disorder	9554 (4.8)	15 985 (1.8)	<.001
Child sex, N (%)			.22
- Female	96 406 (48.6)	423 450 (48.8)	
- Male	101 971 (51.4)	445 162 (51.2)	
Parity, N (%)			<.001
- 0	85 111 (42.9)	376 648 (43.4)	
- 1	69 489 (35.0)	331 010 (38.1)	
- 2	31 007 (15.6)	121 489 (14.0)	
- 3+	12 770 (6.4)	39 465 (4.5)	
Cohort year, N (%)			<.001
- 1995-1997	47 205 (23.8)	133 189 (15.3)	
- 1998-2000	42 271 (21.3)	139 861 (16.1)	

Characteristic	Smokers	Non-smokers	p-value [a]
- 2001-2003	34 592 (17.4)	143 239 (16.5)	
- 2004-2006	29 781 (15.0)	150 893 (17.4)	
- 2007-2009	24 601 (12.4)	153 931 (17.7)	
- 2010-2012	19 927 (10.0)	147 499 (17.0)	

[a] t-tests were performed for normally distributed continuous variables, Wilcoxon rank sum tests were performed for non-normally distributed continuous variables, and χ^2 tests were performed for binary/categorical variables.

5.2.1.1 – *Smoking patterns in the cohort*

As parity increased, changes in smoking from one pregnancy to the next were less likely to occur (see Table C.2.1-1 in Appendix C). For children with parity>1, an index child was less likely to have ID if their mother had smoked in neither the index pregnancy or the previous pregnancy and was more likely to have ID if their mother had smoked in either pregnancy (see Table C.2.1-2).

I present a cross tabulation of the family-averaged exposure variable against case status for ID in Table C.2.1-3. Exposure discordance within a family (i.e. a family-averaged exposure value not equal to 0 or 1) was present for 66,798 individuals (6.3% of the primary analysis sample), across 28,748 different families (4.4% of all families in the primary analysis sample).

207,121 families (31.46% of all families in the primary analysis sample) had a child born before the cohort start date. Information on such children was not observed and so not included in the family-averaged variable; the family-averaged variable in such families may contain error as a result. Families with a child excluded for being born before the cohort start date were less likely to have exposure discordance and were more likely to have all siblings exposed than families in which all children were born after the cohort start date (see Table C.2.1-4).

Only-children were more likely to be exposed to smoking in pregnancy than the first-born child of families with multiple children in the cohort (see Table C.2.1-5). This pattern held true when stratifying separately on birthyear and maternal age indicating that single and multiple-child families may not be comparable.

5.2.1.2 – *Patterns of intellectual disability over time*

By taking a snapshot of prevalence by age in 2017, Figure 5-2 shows that the prevalence increased with age up to an age of 18 where it started to level off. Prevalence was higher among males than females at all ages. The prevalence of ID at ages 6-18 years of age for each cohort year is displayed in Figure 5-3. Prevalence remained approximately steady at each age over time, though some cohort years (e.g. 1999) have slightly greater prevalence at all ages.

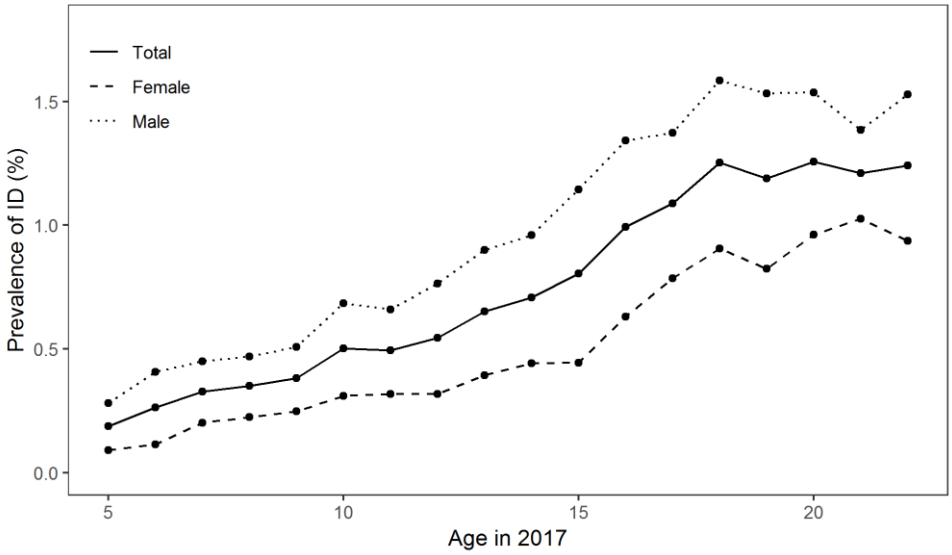

Figure 5-2: Plot of the prevalence of intellectual disability for each age in 2017.

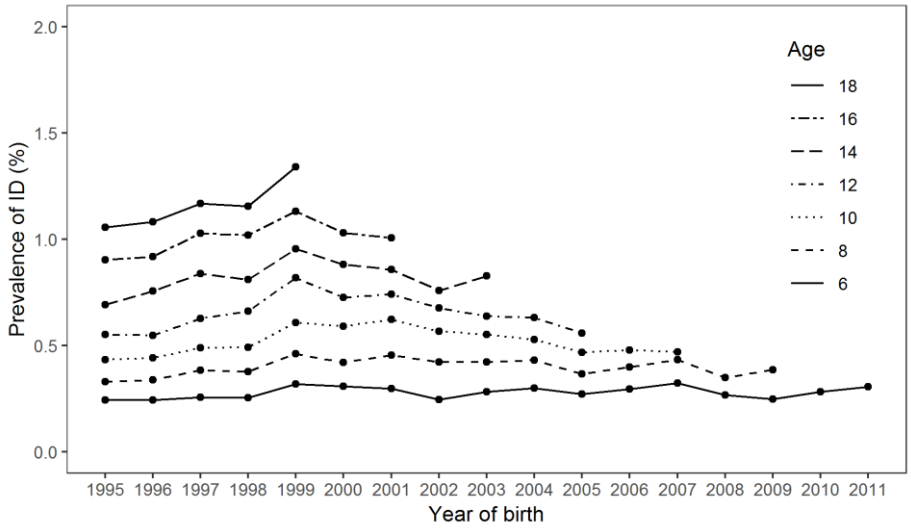

Figure 5-3: Plot of the prevalence of intellectual disability at ages 6-18 for each cohort year.

5.2.2 – Assessment of missing data

Results of the missing data assessment can be viewed in Table C.2.2-1 (Appendix C). Individuals with a diagnosis of intellectual disability had increased odds (OR=1.46; 95% CI=1.34-1.59) of being excluded for having missing data. Individuals whose mother smoked

during pregnancy had lower odds (OR=0.67; 95% CI=0.64-0.70) of exclusion for missing data in other covariates compared to those whose mothers did not smoke. Missingness was driven largely by missing maternal smoking data; 74% of those excluded for having missing data (3.5% of the total sample) had missing exposure information (see Table C.2.2-2). In Table C.2.2-3 and C.2.2-4 I present the distribution of missing data in the smoking variable across cohort year group and parity. These tables show that most of the missing data in the smoking variable was in the years 1995-2000 and that there was slightly more missing data for children with greater parity.

5.2.3 – Primary analyses of the association between maternal smoking and offspring ID

Maternal smoking during pregnancy was associated with increased odds of ID in unadjusted analysis (see Table 5-3; OR=1.91, 95% CI=1.82, 2.00). This was attenuated following adjustment for covariates and confounders (OR=1.35, 95% CI=1.28, 1.42). The within family effect, obtained from the model adjusted for the family-averaged smoking variable, was found to be null before and after adjustment for confounders; before (OR=0.91, 95% CI=0.78, 1.06), after (OR=0.93, 95% CI=0.79, 1.09). The between-family effect showed increased odds of ID before and after confounder adjustment.

Table 5-3: Primary analysis of the association between maternal smoking and offspring intellectual disability.

Model	Coefficient	O.R.	95% CI
Unadjusted	- Population-averaged	1.91	1.82, 2.00
Adjusted for confounders [a]	- Population-averaged	1.35	1.28, 1.42
Adjusted for family smoking variable	- Within-family	0.91	0.78, 1.06
	- Between-family	2.25	1.92, 2.63
Adjusted for confounders [a] and family smoking variable	- Within-family	0.93	0.79, 1.09
	- Between-family	1.51	1.28, 1.79

[a] Adjusted for child sex, parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

5.2.4 – Positive control analyses of the association between maternal smoking and offspring low birth weight

In the positive control analyses (see Table 5-4) I found that maternal smoking in pregnancy was associated with increased odds of low offspring birthweight that was slightly attenuated after adjustment for confounders. Both the within-family effect and between-family effect showed notable attenuation of the association between maternal smoking and low birthweight association, although all estimates remained consistent with increased odds of low birthweight before and after confounder adjustment.

Table 5-4: Positive control analysis of the association between maternal smoking and offspring low birthweight.

Model	Coefficient	O.R.	95% CI
Unadjusted	- Population-averaged	1.88	1.85, 1.92
Adjusted for confounders [a]	- Population-averaged	1.74	1.70, 1.77
Adjusted for family smoking variable	- Within-family	1.21	1.14, 1.27
	- Between-family	1.64	1.54, 1.74
Adjusted for confounders [a] and family smoking variable	- Within-family	1.06	1.00, 1.13
	- Between-family	1.73	1.61, 1.85

[a] Adjusted for child sex, parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

5.2.5 – Secondary analyses

Descriptives of the prevalence of ID across the characteristics investigated in secondary analyses are presented in Table 5-5. The highest severity diagnosis received by cohort members was most often mild or unspecified ID (F70 and F78/9). There was a higher prevalence of ID among those with ASD or ADHD than those without. The prevalence of ID was higher among those whose mother continued to smoke after the 1st trimester than those whose mothers quit during the first trimester or who did not smoke. A pattern of increasing prevalence of ID with greater numbers of cigarettes smoked per day during pregnancy was also observed.

Tables for all secondary analyses are presented in Section C.2.3 of Appendix C. In unadjusted analyses maternal smoking during pregnancy was associated with mild, moderate and unspecified ID (Table C.2.3-1). The odds for the association with mild ID were greater than those for moderate ID. Analyses for comorbidity with ASD and ADHD are presented in Table C.2.3-2 and Table C.2.3-3 respectively. In unadjusted analyses, all combinations of ID with ASD and combinations of ID with ADHD showed an association of increased odds among

those exposed to maternal smoking in pregnancy. Following adjustment for confounders and family averaged exposure, the within-family effects for all comorbidity combinations were null.

I found no evidence for an interaction between smoking in pregnancy and offspring sex which suggests that there is no difference in effect size between male and female offspring. (Table C.2.3-4). After adjustment for confounders, stopping smoking in the first trimester did not differ from not smoking during pregnancy in terms of odds of ID, whereas continuing smoking after the first trimester showed increased odds of ID (Table C.2.3-5). Dosage analyses showed a 5% increase in odds of ID for each additional cigarette smoked per day (Table C.2.3-6). This effect was attenuated to a 3% increase following adjustment for confounders. A null within-family effect was found in all models that adjusted for the family-averaged dose. The between-family effect was a 6% increase in odds per cigarette smoked per day on average across pregnancies, holding fixed the number smoked in each individual pregnancy. This attenuated to a 3% increase after adjustment for confounders.

5.2.6 – Sensitivity analyses

Sensitivity analyses (tables presented in Section C.2.4 of in Appendix C) showed that the results were not substantially influenced by using a stricter definition of the outcome variable, by strong correlation between the individual-level and family-averaged exposure variables or by differing lengths of follow up between cohorts.

Replication of the primary analyses with a stricter ID outcome definition is presented in Table C.2.4-1. The results closely resemble those of the primary analyses though ORs were slightly smaller, and CIs were slightly wider.

Assessment of whether collinearity of the individual and family-averaged exposure variables influenced results is presented in Table C.2.4-2. The within-family effect estimate of the exposure centred on the family averaged exposure is consistent with that from the primary analyses in which the exposure is not centred.

I assessed whether length of follow up influenced the results of analyses. The four primary analyses models, repeated for each cohort year group, show that in all year groups except 1995-1997 and 2010-2012 there was a null within-family effect and a between-family effect that displayed increased odds of ID with increasing family averaged smoking (see Figure 5-4; values displayed in the figure are presented in Table C.2.4-3). In the 1995-1997 group there was a null between-family effect while in the 2010-2012 group the standard errors of both effects were large, suggesting low power to detect an effect. The Cox proportional hazards models provide results that are consistent with those obtained from the logistic GEE models (see Table C.2.4-4).

The results of analyses in a sample restricted to single-child families showed lower adjusted ORs for smoking in pregnancy compared to the primary analyses (see Table C.2.4-5). In comparison, when the sample was restricted to multiple child families, results were comparable to those of the primary analyses.

Table 5-5: Distribution of ID across secondary analysis categories

Category	ID, N(%)	No ID, N(%)
Severity analysis [a]		
- F70	4,436 (55.10)	-
- F71	1,288 (16.00)	-
- F72	451 (5.60)	-
- F73	176 (2.19)	-
- F78/9	1,700 (21.11)	-
Comorbidity analysis [b]		
- ASD		
- Yes	2,840 (13.58)	18,075 (86.42)
- No	5,211 (0.50)	1,040,863 (99.50)
- ADHD		
- Yes	2,033 (9.87)	18,565 (90.13)
- No	6,018 (0.58)	1,040,373 (99.42)
Sex difference analysis [b]		
- Females	2,680 (0.52)	517,176 (99.48)
- Males	5,371 (0.98)	541,762 (99.02)
Timing analysis [b]		
- Non-smoker	4,469 (0.57)	775,546 (99.43)
- Stopped smoking during the first trimester	115 (0.67)	16,972 (99.33)
- Continued smoking after the first trimester	1,986 (1.33)	147,083 (98.67)
Dosage analysis [b]		
- Smoked up to 5 cigarettes per day	399 (0.92)	43,169 (99.08)
- Smoked 6-10 cigarettes per day	713 (1.35)	52,263 (98.65)
- Smoked 11-20 cigarettes per day	642 (1.63)	38,731 (98.37)
- Smoked more than 20 cigarettes per day	144 (2.73)	5,134 (97.27)

[a] Column percentage
[b] Row percentage

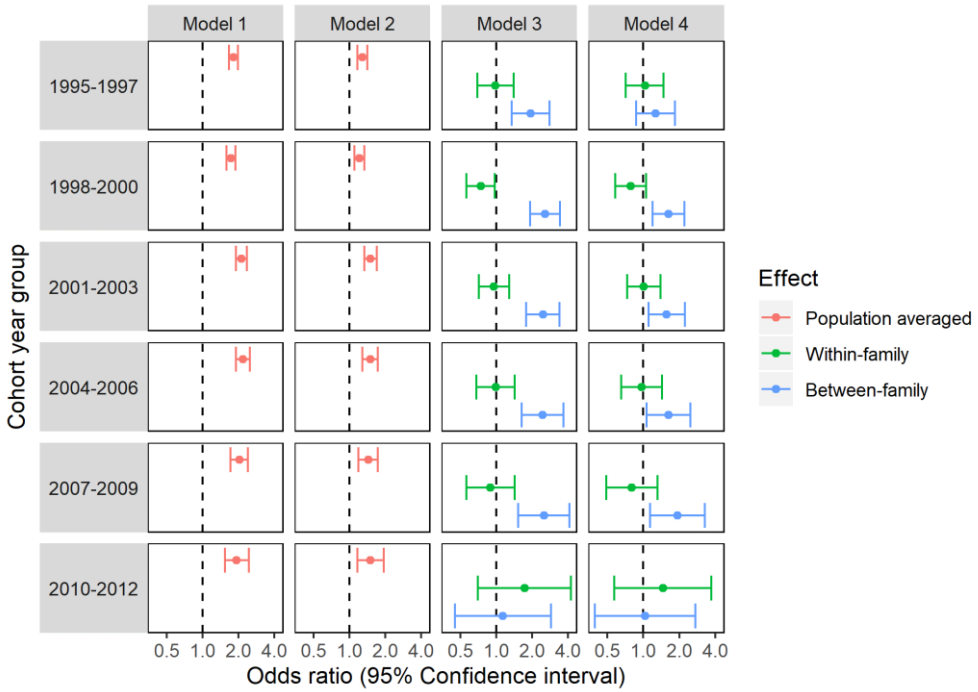

Figure 5-4: Logistic GEE analyses of the association between maternal smoking during pregnancy and offspring ID repeated in each year group category

5.3 – Discussion

Using a large population-based cohort I have provided evidence that the association between maternal smoking during pregnancy and offspring ID is not consistent with a causal relationship. Instead the association appears to be driven by residual confounding. The primary analyses showed that when population averaged associations were decomposed into within-family and between-family effects there was no influence of individual exposure to smoking in pregnancy on the risk of ID. Associations were instead driven by family-level differences in maternal smoking behaviour.

The consistency of results between primary and sensitivity analyses provide evidence that conclusions were robust to (i) measurement error in the outcome variable (ii) collinearity between the individual and family- averaged exposure variables, (iii) differing lengths of follow up between cohort year groups and (iv) potential biases arising from patterns of smoking in the cohort. I validated my analysis approach by performing a positive control analysis in which low birthweight was used as an outcome. Here a causal relationship was expected [8, 9]. I found a small within-family effect suggesting that once family-level differences in exposure were accounted for there was a small increase in the risk of low birthweight for those exposed to smoking in pregnancy.

In secondary analyses I found unadjusted associations with milder and unspecified forms of ID but not more severe forms which may suggest that milder ID is more susceptible to environmental exposures than more severe ID. I also found an effect of exposure timing that could suggest a sensitive period of exposure. Both results need to be treated with caution. Severity may have been measured with substantial error as the unspecified severity group made up the second largest of all groups. Every case in this unspecified diagnosis group belongs to one of the specified diagnosis categories, however, the proportion in each is unclear. Each association estimate in the severity analysis may therefore be biased to an unknown extent. Regarding the timing analysis, I am not aware of a parameterization that would allow the separation of effects into within and between-family effects for a categorical variable and so was not able to control for family-level differences in smoking cessation during pregnancy. The effect of exposure timing may therefore simply reflect familial confounding. Further, smoking cessation during pregnancy is commonly misreported [268] and may be attempted multiple times. The measure of smoking cessation may therefore not adequately capture the timing of exposure during pregnancy.

It is likely that associations with maternal smoking in pregnancy reported in prior studies were due to residual confounding. Braun et al. [98] found strong attenuation of their confounder adjusted association following further adjustment for area level socioeconomic information obtained from data linkage. Comparison with this study suggests that linked data is not always sufficient to account for confounding structures. Further accounting for family structure as in the analyses that estimated a within-family effect, and analyses performed by Lundberg et al. [99], appear to consistently demonstrate that associations between maternal smoking in pregnancy and offspring IQ/ID are unlikely to be causal.

Study strengths included the large population-based sample size which reduced the risk of selection bias and improved generalisability. Extensive data linkage allowed for adjustment for many confounding variables and for the derivation of family- averaged exposure variables which allowed for understanding of residual confounding.

This study has several limitations. I used registry data, and some misclassification in the recording of exposures and outcomes cannot be ruled out. This would be a limitation in any large-scale record linkage study. I had information on a range of potential confounding factors although I was unable to study the role of some potentially relevant factors such as gestational diet quality, alcohol or substance misuse during pregnancy or the role of passive smoke exposure to the mother during pregnancy, or to the child following birth. Although there were little missing data (3.9%), since those excluded were also more likely to have ID, complete case analysis may be biased towards the null compared to a fully observed dataset. Finally, I note

that the reported between-family effect is likely to be imprecise due to the small cluster size of families [102].

The between-family effect showed increased odds of ID for families in which the mother tended to smoke in more pregnancies, holding fixed individual level exposure to smoking. Exploring the between-family effect by investigating correlates of the family-averaged smoking variable further may inform more about the residual confounding structure and what additional variables need to be adjusted for in studies of maternal smoking in pregnancy. Instrumental variable approaches such as Mendelian Randomisation [166, 269, 270] may be informative in exploration of the residual confounding structure as they would specifically assess the influence of the mother's predisposition to smoking behaviours. This is likely to be comparable to the family-averaged smoking variable. The family-averaged exposure variable may, however, reflect genetic confounding in that a genetic propensity for maternal smoking may be associated with offspring ID via pleiotropic mechanisms rather than via maternal smoking in pregnancy. For example, polygenic risk scores (PRS) for ADHD, which are likely to correlate with genetic risk of ID due to the high prevalence of comorbidity, have been found to predict smoking behaviour [271, 272]. In this case standard Mendelian Randomisation would not be appropriate due to violation of the exclusion restriction criteria (the assumption that an instrument is associated with the outcome only via the exposure) [171] and extensions such as multivariable Mendelian Randomisation would be required [273].

5.4 – Conclusions and chapter summary

Based on the consistent findings of no association between maternal smoking and ID across the primary, secondary and sensitivity analyses in family-based analytic approaches, this study provides evidence against a causal effect of maternal smoking during pregnancy on offspring intellectual disability. The persistent between-family effect in the absence of a within-family effect in adjusted analyses provides evidence in support of the role of residual confounding. A lack of causal effect of maternal smoking in pregnancy on offspring ID should not be interpreted as meaning that smoking in pregnancy is safe. It has a range of other demonstrable negative health consequences and these results should not distract from the sustained efforts requires to reduce its prevalence.

Chapter 6 Maternal smoking and smokeless tobacco use during pregnancy and offspring development: sibling analysis in an intergenerational Swedish cohort

The contents of this chapter are currently being revised following peer review as a submission to the International Journal of Epidemiology

The Swedish registries, like the Danish registries analysed in Chapter 5, can be linked to make intergenerational cohorts that contain information on socioeconomic characteristics, pregnancy and birth characteristics and medical diagnoses throughout life. In addition to information on exposure to smoking during pregnancy, the Swedish registries also contain information on exposure to a form of moist smokeless tobacco known as snus. They therefore provide an opportunity to (i) replicate the findings of Chapter 5 in an independent dataset with a lower prevalence of smoking through the use of several causal inference techniques including sibling comparisons and a positive control comparison with fetal growth restriction, and (ii) compare the effect of smoking during pregnancy to an alternative, non-combustible method of nicotine consumption.

If a causal effect of smoking in pregnancy on intellectual disability (ID) does exist, then a cross context comparison between the associations of snus use and smoking in pregnancy with offspring ID can be used to investigate whether effects are the result of nicotine or the combustible components of cigarette smoke. Snus is a moist, smokeless tobacco that is increasingly being used as a smoking cessation aid in Sweden [184, 185], with some suggestion that it is more successful as an aid to stop smoking than nicotine patches or gum [186, 187]. Snus delivers nicotine in quantities that are comparable to cigarette smoke though with slower absorption and higher plasma nicotine concentration over an extended period [184, 188].

The association between snus use in pregnancy and offspring fetal growth restriction is of public health relevance itself as it may guide whether snus has potential to be a smoking alternative during pregnancy for those who have difficulties with cessation. Limited research has been performed on snus use in pregnancy though evidence of associations with preterm delivery [274, 275], offspring born small for gestational age (SGA) [276] and stillbirth [277, 278] have been suggested. Research into snus use in pregnancy and offspring fetal growth

restriction has provided mixed results. Two previous studies investigating snus use in pregnancy and risk of offspring born SGA provided no evidence for an association [278, 279] while another found evidence of increased odds of SGA for those who used snus before and early into pregnancy relative to no use of tobacco products [276]. None of these studies made use of causal inference methods and may be susceptible to the effects of unmeasured confounding.

In this chapter I aimed to (i) use conventional analyses to investigate whether maternal smoking in pregnancy is associated with offspring risk of intellectual disability, (ii) use sibling analyses to investigate whether such associations can be accounted for by characteristics shared between siblings, (iii) use positive control and cross context comparisons to learn more about the nature of the association, and (iv) investigate whether snus use in pregnancy influences fetal growth restriction.

6.1 – Methods

6.1.1 – Cohort definition

The study cohort consisted of all individuals born in Sweden between January 1st, 1999 and December 31st 2010 (n=1,181,264; see Figure 6-1). Information contained in national registries was linked to cohort members and their parents. The registries included the Swedish Medical Birth Registry (MBR) [280], the National Patient Registry (NPR) [281] and the Swedish Longitudinal Integration Database for Health Insurance and Labour Market Studies (LISA) [282]. Data on maternal and paternal identity was obtained from the Swedish Multi-Generation Register (MGR) [283].

Most clinical contacts related to intellectual disability occur in an outpatient setting. The NPR started recording outpatient contact in addition to inpatient admissions in 2001 [281]. By defining the start year of the cohort as 1999 I was able to capture snus use in pregnancy from its earliest recording in the MBR while also capturing diagnoses from 2 years of age onwards for the oldest members in the cohort and from an earlier age for all other cohort members. The end of 2010 was selected as the cut-off for inclusion in the cohort to allow a minimum of 4 years follow up until the end of 2014. The youngest and oldest members of the cohort were followed up until approximately 4 and 14 years of age respectively.

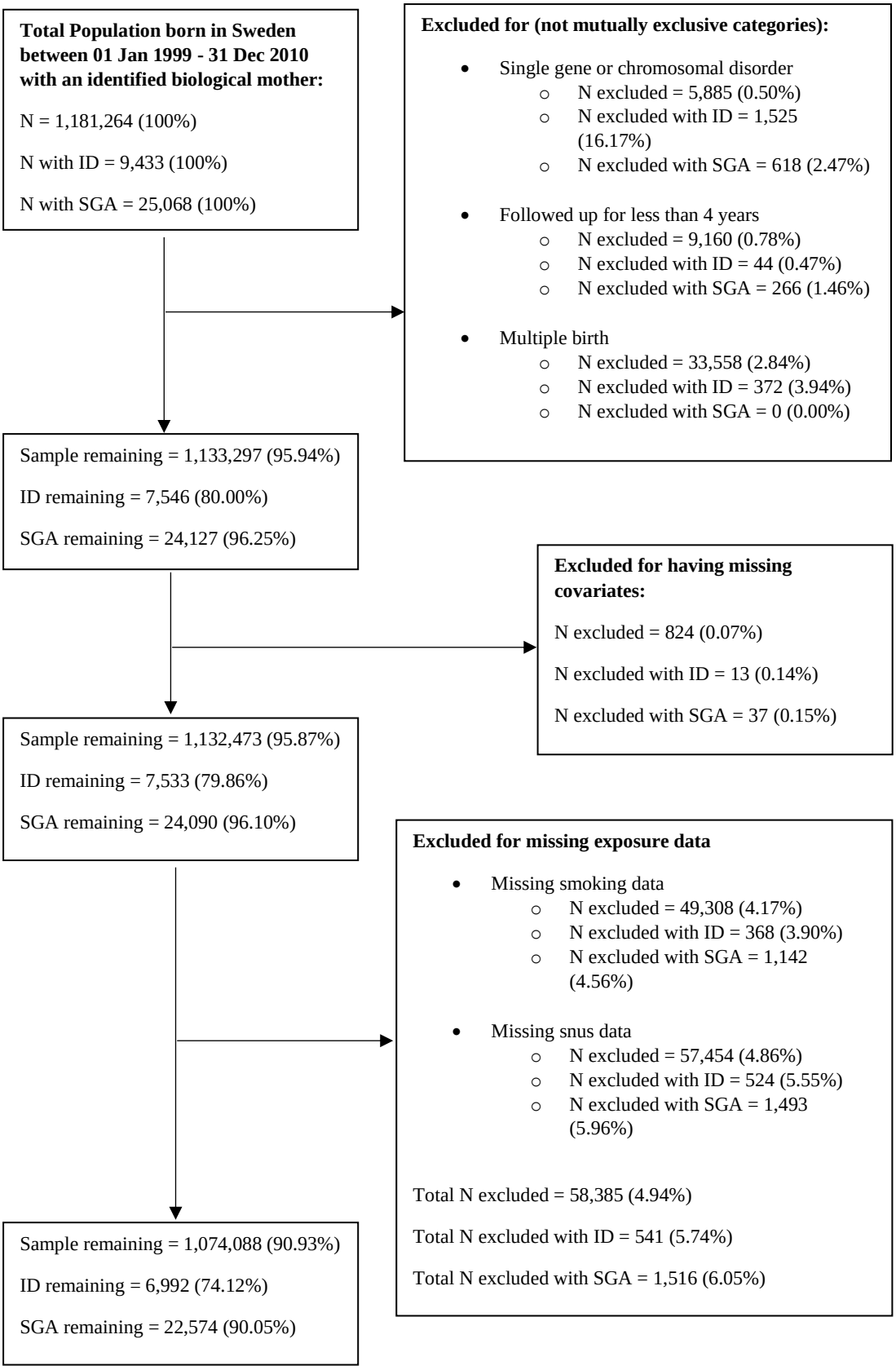

Figure 6-1: Flowchart of cohort derivation. Note: an additional 4147 excluded from SGA analyses due to missing SGA values. These were not removed as the unobserved values may be related to weight for gestational age which is on the hypothesised causal path from smoking to ID.

6.1.2 – Exclusion criteria

Individuals were excluded from the cohort if they had less than 4 years follow up (e.g. if the individual died before the age of 4 or had spent less than 4 years living in Sweden; n=5885), had a genetic or chromosomal abnormality associated with ID that was identified using ICD-10 diagnoses [12] in the NPR (specific diagnosis codes provided in section D1.1 of Appendix D; n=9160) or were part of a multiple birth pregnancy (i.e. twins or triplets etc.; n=33 558).

6.1.3 – Exposure definition: maternal smoking and snus use during pregnancy

Information about maternal smoking and snus use during pregnancy was obtained from the MBR for three time points: i) three months prior to pregnancy, ii) at the first antenatal contact (commonly around 10 weeks of pregnancy) and iii) at 30-32 weeks pregnancy. Missing smoking and snus values at each timepoint were imputed based on non-missing values at other time points in the same pregnancy. Details of the imputation procedure are detailed in Appendix D (section D.1.2).

Two binary variables were created, one for smoking and one for snus use which indicated maternal use at any point during pregnancy (i.e. either at first antenatal contact or at 30-32 weeks). Categorical variables for the timing of exposure were created to indicate those whose mothers' (i) never smoked/used snus, (ii) used only before pregnancy, (iii) quit during pregnancy or (iv) used throughout pregnancy.

Individuals were excluded from analyses if they were missing the binary smoking (n=49 308) or snus (n=57 454) variables. The total number excluded for missing exposure data was 58 385.

6.1.4 – Outcome definitions

6.1.4.1 – *Intellectual disability (ID)*

A binary indicator of ID was defined as having an ICD-10 code of F70-F79, recorded as a primary or secondary diagnosis in the NPR.

6.1.4.2 – *Fetal growth restriction*

Z-scores of birthweight for gestational age were obtained using the Swedish sex-specific reference curve for normal fetal growth [284]. The MBR contains a binary indicator for being SGA, defined as having a z-score value less than -2 (i.e. 2 standard deviations below the mean birthweight for a given gender and gestational age).

6.1.5 – Covariate and confounder definitions

The covariates and confounders adjusted for included child sex and parity, highest education level of either parent at the time of birth, quintiles of income adjusted for family size at the time of birth, any maternal or paternal psychiatric disorders before the birth of the child and maternal country of origin and age at birth. Individuals were excluded from analyses if they were missing data on any of the covariate or confounder variables (n=824). Specific details of the confounder derivations are described below.

6.1.5.1 – *Child sex and parity*

Child sex was defined as the biological sex assigned at birth. Parity was grouped into a categorical variable with 3 levels: 1, 2 and 3+.

6.1.5.2 – *Parental education*

Parental educational attainment was obtained from the Swedish LISA (Longitudinal integration database for health insurance and labour market studies) database [282]. Education was grouped into 3 levels: High school; Gymnasium (age 16-18, equivalent to A levels in the UK); University level. The highest level of either parent at the time of the child's birth was used.

6.1.5.3 – *Parental income*

Household income in the year of the child's birth was obtained from the Swedish LISA database [282]. The value was adjusted for family size and placed into quintiles for each year in order to account for inflation.

6.1.5.4 – *Parental psychiatric disorders*

Indicator variables were derived for diagnoses of anxiety disorders, depressive disorders, psychotic disorders and substance use disorders (excluding nicotine related disorders) in the NPR [281] at any time before the child's birth; see Table D.1.3-1 in Appendix D for the list of ICD-9 and ICD-10 codes used to define these disorders. Due to the low prevalence of each disorder in the cohort I combined the indicator variables into a single variable of any psychiatric disorder at any time before the child's birth.

6.1.5.5 – *Maternal country of origin*

Maternal country of origin was be obtained using the MGR [283]. The variable was categorised into the following levels: Sweden; Scandinavia (Denmark and Finland); Europe; Middle East; Americas (North and South America); Asia; Africa; Oceania.

6.1.5.6 – *Parental age at birth*

Parental age at the time of the child's birth was derived from the parent and child's date of birth and included as a continuous variable.

6.1.6 – Statistical analysis

6.1.6.1 – *Primary analyses*

I repeated the following analyses for the outcomes SGA and ID and the exposures maternal smoking in pregnancy and maternal snus use in pregnancy. For each exposure-outcome combination I used logistic regression. To account for cohort effects of differing lengths of follow up I adjusted for year of birth in all models using an outcome of ID, even those referred to as unadjusted.

I ran four models for each exposure-outcome combination. Model 1 was unadjusted for any covariates. Model 2 adjusted for covariates and confounders. Model 3 adjusted for family level smoking/snus use by including a term equal to the proportion of pregnancies in the cohort in which the mother was recorded as having smoked/used snus, thus making use of model formulation 2 suggested by Begg and Parides [102]. Model 4 adjusted for all covariates, confounders and the family level smoking/snus variable.

Adjustment for a family averaged exposure effect, as in model 3 and 4, allows the calculation of within-family (coefficient of the individual level exposure) and between-family (coefficient of the family level exposure) effects of smoking on child outcomes. The within-family effect is robust against confounders that are shared between the siblings. Failing to find a within-family effect after adjustment for the family-averaged exposure variable is consistent with familial confounding and there being no causal effect of the exposure on the outcome [102, 103].

The family structure present within the cohort meant that the data violates the assumption of independence between observations which can lead to underestimation of standard errors. All models were therefore run using generalised estimating equations (GEE) with exchangeable correlation structures for family groups identified by having the same mother. This means that the analyses accounted for correlations between siblings. Full and half siblings were treated equivalently. Cousins and other relations were treated as independent.

6.1.6.2 – *Secondary analyses*

I assessed whether timing of exposure was associated with the outcomes using four models, repeated for the outcomes ID and SGA. Model 1 and 2, performed unadjusted and adjusted for confounders respectively, were logistic regressions using GEE of the outcome on the categorical timing exposure. Model 3 and 4, again performed unadjusted and adjusted for confounders respectively, were conditional logistic regressions of the outcome on the categorical timing exposure, conditional on family grouping. Model 1 and 2 therefore provide population-averaged estimates while Model 3 and 4 provide within-family estimates.

6.1.6.3 – *Sensitivity analyses*

An assessment of potential biases in sibling designs highlighted that the exposure discordant group, which drives the within-family estimate, is more likely to contain exposure misclassification than the population as a whole [181]. This is because mothers are likely to behave similarly across pregnancies (see section B1 of Appendix B) and if a single sibling is misclassified then all members of that family will incorrectly become part of the exposure discordant group. As a result, the within-family estimate may be biased. To test this, I replaced the exposure status of those not exposed to maternal smoking in pregnancy (i.e. "not exposed" was changed to "exposed") if the mother smoked in later pregnancies. Here I have assumed that it is unlikely that a mother would start smoking in later pregnancies and therefore that the earlier born individuals may be misclassified; this pattern of misclassification has been hypothesised previously [285]. The family averaged exposure was recalculated using the new exposure values of each family member and the primary analyses were repeated. It should be noted that families with a sibling whose exposure was edited were no longer exposure discordant and therefore did not contribute to the within-family coefficient estimate. Further, the change in exposure status was more likely for children of mothers with lower parity than higher parity.

It is possible that differing patterns of change in smoking status may have differing influences on the within-family estimate. To test this, I performed a second sensitivity analysis in which I restricted the sample to include only the first two members of each family included in the cohort (single children were also included in the analysis as they contribute to the between-family estimate). Analyses were then run on this dataset as a whole (restricted cohort 1) and then repeated with further restriction of the exposure discordant group to only those in which the mother stopped smoking in the second pregnancy (restricted cohort 2). This was repeated again, restricting the exposure discordant group to those in which the mother started smoking in the second pregnancy (restricted cohort 3).

6.2 – Results

6.2.1 – Description of the cohort

Descriptive statistics of the cohort, separated by smoking status and by snus use, are presented in Table 6-1. Maternal smoking in pregnancy was more prevalent in the cohort than snus use in pregnancy (8.76% vs. 1.37%). Both smoking and snus use were socially patterned, however the strength of that social patterning was greater for smokers. Smokers were more likely to have low or mid-level education and low income while snus users were more likely to have mid-level education and mid-level income. Both smokers and snus users were more likely to have a psychiatric disorder diagnosis before the birth of their child. Smokers (mean: 28.8 years; SD: 5.95) were younger on average than non-smokers (mean: 30.73 years; SD: 5.00). In comparison, the average age of snus users (mean: 30.4 years; SD: 5.51) was closer to non-users (mean: 30.6 years; SD: 5.11). Further descriptive results separated by categories of family-level smoking/snus use, by timing of smoking/snus use, and by patterns of change in smoking status across pregnancy, can be viewed in Appendix D, Section D.2.1.

Table 6-1: Cohort characteristics by exposure status during pregnancy

		N(%)			
Variable	Level	Non-smokers	Smokers	Non-snus users	Snus users
Total		979949 (100.00)	94139 (100.00)	1059398 (100.00)	14690 (100.00)
Intellectual disability	No	974013 (99.39)	93083 (98.88)	1052533 (99.35)	14563 (99.14)
	Yes	5936 (0.61)	1056 (1.12)	6865 (0.65)	127 (0.86)
Small for gestational age	No	958347 (97.80)	89851 (95.45)	1033866 (97.59)	14332 (97.56)
	Yes	18608 (1.90)	3966 (4.21)	22257 (2.10)	317 (2.16)
Sex	Female	476381 (48.61)	45547 (48.38)	514816 (48.60)	7112 (48.41)
	Male	503568 (51.39)	48592 (51.62)	544582 (51.40)	7578 (51.59)
Parity	1	437416 (44.64)	40127 (42.63)	471244 (44.48)	6299 (42.88)
	2	364099 (37.15)	29042 (30.85)	388035 (36.63)	5106 (34.76)
	3 or more	178434 (18.21)	24970 (26.52)	200119 (18.89)	3285 (22.36)
Highest parental education	High School	36052 (3.68)	13891 (14.76)	49300 (4.65)	643 (4.38)
	Gymnasium	380608 (38.84)	63754 (67.72)	436102 (41.17)	8260 (56.23)
	University	563289 (57.48)	16494 (17.52)	573996 (54.18)	5787 (39.39)
Adjusted family income	1	102675 (10.48)	18678 (19.84)	119926 (11.32)	1427 (9.71)
	2	194158 (19.81)	32984 (35.04)	223192 (21.07)	3950 (26.89)
	3	221203 (22.57)	21894 (23.26)	239088 (22.57)	4009 (27.29)
	4	232245 (23.70)	14094 (14.97)	243171 (22.95)	3168 (21.57)
	5	229668 (23.44)	6489 (6.89)	234021 (22.09)	2136 (14.54)
	No	953227 (97.27)	86834 (92.24)	1026271 (96.87)	13790 (93.87)

		N(%)			
Variable	Level	Non-smokers	Smokers	Non-snus users	Snus users
Maternal anxiety diagnosis	Yes	26722 (2.73)	7305 (7.76)	33127 (3.13)	900 (6.13)
Maternal depression diagnosis	No	961429 (98.11)	89311 (94.87)	1036732 (97.86)	14008 (95.36)
	Yes	18520 (1.89)	4828 (5.13)	22666 (2.14)	682 (4.64)
Maternal psychosis diagnosis	No	976973 (99.70)	93250 (99.06)	1055651 (99.65)	14572 (99.20)
	Yes	2976 (0.30)	889 (0.94)	3747 (0.35)	118 (0.80)
Maternal addiction diagnosis	No	968978 (98.88)	88135 (93.62)	1043047 (98.46)	14066 (95.75)
	Yes	10971 (1.12)	6004 (6.38)	16351 (1.54)	624 (4.25)
Any maternal psychiatric diagnosis	No	933451 (95.26)	80569 (85.59)	1000960 (94.48)	13060 (88.90)
	Yes	46498 (4.74)	13570 (14.41)	58438 (5.52)	1630 (11.10)
Any paternal psychiatric diagnosis	No	946080 (96.54)	84825 (90.11)	1017158 (96.01)	13747 (93.58)
	Yes	33869 (3.46)	9314 (9.89)	42240 (3.99)	943 (6.42)
Any maternal neurodevelopmental diagnosis	No	978680 (99.87)	93377 (99.19)	1057456 (99.82)	14601 (99.39)
	Yes	1269 (0.13)	762 (0.81)	1942 (0.18)	89 (0.61)
Any paternal neurodevelopmental diagnosis	No	978447 (99.85)	93381 (99.19)	1057190 (99.79)	14638 (99.65)
	Yes	1502 (0.15)	758 (0.81)	2208 (0.21)	52 (0.35)
Maternal country of origin	Africa	29336 (2.99)	918 (0.98)	30146 (2.85)	108 (0.74)
	Americas	10549 (1.08)	658 (0.70)	11147 (1.05)	60 (0.41)
	Asia	30982 (3.16)	1470 (1.56)	32234 (3.04)	218 (1.48)
	Europe	47505 (4.85)	7662 (8.14)	54962 (5.19)	205 (1.40)
	Middle East	55131 (5.63)	4542 (4.82)	59480 (5.61)	193 (1.31)
	Oceania	422 (0.04)	26 (0.03)	443 (0.04)	5 (0.03)
	Scandinavia	15363 (1.57)	2228 (2.37)	17380 (1.64)	211 (1.44)
	Swedish	790661 (80.68)	76635 (81.41)	853606 (80.57)	13690 (93.19)
Birth year	1999-2001	209614 (21.39)	20946 (22.25)	228179 (21.54)	2381 (16.21)
	2002-2004	239147 (24.40)	28067 (29.81)	263148 (24.84)	4066 (27.68)
	2005-2007	248260 (25.33)	22508 (23.91)	267039 (25.21)	3729 (25.38)
	2008-2010	282928 (28.87)	22618 (24.03)	301032 (28.42)	4514 (30.73)
Any maternal smoking in pregnancy	No			966623 (91.24)	13326 (90.71)
	Yes			92775 (8.76)	1364 (9.29)
Any maternal snus use in pregnancy	No	966623 (98.64)	92775 (98.55)		
	Yes	13326 (1.36)	1364 (1.45)		

6.2.2 – Missing data assessment

Descriptive tables of the patterns of missing data are presented in Appendix D, section D.2.2. Missing data in covariates (see Table D.2.2-1) was socially patterned with those with lower

income and education and those born in later cohort years more likely to be excluded for having a missing covariate. Missing data in the exposure (see Table D.2.2-2) was more likely in those with lower income and less likely for those born in later cohort years. Smokers were more likely to have been excluded for having missing snus data and vice versa. Both diagnosis of ID and being born SGA were associated with increased risk of exclusion for missing covariate and exposure data.

6.2.3 – Primary analyses

Table 6-2 shows the results for the primary analyses using offspring ID as an outcome. Conventional analyses (i.e. Models 1 and 2) showed that both smoking and snus use in pregnancy were associated with increased odds of ID following adjustment for confounders. When separated into within-family and between-family effects, there was evidence of between-family but not within-family effects of smoking and snus use in pregnancy before and after adjustment for confounders.

The results of the primary analyses using offspring SGA as the outcome are presented in Table 6-3. Smoking in pregnancy was associated with a population-averaged increased odds of offspring SGA after adjustment for confounders. Model 3 and 4 showed that smoking in pregnancy was associated with increased odds of SGA for both the within-family and between-family effects. Snus use in pregnancy was not associated with offspring SGA in any model. For most models, the confidence intervals for estimates of the effect of snus use were not compatible with that of smoking, suggesting that the absence of an association between snus use and SGA was not the result of a lack of power for a rarer exposure.

Table 6-2: Primary analysis of the association between exposure and offspring ID.

		Smoking in pregnancy		Snus use in pregnancy	
Model	Coefficient	O.R. [a]	95% CI	O.R. [a]	95% CI
1 - Conventional unadjusted [b]	Population-averaged	1.80	(1.68-1.92)	1.37	(1.14-1.64)
2 - Conventional adjusted [c]	Population-averaged	1.24	(1.16-1.33)	1.28	(1.07-1.53)
3 - Within-between unadjusted [b, d]	Within-family	0.92	(0.74-1.14)	0.89	(0.61-1.31)
	Between-family	2.12	(1.68-2.66)	1.72	(1.12-2.62)
4 - Within-between adjusted [c, d]	Within-family	0.92	(0.75-1.14)	0.88	(0.60-1.31)
	Between-family	1.39	(1.11-1.75)	1.59	(1.04-2.45)

[a] Estimates produced using a total sample size of 1,074,088 individuals from 705,862 families including 6,992 cases of ID.

[b] Model adjusted for year of birth.

[c] Model adjusted for year of birth, sex, parity, highest parental education, income, parental psychiatric history, maternal country of origin and maternal age at birth.

[d] Model adjusted for family averaged exposure.

Table 6-3: Primary analysis of the association between exposure and offspring SGA.

		Smoking in pregnancy		Snus use in pregnancy	
Model	Coefficient	O.R. [a]	95% CI	O.R. [a]	95% CI
1 - Conventional unadjusted	Population-averaged	2.26	(2.18-2.34)	1.02	(0.91-1.14)
2 - Conventional adjusted [b]	Population-averaged	2.18	(2.10-2.27)	1.05	(0.93-1.17)
3 - Within-between unadjusted [c]	Within-family	1.67	(1.49-1.87)	1.01	(0.81-1.27)
	Between-family	1.40	(1.24-1.59)	1.02	(0.78-1.32)
4 - Within-between adjusted [b, c]	Within-family	1.43	(1.27-1.62)	1.07	(0.84-1.36)
	Between-family	1.61	(1.41-1.84)	0.97	(0.74-1.28)

[a] Estimates produced using a total sample size of 1,070,772 individuals from 704,487 families including 22,574 cases of SGA.

[b] Model adjusted for year of birth, sex, parity, highest parental education, income, parental psychiatric history, maternal country of origin and maternal age at birth.

[c] Model adjusted for family averaged exposure.

6.2.4 – Secondary analyses

The results for offspring ID are presented in Table 6-4. An exposure-duration-response of increased odds of ID was found for smoking and using snus later into pregnancy in conventional models only (Models 1 and 2). In conditional logistic models that calculated within-family estimates of the exposure-ID association, no association was found for smoking for any exposure timing. Within-family estimates of the snus use-ID association showed evidence of decreased odds of ID among those who quit using snus during pregnancy compared to those who did not use snus at any time.

Table 6-5 shows the results of the timing analyses for offspring SGA as the outcome. Smoking longer into pregnancy was associated with a duration-responsive increase in odds of offspring SGA in conventional and conditional logistic analyses (Models 1-4). For smoking (Model 2) and snus use (Models 1 and 2) in pregnancy, there was evidence for a reduced risk of offspring SGA in conventional models for mothers who gave up using before pregnancy compared to those who did not use snus at any time. There was no other evidence for an association between snus use and offspring SGA.

6.2.5 – Sensitivity analyses

In the first sensitivity analysis I tested whether the results would be influenced by exposure misclassification in the exposure discordant group. The results are presented in Table 6-6. There were 7126 individuals who were reported as not exposed to smoking had their exposure status changed (0.66% of unexposed individuals). 13,178 individuals (34.68% of exposure discordant individuals) from 5547 families (35.45% of exposure discordant families) who had been classed as exposure discordant became exposure concordant in the sensitivity analysis. Comparison of the results of the sensitivity analysis with the primary analysis shows that occurrence of the extreme hypothetical situation, in which all unexposed individuals with exposed younger siblings were misclassified, would only influence the results of the SGA within-between model that was not adjusted for confounders. No other models showed notable changes to parameter estimates.

In the second sensitivity analysis I tested whether there were differing effects for those who started smoking between pregnancies versus those who stopped smoking between pregnancies by applying different restrictions to the cohort. Table 6-7 indicates that the different patterns of change in smoking across pregnancies may have slightly different strengths of association; the within-family estimates of Model 4 across restricted cohorts show that those who stopped smoking between pregnancies had a smaller OR than those who started smoking between pregnancies. This was true for ID and SGA. These differences in effect size were not large enough to substantially change the conclusions from the analyses.

Table 6-4: Secondary analysis of the association between exposure timing and offspring ID.

		Smoking in pregnancy		Snus use in pregnancy	
Model	Coefficient	O.R. [a]	95% CI	O.R. [b]	95% CI
1 - Conventional unadjusted [c]	Non-user	1.00		1.00	
(population-averaged estimates)	User before pregnancy only	1.04	(0.96-1.13)	0.76	(0.60-0.97)
	Quit during pregnancy	1.40	(1.19-1.64)	1.05	(0.75-1.47)
	Used late into pregnancy	1.79	(1.65-1.95)	2.05	(1.53-2.76)
2 - Conventional adjusted [c, d]	Non-user	1.00		1.00	
(population-averaged estimates)	User before pregnancy only	0.90	(0.82-0.98)	0.85	(0.67-1.08)
	Quit during pregnancy	1.04	(0.89-1.22)	1.02	(0.73-1.43)
	Used late into pregnancy	1.17	(1.07-1.28)	1.82	(1.36-2.45)
3 - Unadjusted conditional logistic [c]	Non-user	1.00		1.00	
(within-family estimates)	User before pregnancy only	0.93	(0.76-1.15)	0.87	(0.54-1.41)
	Quit during pregnancy	0.95	(0.64-1.39)	0.43	(0.19-0.96)
	Used late into pregnancy	1.07	(0.76-1.51)	1.18	(0.52-2.68)
4 - Adjusted conditional logistic [c, d]	Non-user	1.00		1.00	
(within-family estimates)	User before pregnancy only	0.91	(0.73-1.13)	0.91	(0.56-1.47)
	Quit during pregnancy	0.93	(0.63-1.38)	0.41	(0.18-0.93)
	Used late into pregnancy	1.02	(0.72-1.44)	1.10	(0.49-2.52)

[a] Estimates conventional models produced using a total sample size of 1,054,561 individuals from 696,247 families including 6,696 cases of ID. Estimates for conditional logistic models produced using a total sample size of 8,493 individuals including 3,644 cases of ID.

[b] Estimates for conventional models produced using a total sample size of 1,068,173 individuals from 703,231 families including 6,946 cases of ID. Estimates for conditional logistic models produced using a total sample size of 8,879 individuals including 3,797 cases of ID.

[c] Model adjusted for year of birth.

[d] Model adjusted for year of birth, sex, parity, highest parental education, income, parental psychiatric history, maternal country of origin and maternal age at birth.

Table 6-5: Secondary analysis of the association between exposure timing and offspring SGA.

		Smoking in pregnancy		Snus use in pregnancy	
Model	Coefficient	O.R. [a]	95% CI	O.R. [b]	95% CI
1 - Conventional unadjusted (population-averaged estimates)	Non-user	1.00		1.00	
	User before pregnancy only	1.04	(0.99-1.09)	0.82	(0.73-0.92)
	Quit during pregnancy	1.46	(1.34-1.60)	0.92	(0.75-1.11)
	Used late into pregnancy	2.42	(2.31-2.53)	0.98	(0.77-1.23)
2 - Conventional adjusted [c] (population-averaged estimates)	Non-user	1.00		1.00	
	User before pregnancy only	0.90	(0.86-0.94)	0.78	(0.70-0.88)
	Quit during pregnancy	1.30	(1.19-1.42)	0.92	(0.76-1.12)
	Used late into pregnancy	2.37	(2.26-2.48)	1.04	(0.82-1.32)
3 - Unadjusted conditional logistic (within-family estimates)	Non-user	1.00		1.00	
	User before pregnancy only	1.76	(1.54-2.01)	1.17	(0.86-1.58)
	Quit during pregnancy	2.00	(1.56-2.56)	1.11	(0.72-1.71)
	Used late into pregnancy	2.86	(2.29-3.56)	1.10	(0.56-2.14)
4 - Adjusted conditional logistic [c] (within-family estimates)	Non-user	1.00		1.00	
	User before pregnancy only	0.96	(0.83-1.11)	0.92	(0.66-1.29)
	Quit during pregnancy	1.13	(0.86-1.48)	1.28	(0.79-2.07)
	Used late into pregnancy	1.79	(1.41-2.26)	1.54	(0.75-3.16)

[a] Estimates for conventional models produced using a total sample size of 1,051,383 individuals from 694,917 families including 21,628 cases of SGA. Estimates for conditional logistic models produced using a total sample size of 22,093 individuals including 10,049 cases of SGA.

[b] Estimates for conventional models produced using a total sample size of 1,064,837 individuals from 701,858 families including 22,431 cases of SGA. Estimates for conditional logistic models produced using a total sample size of 23,066 individuals including 10,473 cases of SGA.

[c] Model adjusted for year of birth, sex, parity, highest parental education, income, parental psychiatric history, maternal country of origin and maternal age at birth.

Table 6-6: Sensitivity analysis assuming all those who had younger siblings who were exposed to smoking were also exposed themselves.

			Primary analyses		Sensitivity analyses	
Outcome	Model	Coefficient	O.R.	95% CI	O.R.	95% CI
Intellectual disability [a]	1 - Conventional unadjusted [c]	Population-averaged	1.80	(1.68-1.92)	1.80	(1.68-1.92)
	2 - Conventional adjusted [d]	Population-averaged	1.24	(1.16-1.33)	1.24	(1.16-1.33)
	3 - Within-between unadjusted [c, e]	Within-family	0.92	(0.74-1.14)	0.90	(0.68-1.18)
		Between-family	2.12	(1.68-2.66)	2.11	(1.58-2.81)
	4 - Within-between adjusted [d, e]	Within-family	0.92	(0.75-1.14)	0.90	(0.68-1.18)
		Between-family	1.39	(1.11-1.75)	1.42	(1.06-1.89)
Small for gestational age [b]	1 - Conventional unadjusted	Population-averaged	2.26	(2.18-2.34)	2.21	(2.14-2.29)
	2 - Conventional adjusted [d]	Population-averaged	2.18	(2.10-2.27)	2.11	(2.03-2.19)
	3 - Within-between unadjusted [e]	Within-family	1.67	(1.49-1.87)	2.39	(2.07-2.76)
		Between-family	1.40	(1.24-1.59)	0.92	(0.79-1.07)
	4 - Within-between adjusted [d, e]	Within-family	1.43	(1.27-1.62)	1.37	(1.18-1.60)
		Between-family	1.61	(1.41-1.84)	1.60	(1.36-1.87)

[a] Estimates produced using a total sample size of 1,074,088 individuals from 705,862 families including 6,992 cases of ID.

[b] Estimates produced using a total sample size of 1,070,772 individuals from 704,487 families including 22,574 cases of SGA.

[c] Model adjusted for year of birth.

[d] Model adjusted for year of birth, sex, parity, highest parental education, income, parental psychiatric history, maternal country of origin and maternal age at birth.

[e] Model adjusted for family averaged exposure.

Table 6-7: Sensitivity analysis of a restricted cohort of the first two pregnancies study period.

			Restricted dataset of 1st two children in cohort (Restricted cohort 1) [a]		Discordant sample restricted to those who stop smoking in second pregnancy (Restricted cohort 2) [b]		Discordant sample restricted to those who start smoking in second pregnancy (Restricted cohort 3) [c]	
Outcome	Model	Coefficient	O.R.	95% CI	O.R.	95% CI	O.R.	95% CI
Intellectual disability	1 - Conventional unadjusted [d]	Population-averaged	1.77	(1.65-1.90)	1.80	(1.67-1.93)	1.88	(1.75-2.02)
	2 - Conventional adjusted [e]	Population-averaged	1.23	(1.14-1.32)	1.23	(1.14-1.33)	1.27	(1.17-1.37)
	3 - Within-Between unadjusted [d, f]	Within-family	0.87	(0.69-1.11)	0.74	(0.55-1.01)	1.03	(0.71-1.51)
		Between-family	2.20	(1.71-2.82)	2.58	(1.89-3.52)	1.86	(1.27-2.73)
	4 - Within-Between adjusted [e, f]	Within-family	0.88	(0.70-1.11)	0.74	(0.54-1.00)	1.08	(0.75-1.56)
		Between-family	1.45	(1.13-1.85)	1.74	(1.27-2.38)	1.18	(0.82-1.71)
Small for Gestational age	1 - Conventional unadjusted	Population-averaged	2.27	(2.19-2.36)	2.37	(2.28-2.46)	2.29	(2.21-2.39)
	2 - Conventional adjusted [e]	Population-averaged	2.17	(2.09-2.26)	2.20	(2.11-2.29)	2.27	(2.17-2.36)
	3 - Within-Between unadjusted [f]	Within-family	1.61	(1.42-1.82)	2.24	(1.92-2.60)	0.87	(0.71-1.07)
		Between-family	1.47	(1.29-1.68)	1.06	(0.91-1.25)	2.74	(2.23-3.37)
	4 - Within-Between adjusted [e, f]	Within-family	1.40	(1.22-1.59)	1.31	(1.11-1.54)	1.59	(1.27-1.98)
		Between-family	1.64	(1.43-1.89)	1.77	(1.49-2.10)	1.45	(1.16-1.81)

[a] Estimates produced using a total sample size of 967,361 individuals for ID analyses and 964,371 individuals for SGA analyses (cases of ID = 6,352; SGA = 20,952).

[b] Estimates produced using a total sample size of 957,371 individuals for ID analyses and 954,411 individuals for SGA analyses (cases of ID = 6,265; SGA = 20,712).

[c] Estimates produced using a total sample size of 949,535 individuals for ID analyses and 946,604 individuals for SGA analyses (cases of ID = 6,219; SGA = 20,484).

[d] Model adjusted for year of birth.

[e] Model adjusted for year of birth, sex, parity, highest parental education, income, parental psychiatric history, maternal country of origin and maternal age at birth.

[f] Model adjusted for family averaged exposure.

6.3 – Discussion

The results of this study provide evidence that supports a causal role of smoking but not snus use in offspring fetal growth restriction, as measured by SGA, and no evidence of a causal influence of smoking and snus use in pregnancy on risk of offspring ID. Both population-averaged and within-family effect estimates suggested a role of smoking on fetal growth restriction. The within-family effect can be interpreted as meaning that individual level exposure to maternal smoking in pregnancy, holding fixed shared familial genetics and environment, is associated with being born SGA. In comparison, no association was found between snus use in pregnancy and offspring fetal growth restriction, even in unadjusted conventional models. Maternal smoking and snus use in pregnancy were both associated with increased population-averaged odds of ID. In both cases this was shown to be driven by the between-family effect and not the within-family effect; a finding that is not consistent with a causal effect.

The results of the timing analyses supported the conclusions of the causal nature of each exposure-outcome association. A duration-response was found for the within-family effect estimates of smoking in pregnancy and offspring SGA but not for any other investigated association. I did however obtain some unusual results for those who quit smoking or snus use. Compared to no use, giving up smoking or snus before or during pregnancy was associated with reduced odds of SGA and ID in some models. While a protective effect of nicotine at critical time points may be possible, I believe that these results could potentially be explained by the characteristics of mothers who are able to quit using an addictive substance at an important time in order to benefit their child's health.

This study strengthens the current body of evidence for a causal effect of maternal smoking in pregnancy on fetal growth restriction and provides strong support to the suggestion that the association between smoking in pregnancy and offspring ID is the result of residual confounding [98, 99]. To my knowledge, no prior research has investigated the association between prenatal exposure to snus and offspring ID.

This study used a similar Swedish cohort and an identical definition of SGA to the studies by Baba et al [276], who found an association between maternal snus use in pregnancy and offspring born SGA, and Wikström et al [278] who did not. The biggest difference between the sample in this study and that of Baba et al was the exclusion criteria; we excluded those with metabolic, genetic and chromosomal abnormalities where they did not. It is possible that these exclusion criteria lead to the differences in associations found if either these disorders play a mediatory role in the association or if they are common effects of snus use in pregnancy and being born SGA and thereby introduce collider bias towards the null. I repeated the snus

analysis without these exclusion criteria. The results (presented in section D.3 of Appendix D) show no substantial difference from those presented in the primary and secondary analyses which suggests that different exclusion criteria did not lead to the difference in findings.

6.3.1 – Strengths and limitations

The present study suffers from limitations that all registry-based studies are subject to, including potential misclassification, missing data and residual confounding. Sensitivity analysis (i) in Table 6-6 suggested that the results would not be substantially changed if there were high levels of exposure misclassification in the exposure-discordant group, though I only tested one scenario of exposure misclassification.

Complete case analysis was used, which may be biased for logistic regression when the missing data is related to both the exposure and the outcome [150], as is the case for this dataset. Previous work has shown that there is smaller bias at smaller proportions of missing data [218]. The small quantity of missing data in the dataset may limit the bias, however the strong associations between smoking/snus use status and missing data for snus use/smoking data may have led to greater quantities of bias. Given that smokers/snus users and those with ID were both more likely to have been excluded for missing data this may have biased the association towards the null.

This study attempted to account for residual confounding by using a sibling design which holds fixed shared familial genetics and environment. I was unable to easily control for the non-shared confounders of siblings, which have been shown to bias the results of sibling designs [181], due to the varying size of families. The antenatal nature of the associations that I am investigating, however, mean that the non-shared confounders will be limited to changes in environment between pregnancies as the mother's genetics will not change. Sibling designs are also subject to bias from carry-over effects, where the outcome of the first sibling influences the exposure of the second sibling [182]. The difference in effect estimates between cohorts that restricted to specific patterns of changes in smoking between pregnancies (sensitivity analysis (ii) in Table 6-7) was small suggesting a limited influence of bias from carry-over effects.

A key strength of this study is the relationship between the associations investigated. There is strong evidence for a causal effect of smoking in pregnancy on fetal growth restriction while the causal nature of the association between smoking in pregnancy and ID is unclear. I found evidence using within-between models of a causal influence of smoking in pregnancy on fetal growth restriction but not ID. As this is in line with previous findings the former association can be thought of as a positive control of the latter association, thereby strengthening confidence in the evidence produced for the non-causal influence of exposure to maternal smoking in

pregnancy on offspring ID. Further, had a causal effect been suggested, comparison of the influence of smoking and snus use in pregnancy on offspring ID would have provided a cross context comparison which could have been useful for identifying if nicotine or combustible components of smoking were involved in biological mechanisms.

6.4 – Conclusions and chapter summary

This study has provided no evidence for a causal effect of smoking or snus use in pregnancy on risk of offspring ID, instead suggesting that associations are the result of residual confounding. Further no evidence was found for a causal effect of snus use in pregnancy on risk of offspring being born SGA however a causal effect of smoking in pregnancy on offspring SGA was supported. Neither finding suggests that smoking or snus use in pregnancy is safe. Smoking in pregnancy has well established negative effects on offspring health while research into snus use in pregnancy is in its infancy. Further assessments of the health costs and benefits of snus use in pregnancy relative to smoking need to be performed before guidance can be given regarding whether snus use is a suitable alternative to smoking or even whether it should be used as a cessation aid during this period.

Chapter 7 Applying the negative control design and Mendelian Randomisation to the Avon Longitudinal Study of Parents and Children

So far in this thesis I have set out the evidence for an association between maternal smoking during pregnancy and offspring risk of intellectual disability (ID), that this association is potentially the result of confounding factors and that further investigation is required (Chapter 1). I have established how we can investigate the causal nature of such an association using observational data (Chapter 2), investigated how assortative mating can bias results from the negative control design (Chapter 3) and how the proportion of missing data influences bias and efficiency of multiple imputation (MI) analysis (Chapter 4). In the last two chapters I used sibling comparisons to investigate the overarching research question using Danish and Swedish registry data.

The current chapter aims to use a UK based cohort study, the Avon Longitudinal Study of Parents and Children (ALSPAC) to further investigate the association using observational, negative control and Mendelian Randomisation (MR) analyses. The cohort, described in greater detail below, contains a wealth of data on mother, partner and child collected from pregnancy onwards. Genetic information has also been recorded on both mother and child providing a rare opportunity to explore the cross generational genetic associations that are required for implementing MR to assess the causal nature of intergenerational exposure-outcome associations.

For this chapter the term paternal is replaced with partner as not all partners in ALSPAC are the biological father of the child. It has been argued that when the exposure is a behaviour, such as smoking, whether the partner is the biological father is not important to the negative control design [8]. Where the exposure is biological, such as body mass index, non-paternity may be of greater relevance.

The chapter is set out as follows. First I explore the data available within the ALSPAC study to create an appropriate, well defined measure of ID. Multiple measures, from different time points and from different sources, are available for each participant of the cohort. A core focus of the investigation in this chapter revolved around creating a multi-sourced indicator variable that

could overcome some of the difficulties in defining ID as highlighted in Chapter 1. Previous studies of neurodevelopmental conditions have shown that multi-source methods of identifying cases may reduce measurement error [21, 22]. This is followed by description of the methods and results of the investigation into the causal nature of the association of maternal smoking during pregnancy and offspring risk of ID. An attempt is then made to triangulate across the evidence from the three analysis methods.

7.1 – Identifying individuals with intellectual disability

7.1.1 – Cohort specification

The ALSPAC cohort [242, 243] recruited 14,541 pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992. The core sample of pregnancies (also referred to as phase I) contained a total of 14,676 fetuses that resulted in 14,062 live births; 13,988 of these children were alive at 1 year of age. Data has been collected on the cohort since its inception and is still ongoing.

Attempts were made to bolster the initial core sample with eligible cases who had failed to join the study originally. These attempts were made in 1999 when the oldest children were approximately 7 years of age (phase II recruitment), opportunistically from 1999-2012 (phase III) and then from 2012 onwards with specific focus on recruiting second generation pregnancies (phase IV) [286]. Children recruited during phases II-IV were not included in analyses in this study as pregnancy information would be recorded retrospectively, increasing the risk of information bias.

There were 15,659 total ALSPAC mother-child pairs across phase I-IV recruitment. Of these, 795 had no NHS number and so could not be linked to the UK Secure eResearch Platform (UKSeRP) where the data were held, 1 participant withdrew consent at this stage. Of the remaining, 931 were not recruited during phase I, 68 were not alive at 1 year of age and 355 were not singleton births. This left a sample of 13,509 mother child pairs. A cohort flow diagram is presented in Figure 7-1, that describes the exclusion process for each stage of the study.

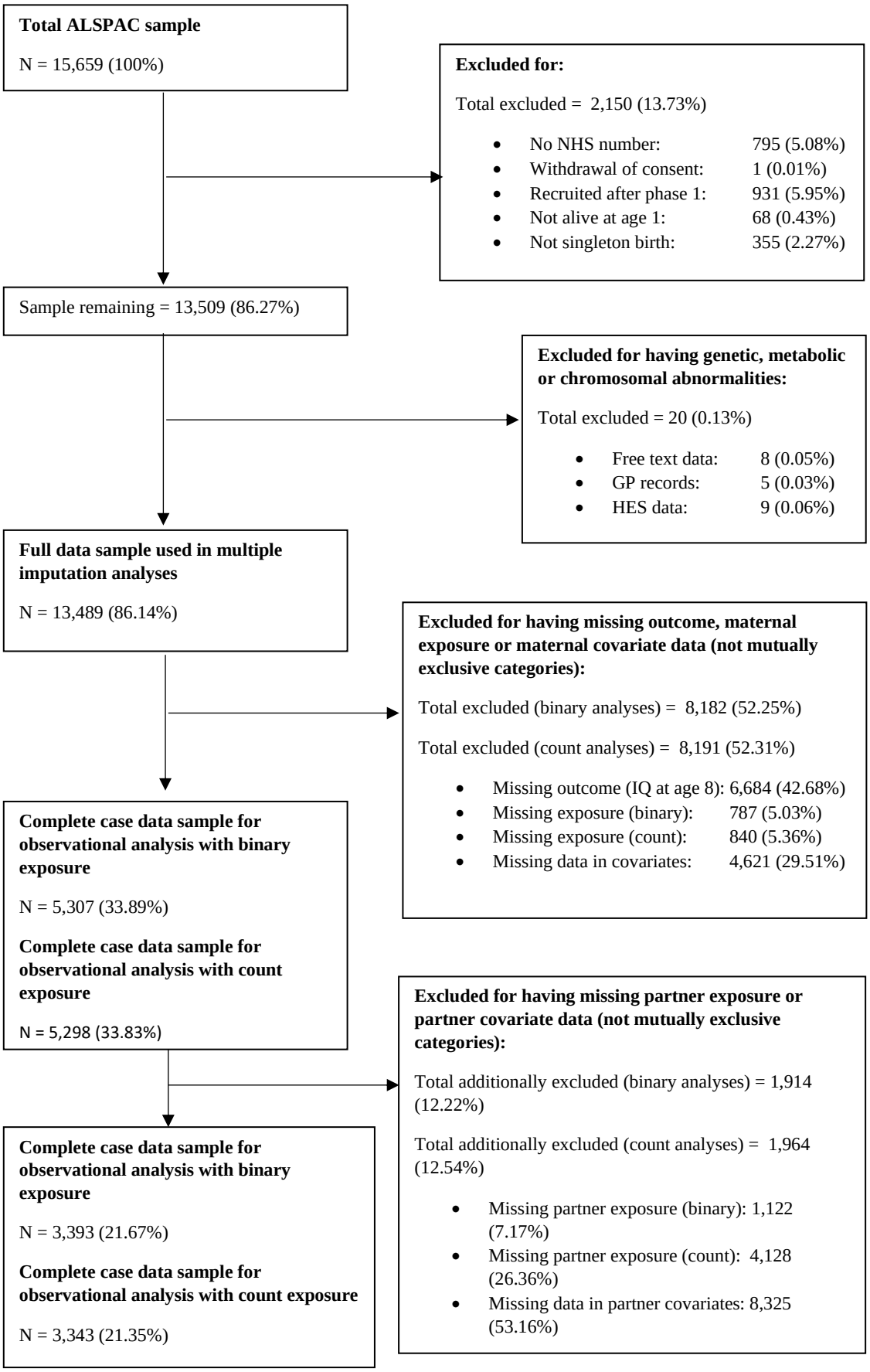

Figure 7-1: Flowchart of cohort derivation

7.1.2 – Derivation of a multi-sourced variable for intellectual disability

It was anticipated that having ID may lead to reduced or discontinued participation in the ALSPAC cohort. This would give rise to a missing not at random (MNAR) missing data pattern (see Chapter 2 for an explanation of missing data patterns) that was dependent on the outcome variable. Such a pattern would lead to bias under complete case analysis [150]. I therefore attempted to create a multi-sourced indicator variable for ID that made use of data recorded as part of the ALSPAC study as well as data from external linked sources. The linked data was intended to retain information on those who no longer participated in ALSPAC, thereby reducing bias from missing data, and increasing the accuracy of the measure of ID.

Data from ALSPAC sources included measures of IQ taken at age 8 and 15 and free text fields in child-based questionnaires where the responder could record additional information. The linked sources included statements of special educational needs recorded in the Pupil Level Annual School Census (PLASC), Read codes related to ID recorded in General Practitioner (GP) records, International Classification of Disease (ICD) [12] diagnosis codes contained in Hospital Episode Statistics (HES) data and interactions with mental health services for reasons related to ID held in the Mental Health Services Data Set (MHSDS). Data linkage has previously been undertaken in the Identification of Developmental Impairments (IDI) project led by Emond [287] which identified neurodevelopmental disorders up to a maximum age of 11 years using ICD-10 diagnoses. Further details of each source of information is provided in the subsections below as well as the derivation of the outcome variable used for analyses.

At the time of writing, linked health records (including GP records, HES data and MHSDS data) were only available for individuals who had explicitly consented for linkage to data records. Beginning in November 2017, I began the process of a Confidentiality Advisory Group (CAG) application [288] to obtain access to the information of those who had not explicitly consented to data linkage via use of Section 251 of the National Health Service Act 2006 [289]. The CAG application, submitted by the ALSPAC data linkage team [290], via the Integrated Research Application System [291] (CAG reference: 20.CAG/0056; IRAS project ID: 268410), has been provisionally supported subject to satisfactory responses to requests for further information and compliance. At the time of analysis (July 2020) full approvals were still not in place and based on previous experience full access to non-explicit consenters data are likely to require several additional months. As a result linked health records for non-explicit consenters (described herein as the Section 251 sample) were not included.

7.1.2.1 – *IQ scores*

IQ at age 8 years was measured using the Wechsler Intelligence Scale for Children - III [245] as part of a half day battery of mainly psychological and psychometric testing. IQ at age 15 years was measured using the Wechsler Abbreviated Scale of Intelligence [244] as part of a 4 hour battery of testing. Data was available for 6,613 (50.43%) of individuals at age 8 and for 4,869 (36.04%) of individuals at age 15. From the IQ scores binary variables were created indicating if IQ was below 70 at each age. A second variable was created indicating a less stringent cut off of IQ below 85, equivalent to one population standard deviation below the population average of 100.

7.1.2.2 – *Free text fields*

ALSPAC contains free text responses to many questions answered by participants and their guardians. For example, at age 9 guardians of participant children were asked whether the children had been identified as having any particular problems at school and to describe in text each type of school problem. A search was performed across all free text fields contained in ALSPAC for terms related to ID (see Table E.1.1-1 in Appendix E for the search terms used and number of hits). A review of all free text responses for each individual identified with relevant free text fields (n=208) was performed to check if the text indicated whether the child was likely to have ID or not. Any queries were checked by a clinician who specialises in neurodevelopmental disorders (Dr Dheeraj Rai). The search terms identified several individuals who had specific learning difficulties such as dyslexia or difficulties specific to maths and literacy ability; the terms also identified individuals who were explicitly stated to not having a learning disability. Neither of these groups were classified as having ID. Following the review of all free text fields for each identified individual, 97 individuals were classed as having ID and 111 individuals were classed as not having ID. Free text data was available for 12,104 individuals in the sample.

7.1.2.3 – *Pupil level annual school census statements of special educational needs*

Statements of special educational needs (SEN) for cognition and learning needs [292] were used to indicate ID. These statements were recorded in 2003/4 when the vast majority of the sample children were in school years 6-8 (ages 11-13). The cognition and learning needs category include those with a statement for moderate to profound learning difficulties, but also includes individuals with specific learning difficulties related to problems learning to read, write, spell or manipulate numbers. This latter group is not of interest to the research question in this thesis. The data available to researchers is a binary indicator of all those who have a SEN statement included within the cognition and learning needs category, including both those with specific

(not of interest) and non-specific (of interest) learning difficulties. To attempt to address this issue, teacher reports of specific learning difficulties, recorded during approximately the same time period, were used to supplement the information. Individuals whose teacher reported a specific learning difficulty were edited to have an indicator of not having ID according to the PLASC. It should be acknowledged that some people who have an intellectual disability may also have a specific learning difficulty such as dyslexia but I decided to take the most conservative approach to identify cases.

Data on the PLASC is available for 11,458 (81.5%) of the 14,062 live born children in the core ALSPAC sample. Those who did not have a PLASC record either did not attend state school in England (includes those attending independent schools, schools outside of England or those educated at home) or could not be matched (for example if their name was changed without ALSPAC being informed). Absence of PLASC information may therefore be associated with socioeconomic variables. Data were available for 9,695 (71.8%) of the sample for analysis.

7.1.2.4 – *GP records*

GP records contain information in the form of read codes [293, 294]. These are a hierarchically coded thesaurus of clinical terms that have been in use by the NHS since 1985. The codes are entered into a computerised system by clinicians or practice staff from general practice or secondary care consultations. A list of version 2 read codes was created by checking for terms related to intellectual disability or its synonyms using the UK Read Browser, previously accessible from NHS digital's Technology Reference data Update Distribution. The list of read codes identified was cross checked against a list of codes selected in a previous study looking at incidence of mental illness and challenging behaviour in individuals with ID [295]. Terms that appeared in either list were used (see Table E.1.1-2 in Appendix E for the read codes used). Data was available for 4,659 individuals (34.5% of the analysis sample) who had explicitly consented, pending for 7,001 (51.8%) who required a S251 approval and completely unavailable for 1,849 (13.7%) individuals.

7.1.2.5 – *Hospital episode statistics*

Details of all admissions, attendances at accident and emergency and any outpatient appointments at NHS hospitals in England are collected in the HES database [296]. The HES dataset recorded all diagnoses up until 1995 using ICD-9 and all diagnoses in subsequent years as ICD-10 codes [12, 297]. Diagnoses of 317-319 (ICD-9) and F70-F79 (ICD-10) made during hospital interactions were used to indicate ID (full list presented in Appendix E, Table E.1.1-3). Data availability was the same as for GP records.

7.1.2.6 – *Mental health services data set*

The MHSDS collects data on all interactions between patients and specialist secondary mental health care services [298]. Patients are assigned to mental health clusters using the Health of the Nation Outcome Scales [299] which can be used to indicate the nature of the mental health care. Information regarding intellectual disability can be found within care clusters 18-21 which relate to cognitive impairment. All individuals that had more than one recorded final clinician allocated cluster related to cognitive impairment were indicated as having ID. Less than 5 cases were indicated using this method. All were contained within cluster 18.

MHSDS data was only available for 58 individuals (0.4% of the total sample) who had a relevant read code found in GP records or ICD code found in HES data. The sample for who MHSDS data was available was therefore a subsample of the explicitly consenting sample of 4,659 individuals defined in Section 7.1.2.4 .

7.1.2.7 – *IDI project*

The IDI project has been described in detail elsewhere [287]. Briefly, the project identified individuals in the ALSPAC cohort with any form of developmental delay determined by ICD-10 diagnosis codes between 1991 and 2003 (maximum possible age 13). These diagnoses were obtained from computerised medical records of NHS trusts in the local Bristol area (North Bristol Trust, United Bristol Healthcare Trust, Weston Area Health Trust and Royal United Hospital, Bath). The codes F70-F79 were used to select those with a diagnosis of ID.

It was not possible to determine the exact overlap between the IDI project sample and the analysis sample of the current project. This was due to the data retained from the IDI project only containing information on those who had an identified ICD-10 diagnosis and not all those for whom medical records were available at the time of the project. The documentation for the IDI project (which can be obtained from the ALSPAC useful data repository) states that 13,898 of the 14,062 live born individuals who make up the core ALSPAC sample were eligible for the IDI project. It was therefore assumed that data was available on IDI diagnoses for all participants.

7.1.3 – Exclusion criteria

Individuals who had a genetic, metabolic or chromosomal abnormality that was associated with ID were excluded from analyses as this is a group in which ID is likely regardless of exposure to maternal smoking during pregnancy. These disorders were identified using free text information, GP records and HES data. The free text records of individuals with text relevant to

ID were screened for mentions of known genetic causes of ID. Read codes and ICD codes for genetic disorders related to ID were obtained from GP records and HES data respectively. A list of the codes used is presented in Table E.1.2-1 in Appendix E. If a participant had any of these codes they were excluded from the data. In total 20 participants were excluded for genetic, metabolic or chromosomal abnormalities (8 using free text data, 5 using GP records and 9 using HES data) leaving a sample of 13,489 participants. There was low concordance between sources of information; only 2 (10%) of these individuals met the exclusion criteria based on information from more than one source.

7.1.4 – Assessment of the validity of the ID variables

The information available to create a multi-sourced indicator of ID were therefore the following eight items:

1. An IQ less than 70 at age 8
2. IQ less than 70 at age 15
3. Free text fields that suggest the child has ID
4. A statement of SEN for cognitive and learning needs
5. A relevant read code from linked GP records
6. A relevant ICD-9 or ICD-10 diagnosis code from HES data
7. Multiple records indicating use of learning disability care services in the MHSDS
8. An ICD-10 diagnosis found in the IDI project

The distribution of IQ scores for those with ID indicated by each source of information is presented in Table 7-1. The mean IQ at age 8 was less than 70 among those who were indicated as having ID from the IDI project and from HES data but was greater than 70 for those with ID indicated from free text data, SEN statements and GP records. This may suggest that different severities of ID are being identified by the different sources of information or may indicate that some sources contain substantial measurement error and are therefore identifying individuals who are within the normal range of IQ. This measurement error is likely to be amplified in health records data since data is only available for explicit consenters and those with lower IQs are by definition less likely to be able to provide explicit consent. Table 7-2 shows that there was low concordance for ID between sources. Again, this could be due to low agreement between sources as to what counts as ID but could also signify poor overlap in available data between sources.

Consenter status for linked health records (GP records, HES data and MHSDS data) may also influence the ability to identify cases of ID. Table 7-3 presents the number of variables available to identify ID, as well as average IQ scores at age 8 and 15, across categories of consent status.

The section 251 group had on average fewer available sources of information (excluding linked health data sources) than the explicit consenters. The section 251 group also had lower average IQ scores at age 8 and 15 than the explicit consenters. This may suggest that the section 251 group contains more severe cases of ID than the explicit consenters, some of which may not be being picked up due to lack of availability of linked health data.

7.1.5 – Deciding how to define intellectual disability for analysis

The assessment of validity for the different ID variables has highlighted that some variables may contain substantial measurement error while others may be missing the more severe cases of ID due to consent status. Further time than is allowed by the PhD programme is required to obtain the legal approval for the use of the section 251 sample's health records and investigate more thoroughly which variables can be used to derive a valid ID variable. As a result, the intended approach of creating a multi-sourced variable for ID could not be integrated into this thesis.

Instead, I decided to use an IQ of less than 70 at age 8 to define ID in this investigation, using MI to account for missing data. As has been highlighted, those with missing data in this variable are likely to have lower IQ, thereby resulting in a MNAR missing data pattern that could bias analyses. Previous work by Cornish et al. [300] has used linked data as auxiliary information for missing IQ data in ALSPAC to improve the validity of the missing at random (MAR) assumption of MI analyses. The approach was shown to reduce bias and improve efficiency of estimates in their simulation study. Full details of the MI procedure can be found in the Methods section below.

The chosen approach runs counter to the suggestions of O'Brien that IQ testing alone without information on functional impairments should not be used to indicate ID [18]. The results of this chapter should therefore be interpreted with caution and used as an exemplar to demonstrate the approaches until further work including the section 251 sample is possible.

Table 7-1: Distribution of IQ for each source of ID.

			IQ at age 8					IQ at age 15				
Source	Description	N with ID	N with IQ data	<70	70-84	≥85	Mean (SD)	N with IQ data	<70	70-84	≥85	Mean (SD)
Free text	Free text data in ALSPAC questionnaires indicating ID	82	34	12 (35.29)	13 (38.24)	9 (26.47)	75.74 (17.507)	27	10 (37.04)	11 (40.74)	6 (22.22)	73.81 (14.597)
SEN statement	Statement of special educational needs from the PLASC	297	76	22 (28.95)	28 (36.84)	26 (34.21)	80.17 (16.158)	46	15 (32.61)	17 (36.96)	14 (30.43)	78.41 (16.198)
GP records	A read code related to ID found in GP records	31	20	6 (30.00)	6 (30.00)	8 (40.00)	83.2 (24.054)	20	7 (35)	a	a	77.25 (16.457)
HES data	An ICD-9 or ICD-10 diagnosis recorded in HES data	≤5	≤5				63.33 (16.921)					a
MHSDS data	Multiple records indicating use of learning disability care services	≤5					a					a
IDI data	ICD-10 diagnosis identified	110	23	17 (73.91)	≤5	≤5	63.04 (12.115)	14	8 (57.14)	≤5	≤5	64.21 (10.772)

[a] Count too low to be presented. For columns indicating a count the value may be 0.

Table 7-2: Cross tabulation of ID obtained from each source.

ID Source	Total with data available	Total with ID from source	IQ < 70 at age 8	IQ < 85 at age 8	IQ < 70 at age 15	IQ < 85 at age 15	Free text	SEN statement	GP Read code	HES ICD 9/10 diagnosis	MHSDS code	IDI project ICD-10 diagnosis
IQ < 70 at age 8	6805	113		113 (100)	20 (17.70)	50 (44.25)	12 (10.62)	22 (19.47)	6 (5.31)	≤5	≤5	17 (15.04)
IQ < 85 at age 8	6805	804	113 (14.05)		68 (8.46)	271 (33.71)	25 (3.11)	50 (6.22)	12 (1.49)	≤5	≤5	22 (2.74)
IQ < 70 at age 15	4862	134	20 (14.93)	68 (50.75)		134 (100)	10 (7.46)	15 (11.19)	7 (5.22)	≤5	≤5	8 (5.97)
IQ < 85 at age 15	4862	1147	50 (4.36)	271 (23.63)	134 (11.68)		21 (1.83)	32 (2.79)	10 (0.87)	≤5	≤5	13 (1.13)
Free text	12084	82	12 (14.63)	25 (30.49)	10 (12.20)	21 (25.61)		13 (15.85)	8 (9.76)	≤5	≤5	27 (32.93)
SEN statement	9679	297	22 (7.41)	50 (16.84)	15 (5.05)	32 (10.77)	13 (4.38)		9 (3.03)	≤5	≤5	39 (13.13)
GP Read code	4646	31	6 (19.35)	12 (38.71)	7 (22.58)	10 (32.26)	8 (25.81)	9 (29.03)		≤5	≤5	10 (32.26)
HES ICD-9/10 diagnosis	4646	≤5	≤5	≤5	≤5	≤5	≤5	≤5	≤5		≤5	≤5
MHSDS code	58	≤5	≤5	≤5	≤5	≤5	≤5	≤5	≤5	≤5		≤5
IDI ICD-10 code	13489	110	17 (15.45)	22 (20.00)	8 (7.27)	13 (11.82)	27 (24.55)	39 (35.45)	10 (9.09)	≤5	≤5	

Percentages in each row are out of the total with ID from source

Where counts are ≤5, the count may be equal to 0

Table 7-3: Descriptive statistics across categories of consent.

Statistic	Explicit consenter	Section 251	Explicit non-consenter	No data linkage available
N	4646	7001	346	1503
Mean (SD) number of ID variables available [a]	4.30 (0.88)	3.26 (0.99)	3.30 (0.97)	2.02 (0.78)
Median (IQR) number of ID variables available [a]	5 (4-5)	3 (3-4)	3 (3-4)	2 (2-2)
N (%) with available IQ score at age 8	3678 (79.16)	2586 (36.94)	242 (69.94)	299 (19.89)
Mean (SD) IQ score at age 8	107.81 (15.97)	100.10 (15.96)	103.28 (17.54)	98.39 (15.57)
N (%) with available IQ score at age 15	3300 (71.03)	1303 (18.61)	164 (47.40)	96 (6.39)
Mean (SD) IQ score at age 15	96.16 (12.90)	90.56 (12.57)	93.16 (12.46)	87.88 (12.82)

[a] Sources of data included were IQ at age 8 and 15, free text data, SEN statement and diagnosis in the IDI project (i.e. excluded linked health data) in order to be able to compare across consenter status.

7.2 – Methods for assessing the association between maternal smoking during pregnancy and offspring ID

7.2.1 – Ethical approval

Ethical approval for the study (project B3010) was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees - http://www.bristol.ac.uk/alspac/researchers/research-ethics/.

7.2.2 – Definition of exposure to smoking during pregnancy

Maternal smoking is reported at several time points during pregnancy in ALSPAC. At 18 weeks gestation both mothers and partners were asked the following questions:

Question 1. Have you ever been a smoker?
Question 2. Have you now stopped smoking?
Question 3. How many times per day did you smoke in the first 3 months of your pregnancy? (For the partner questionnaire this question was worded as "How many times per day did you smoke at the start of your partner's pregnancy?" but treated equivalently)
Question 4. How many times per day did you smoke in the last 2 weeks?

This time point was selected because it is the only time point at which both maternal and partner smoking were reported, and it uses identical phrasing for all questions but one.
Binary measures of smoking in pregnancy were created for both mother and partner. Smokers were defined as having met Condition 1 or 2 below while a non-smoker was defined as having met neither condition provided data was available for at least one of the conditions:

Condition 1. Reported "yes" to Question 1 and had not reported "yes" to question 2 (i.e. reported "no" or left Question 2 missing).
Condition 2. The number of cigarettes reported per day in Question 3 or 4 was greater than 0.

A continuous measure of the number of times smoked per day obtained from question 4 (i.e. gestational weeks 16-18). This was captured in categories (0, 1-4, 5-9, 10-14, 15-19,20-24, 25-

29, 30+). The lowest number from each category was used to create a conservative estimate of maternal/partner self-report

ALSPAC also contains maternal report of partner smoking at 18 weeks gestation. Mothers were asked “Does your partner smoke?” and “If yes, about how many times per day does your partner smoke at the moment?” with responses captured in the same categories as for the mother’s own smoking (except with no option for 0 cigarettes smoked at the moment). This may cause issue as the partner may be a smoker but have given up smoking during the pregnancy period. Previous work, however, has shown high concordance between partner’s self-report and maternal report of partner smoking but low concordance for quantity of smoking [301]. Maternal report was therefore used to impute missing partner smoking status (but not quantity) where possible. Here there is a potential trade off in the negative control design between bias from increased error in the paternal report and bias from missing data.

7.2.3 – Covariate variable definitions

The variables used as covariates in models were the following:

1. Child sex assigned at birth
2. Maternal age at the time of birth
3. Maternal parity
4. Maternal depressive symptoms at 18 weeks gestation
5. Maternal alcohol use recorded at 18 weeks gestation
6. Maternal reported financial difficulties recorded at 32 weeks gestation
7. Maternal education recorded at 32 weeks gestation
8. Maternal occupational class recorded at 32 weeks gestation

Maternal age was grouped as under 25, 25-29, 30-34 and 35 or over. Parity was grouped as 0 (nulliparous), 1, or 2 or greater. Maternal depressive symptoms were measured using the Edinburgh Postnatal Depression Scale. A cut off of 12 was used to create a binary indicator as this has previously been shown to have a high correlation with depression [302]. Maternal alcohol use was included as a binary measure indicating the report of any alcohol use.

Financial difficulties during pregnancy were recorded on a scale from 0-15 with 15 indicating maximum financial difficulty. Mothers were asked “how difficult at the moment do you find it to afford these items?” for each of food, clothing, heating, rent or mortgage and things you will need for the baby. A score greater than 9 was used to indicate that financial difficulties were experienced during pregnancy. This score reflected the 10% of the sample that reported the greatest financial difficulty.

Maternal education was grouped as vocational, CSE/O level and A level/degree. Maternal occupational class was grouped into manual or non-manual occupation according to the definitions provided by the 1991 British Office of Population Census Surveys job codes) [303].

Categorical variables were used in the place of continuous variable for maternal age at birth, maternal depressive symptoms, and financial difficulties. This was done to allow for the possibility of non-linear associations between these variables and the outcome.

The covariates were grouped together for the purpose of investigating different model adjustment strategies. Maternal characteristics included maternal age at birth, parity, maternal depressive symptoms, and maternal alcohol use during pregnancy. Socioeconomic characteristics included financial difficulties, education, and occupational class. Child sex was included with maternal characteristics during adjustments.

Equivalent partner covariates were collected and derived in the same manner as for maternal covariates. Partner education, depression and alcohol use were all reported by the partner at 18 weeks gestation using the same questions/measures as for the mother. Partner occupational class was reported by the mother at the same time as reporting her own occupational class.

7.2.4 – Genotype information

Data on maternal genetic information was extracted from the ALSPAC genetic database [243]. Genotyping of ALSPAC mothers was performed using the Illumina human660W-quad at Centre National de G*énotypage* and genotypes were called using the Illumina GenomeStudio algorithm. Quality control was undertaken using PLINK (v1.07) on an initial set of 10,015 subjects and 557,124 directly genotyped SNPs. SNPs were removed if they had a proportion of missing data greater than 5%, had a Hardy-Weinberg-Equilibrium p-value lower than 10^{-6} or had a minor allele frequency less than 1%. Samples were excluded if they had greater than 5% missing data, had indeterminate X chromosome heterozygosity or had extreme autosomal heterozygosity. The quality control process is described in further detail elsewhere [304]. SNP imputation was carried out against the 1000 Genome Project database [305]. Quality control and SNP imputation were undertaken by the ALSPAC team before access was granted to the data.

7.2.5 – Observational analyses

All analyses were performed using R version 3.5.3 [306]. Logistic regression models of outcome (IQ at age 8 less than 70) on exposure were repeated for the binary measure of

smoking in pregnancy and the number of cigarettes smoked per day. Models were performed using four adjustment strategies: (i) unadjusted, (ii) adjusted for maternal characteristics (maternal age at birth, parity, maternal depressive symptoms, maternal alcohol use during pregnancy and child sex), (iii) adjusted for socioeconomic factors (financial difficulties, education, and occupational class) and (iv) adjusted for both maternal characteristics and socioeconomic factors.

7.2.6 – Negative control analyses

Logistic regression models of IQ at age 8 less than 70 on maternal and partner smoking during pregnancy, mutually adjusted for each other, were fitted using the same four adjustment strategies as for the observational analyses. The models were repeated for the binary and count forms of the exposure variable.

7.2.7 – Multiple imputation procedure

All observational analyses and negative control analyses were conducted as complete case analyses and repeated using multiply imputed data. MI was implemented in order to account for the substantial missing data in outcome and covariate variables and reduce consequent bias. Previous work has shown that, provided the data meet the MAR assumption, MI can produce unbiased results even at large proportions of missing data [157] (see Chapter 4).

Data were imputed using fully conditional specification (FCS) [156, 307]; this method is also commonly known as multivariate imputation using chained equations (MICE) and was carried out using the R package 'mice' [308] with 100 imputations and 5 iterations (to check convergence of parameter estimates). The exposure, outcome and all maternal and paternal covariates were included in the imputation model in order to maintain consistency between the imputation model and the most complex analysis model (the fully adjusted negative control model). Each variable was included as a predictor of all other variables. Binary variables (maternal and partner smoking status during pregnancy, financial difficulties, depression and alcohol use) were imputed using logistic regression, unordered categorical variables (maternal and partner occupational class) were imputed using multinomial logistic regression, ordered categorical variables (parity group and age group) were imputed using proportional odds models, and numeric variables (maternal and partner count exposure variables) were imputed using predictive mean matching.

As socioeconomic variables were often missing (see the missing data assessment in Section 7.3.2), two auxiliary variables for socioeconomic status were also included in the imputation

model. These were home ownership status and present maternal marital status which were both recorded at 6 weeks gestation. Homeownership status was grouped as (i) owned/mortgaged, (ii) rented from the council, (iii) privately rented, or (iv) other. Maternal marital status was grouped as (i) never married, (ii) previously married but not currently married (this included those who were widowed, separated and divorced), (iii) those who were in their first marriage, and (iv) those who were in their second or third marriage. Data on homeownership and marital status were available for 12,580 (93.3%) and 12,636 (93.7%) of the analysis sample respectively.

Missing data in the outcome variable, IQ at age 8 less than 70, was imputed using all exposure and covariate data. Indicator variables for ID from free text information, SEN statements and IDI project diagnoses were all included in the imputation model as auxiliary variables. Each of these three auxiliary variables had substantially less missing data than IQ at age 8, were likely to predict the missing IQ scores, thereby reducing error in the model, and most importantly were likely to predict the probability of having a missing value in IQ score, aiding the likelihood of meeting the MAR assumption and reducing bias from missing data [156].

The outcome was a binary variable derived from an underlying continuous variable. Several options for the imputation of this variable are available. The variable could be (i) imputed as a continuous variable and then dichotomised, (ii) imputed directly as a binary variable or (iii) the continuous form of the variable is imputed and then the binary form derived passively in the imputation procedure. It should be noted that for method (i) the imputation model is not compatible with the analysis model while for methods (ii) and (iii) the imputation and analysis models are compatible. Compatibility with the analysis model is widely regarded as a requirement for valid MI estimates [309, 310]. However, a recent simulation study has shown that for outcome variables, imputing as a continuous variable, then dichotomising (i.e. method (i)) leads to less bias than dichotomising then imputing (method (ii)), particularly as the quantity of missing data increases [311]. Passive imputation (method (iii)) has the advantage over the other methods in that it imputes the continuous measure, and so should be unbiased, while maintaining compatibility between the imputation and analysis models. Neither method (ii) or (iii) would converge in practice however and therefore I imputed IQ at age 8 as a continuous variable using a linear regression model and dichotomised post imputation.

To test whether lack of compatibility between imputation and analysis models for the outcome variable substantially altered conclusions, a sensitivity analysis across the three different methods of imputation was conducted. For each method I used a simplified version of the imputation model to aid in convergence of the models. The simplified imputation model included only the outcome, the maternal binary exposure, the maternal covariate variables and the auxiliary variables. The observational analyses were then repeated for each outcome imputation method.

In contrast to the findings for a dichotomised outcome variable, deriving a binary covariate variable from a continuous variable post imputation can lead to substantial bias in effect estimates [312]. Imputing the binary form of the variable only in the imputation model was shown not to exhibit such bias. I therefore imputed covariate variables in the form that the were to be used in the analysis model. This was done under the assumption that the findings for binary variables extrapolate to unordered and ordered categorical variables, which has not been investigated so far.

7.2.8 – Mendelian randomisation analyses

Two-sample MR [173, 174] was used to investigate the causal nature of the association between maternal smoking during pregnancy and offspring intellectual disability. The two-sample MR framework combines summary statistics for SNP-exposure associations and for SNP-outcome associations. Summary statistics for smoking initiation and number of cigarettes smoked per day were obtained from analyses of the GWAS and Sequencing Consortium of Alcohol and Nicotine (GSCAN) [168]. SNPs below a p-value threshold of 10^{-8} were extracted to proxy for these respective exposure phenotypes.

To my knowledge, no GWAS has so far been performed that assesses the association between maternal genetic variants and offspring phenotypic ID. To produce summary statistics for SNP-outcome associations, logistic regression models of offspring ID (defined as IQ at age 8 less than 70) were fitted against the number of copies of maternal alleles for a given SNP using ALSPAC data. Regression models were fitted separately for each SNP identified in the GSCAN study. Data were restricted to mothers of European ancestry to prevent potential bias from population stratification. Each model was adjusted for offspring sex and the first 20 principal components. Imputation of missing outcome data was not performed in the GWAS analyses.

The SNP-exposure and SNP-outcome summary statistics were harmonised to ensure that the same strand of DNA (and therefore the same allele) was used in each association. Palindromic SNPs - genetic variants in which the alleles are also corresponding nucleotide pairs - can negatively impact harmonisation. For example, a variant with alleles adenine and thymine (A/T) on one strand of DNA will correspond to the same pair of nucleotides on the opposing strands, making it difficult to determine the reference strand used in each association. In these situations, the effect allele frequencies were used to inform which strand was used for palindromic SNPs by manually checking whether the frequency in each data source was approximately equal. However, it is not possible to determine the reference strand if an effect allele frequency is close to 50% [174] – the allele frequency on the opposing strand will be too similar to confidently determine which strand was the reference. Therefore, all palindromic SNPs in which either

GSCAN or ALSPAC had an effect allele frequency between 40% and 60% were removed to avoid using SNPs with an ambiguous effect allele. Allowing such alleles to remain in analyses may bias estimates if the incorrect allele is identified as the reference allele.

Analyses were conducted using R code developed as part of the MR-Base project [313]. The causal effect of smoking behaviour (repeated for smoking initiation and number of cigarettes smoked per day) on offspring ID was obtained for each SNP identified in GSCAN as having $p<10^{-8}$ using a ratio of the beta coefficient from the SNP-outcome association relative to the beta coefficient from the SNP-exposure association. This is a ratio of the mean change in log-odds of ID with each allele for a given SNP, relative to the mean change in units of smoking behaviour per allele for the SNP (the units of smoking behaviour were risk of smoking initiation and number of cigarettes smoked per day). An average effect across all SNPs was obtained using the inverse variance weighted (IVW) method [173].

In sensitivity analyses MR-Egger [175, 176] weighted median [177, 178] and weighted mode [179] estimators were fit to check for consistency with the IVW estimate. Deviation from the IVW estimate would indicate that the MR assumptions described in Chapter 2 (Section 2.3.2) may have been violated.

7.3 – Results

7.3.1 – Cohort descriptives

Descriptives of the cohort separated by maternal and partner smoking status during pregnancy are presented in Table 7-4. The table shows that 25.2% of participants were exposed to maternal smoking during pregnancy, 68.9% were not exposed and 5.83% had no exposure data available. Maternal smokers were more likely to be younger, have prenatal depression symptoms, have used alcohol during pregnancy, have a lower level of education, have a manual occupation and to have experienced financial difficulties during pregnancy.

Exposure to partner smoking was more common than to maternal smoking (36% vs. 25%), though was more often missing . Partners tended to smoke more cigarettes per day than mothers if they did smoke (median number smoked [inter quartile range]: 10 [5-20] vs 5 [0-10]). The overall pattern of confounder distributions between smokers and non-smokers were similar between mothers and partners for most characteristics; smokers tended to be younger, have depression during pregnancy, have a lower education, a manual occupation and have experienced financial difficulties. The actual distributions were not similar, however, as partners tended to be older than mothers, were more likely to have been educated to A/level or degree

standard, and work a manual job. It is unclear whether this disparity is due to actual differences in the distributions or due to substantially lower responses from partners than mothers. Partners were much less likely to respond to questions on alcohol consumption and depression. The number of partners reporting depression were substantially lower for partners than mothers while alcohol use was more common among partners. A cross tabulation of maternal and partner smoking is presented in Table 7-5, which shows that there is positive assortative mating between parents (I_{PSI} = 0.41; see Chapter 3 for further details on I_{PSI}).

Table 7-4: Descriptive statistics of the cohort separated by maternal smoking status

	Maternal			Partner		
	Non-smoker	Smoker	Missing data	Non-smoker	Smoker	Missing data
Total, N (%)	9298 (68.93)	3404 (25.24)	787 (5.83)	7453 (55.25)	4914 (36.43)	1122 (8.32)
Number of times smoked per day, Median (IQR)	-	5 (0-10)	-	-	10 (5-20)	-
Child Sex, N (%)						
Female	4574 (49.19)	1593 (46.8)	367 (46.63)	3637 (48.8)	2373 (48.29)	524 (46.7)
Male	4724 (50.81)	1811 (53.2)	420 (53.37)	3816 (51.2)	2541 (51.71)	598 (53.3)
Parity, N (%)						
0	4119 (44.3)	1491 (43.8)	0 (0)	3304 (44.33)	2102 (42.78)	204 (18.18)
1	3315 (35.65)	1029 (30.23)	0 (0)	2660 (35.69)	1614 (32.84)	70 (6.24)
>= 2	1741 (18.72)	786 (23.09)	0 (0)	1377 (18.48)	1088 (22.14)	62 (5.53)
Missing	123 (1.32)	98 (2.88)	787 (100)	112 (1.5)	110 (2.24)	786 (70.05)
Age, N (%)						
< 25	1609 (17.3)	1321 (38.81)	344 (43.71)	324 (4.35)	441 (8.97)	40 (3.57)
25-39	3742 (40.25)	1202 (35.31)	263 (33.42)	1557 (20.89)	911 (18.54)	56 (4.99)
30-34	2895 (31.14)	647 (19.01)	127 (16.14)	1786 (23.96)	761 (15.49)	39 (3.48)
>= 35	1052 (11.31)	234 (6.87)	53 (6.73)	1113 (14.93)	559 (11.38)	24 (2.14)
Missing	0 (0)	0 (0)	0 (0)	2673 (35.86)	2242 (45.62)	963 (85.83)
Depression, N (%)						
No	7682 (82.62)	2416 (70.98)	0 (0)	5728 (76.85)	3407 (69.33)	≤5
Yes	933 (10.03)	686 (20.15)	0 (0)	174 (2.33)	208 (4.23)	≤5
Missing	683 (7.35)	302 (8.87)	787 (100)	1551 (20.81)	1299 (26.43)	1119 (99.73)
Alcohol use in pregnancy, N (%)						
No	4366 (46.96)	1332 (39.13)	0 (0)	275 (3.69)	181 (3.68)	≤5
Yes	4824 (51.88)	2034 (59.75)	0 (0)	5544 (74.39)	3364 (68.46)	≤5
Missing	108 (1.16)	38 (1.12)	787 (100)	1634 (21.92)	1369 (27.86)	1119 (99.73)
Education, N (%)						
Vocational	772 (8.3)	376 (11.05)	37 (4.7)	418 (5.61)	365 (7.43)	≤5
CSE/O level	4385 (47.16)	1988 (58.4)	190 (24.14)	2235 (29.99)	1943 (39.54)	a
A level/ Degree	3576 (38.46)	605 (17.77)	55 (6.99)	3229 (43.32)	1234 (25.11)	≤5

	Maternal			Partner		
	Non-smoker	Smoker	Missing data	Non-smoker	Smoker	Missing data
Missing	565 (6.08)	435 (12.78)	505 (64.17)	1571 (21.08)	1372 (27.92)	1097 (97.77)
Occupation, N (%)						
Non-manual	6169 (66.35)	1505 (44.21)	101 (12.83)	4152 (55.71)	1650 (33.58)	91 (8.11)
Manual	1200 (12.91)	685 (20.12)	49 (6.23)	2363 (31.71)	2152 (43.79)	151 (13.46)
Missing	1929 (20.75)	1214 (35.66)	637 (80.94)	938 (12.59)	1112 (22.63)	880 (78.43)
Financial difficulties, N (%)						
No	7904 (85.01)	2397 (70.42)	192 (24.4)	6420 (86.14)	3677 (74.83)	396 (35.29)
Yes	629 (6.76)	497 (14.6)	44 (5.59)	466 (6.25)	598 (12.17)	106 (9.45)
Missing	765 (8.23)	510 (14.98)	551 (70.01)	567 (7.61)	639 (13)	620 (55.26)

a - Data greater than 5 but cannot be presented

Note that values labelled as ≤5 may include 0

Table 7-5: Cross tabulation of maternal and partner smoking during pregnancy

		Maternal smoking status, N (%)	
		Non-smoker	Smoker
Partner smoking status	Non-smoker	6344 (71.24)	9255 (30.08)
	Smoker	2561 (28.76)	2150 (69.92)

IPSI value = 0.41

7.3.2 – Missing data assessment for complete case analysis

The distributions of exposure, outcome and covariate variables among participants with no missing data in any variables as compared to participants with data missing in at least one variable are presented in Table 7-6. Using complete case analysis would lose 61% of observations to missing data in any variable for observational analyses. This increased to 75% when paternal variables were included, as in the negative control analyses.

Table 7-6 can be used to identify associations with the probability of being a missing value and therefore guide understanding of the likely biases that would arise in complete case analyses. Odds ratios for missing data in any variable were produced using unadjusted logistic regression models. IQ <70 at age 8 was associated with increased odds of being a missing value. Smokers and heavier smokers were also more likely to be excluded from complete case analyses for missing data. Younger mothers, those with greater parity, those with depression, and those with a manual occupation were also more likely to have missing data. Those who used alcohol and those who had A levels or a degree were less likely to have missing data in any variable.

There were some differences between mothers and partners in the strength of the relationship between variables and being excluded from complete cases analyses for missing data in the negative control analyses. Mothers were more likely than partners to be excluded for missing data in another variable if they were smokers (maternal OR [95 % CI] = 2.53 [2.28-2.81], partner OR [95 % CI] = 1.94 [1.78-2.11]). Partners were less likely to have missing data if they drank alcohol during pregnancy whereas for mothers there was no association between alcohol consumption and exclusion from complete case analyses. The association with missing data was similar between mothers and partners for all other covariates.

Table 7-7 shows the number of observations with missing data in each variable separated by the number of variables missing. Data are presented for observations with at least one missing value (i.e. only observations excluded from complete case analysis). This table can be used to show the co-occurrence of missing data in multiple variables. Of those excluded from complete case analysis maternal smoking was missing 4% of the time and was never missing on its own. No IQ information was available for 37% of those excluded from complete case analysis. Socioeconomic variables were often missing (ranging from 8% for maternal education to 21% for occupation) and were commonly missing together. Maternal depression, alcohol use and child parity were also frequently missing.

Table 7-6: Descriptive statistics for missing data in any variable for observational and negative control analyses

	Observational analysis			Negative control analysis		
	Non-missing, N (%)	Missing, N (%)	OR (95% CI) [a]	Non-missing, N (%)	Missing, N (%)	OR (95% CI) [a]
Total	5307 (39.34)	8182 (60.66)	-	3393 (25.15)	10096 (74.85)	-
Outcome						
IQ at age 8 < 70	65 (1.22)	48 (3.20)	2.67 (1.83-3.89)	29 (0.85)	84 (2.46)	2.93 (1.91-4.48)
Exposure						
Maternal smoking (Yes)	923 (17.39)	2481 (33.55)	2.40 (2.20-2.61)	513 (15.12)	2891 (31.06)	2.53 (2.28-2.81)
Maternal smoking quantity among smokers [b]	5 (0-10)	5 (0-15)	1.08 (1.07-1.09)	3 (0-10)	5 (0-10)	1.09 (1.08-1.11)
Partner smoking (Yes)				976 (28.77)	3938 (43.88)	1.94 (1.78-2.11)
Partner smoking quantity among smokers [b]				10 (1-15)	10 (5-20)	1.04 (1.04-1.05)
Covariates						
Child sex						
Female	2638 (49.71)	3896 (47.62)	Ref	1688 (49.75)	4846 (48.00)	Ref
Male	2669 (50.29)	4286 (52.38)	1.09 (1.01-1.17)	1705 (50.25)	5250 (52.00)	1.07 (0.99-1.16)
Parity						
0	2651 (49.95)	2959 (41.25)	Ref	1784 (52.58)	3826 (42.10)	Ref
1	1873 (35.29)	2471 (34.44)	1.18 (1.09-1.28)	1162 (34.25)	3182 (35.01)	1.28 (1.17-1.39)
>= 2	783 (14.75)	1744 (24.31)	2.00 (1.81-2.20)	447 (13.17)	2080 (22.89)	2.17 (1.93-2.44)
Maternal age						
<25	687 (12.95)	2587 (31.62)	Ref	418 (12.32)	2856 (28.29)	Ref
25-29	2151 (40.53)	3056 (37.35)	0.38 (0.34-0.42)	1386 (40.85)	3821 (37.85)	0.40 (0.36-0.45)
30-34	1810 (34.11)	1859 (22.72)	0.27 (0.25-0.30)	1181 (34.81)	2488 (24.64)	0.31 (0.27-0.35)
>= 35	659 (12.42)	680 (8.31)	0.27 (0.24-0.31)	408 (12.02)	931 (9.22)	0.33 (0.29-0.39)
Partner age						
<25				188 (5.54)	617 (14.63)	Ref
25-29				1087 (32.04)	1437 (34.07)	0.40 (0.34-0.48)
30-34				1181 (34.81)	2488 (24.64)	0.31 (0.27-0.35)
>= 35				822 (24.23)	874 (20.72)	0.32 (0.27-0.39)
Maternal Depression						
No	4755 (89.60)	5343 (83.35)	Ref	3069 (90.45)	7029 (84.44)	Ref

	Observational analysis			Negative control analysis		
	Non-missing, N (%)	Missing, N (%)	OR (95% CI) [a]	Non-missing, N (%)	Missing, N (%)	OR (95% CI) [a]
Yes	552 (10.40)	1067 (16.65)	1.72 (1.54-1.92)	324 (9.55)	1295 (15.56)	1.75 (1.53-1.99)
Partner Depression						
No				3293 (97.05)	5845 (95.40)	Ref
Yes				100 (2.95)	282 (4.60)	1.59 (1.26-2.00)
Maternal Alcohol Use						
No	2323 (43.77)	3375 (46.56)	Ref	1496 (44.09)	4202 (45.86)	Ref
Yes	2984 (56.23)	3874 (53.44)	0.89 (0.83-0.96)	1897 (55.91)	4961 (54.14)	0.93 (0.86-1.01)
Partner Alcohol Use						
No				95 (2.80)	361 (6.04)	Ref
Yes				3298 (97.20)	5613 (93.96)	0.45 (0.36-0.56)
Maternal Race						
Non-White	86 (1.63)	225 (3.40)	Ref	42 (1.24)	269 (3.16)	Ref
White	5203 (98.37)	6383 (96.60)	0.47 (0.36-0.60)	3342 (98.76)	8244 (96.84)	0.39 (0.28-0.53)
Partner Race						
Non-White				43 (1.27)	230 (3.82)	Ref
White				3335 (98.73)	5789 (96.18)	0.32 (0.23-0.45)
Maternal Occupation						
Non-Manual	4471 (84.25)	3304 (75.06)	Ref	2921 (86.09)	4854 (76.85)	Ref
Manual	836 (15.75)	1098 (24.94)	1.78 (1.61-1.97)	472 (13.91)	1462 (23.15)	1.86 (1.66-2.09)
Partner Occupation						
Non-Manual				2277 (67.11)	3616 (50.46)	Ref
Manual				1116 (32.89)	3550 (49.54)	2.00 (1.84-2.18)
Maternal Education						
Vocational	435 (8.20)	750 (11.23)	Ref	272 (8.02)	913 (10.63)	Ref
CSE/ O level	2412 (45.45)	4151 (62.17)	1.00 (0.88-1.13)	1448 (42.68)	5115 (59.54)	1.05 (0.91-1.22)
A level/Degree	2460 (46.35)	1776 (26.60)	0.42 (0.37-0.48)	1673 (49.31)	2563 (29.83)	0.46 (0.39-0.53)
Partner Education						
Vocational				224 (6.60)	559 (9.23)	Ref
CSE/ O level				1230 (36.25)	2972 (49.08)	0.97 (0.82-1.15)
A level/Degree				1939 (57.15)	2525 (41.69)	0.52 (0.44-0.62)
Financial Difficulties						

	Observational analysis			Negative control analysis		
	Non-missing, N (%)	Missing, N (%)	OR (95% CI) [a]	Non-missing, N (%)	Missing, N (%)	OR (95% CI) [a]
No	4954 (93.35)	5539 (87.15)	Ref	3207 (94.52)	7286 (88.10)	Ref
Yes	353 (6.65)	817 (12.85)	2.07 (1.82-2.36)	186 (5.48)	984 (11.90)	2.33 (1.98-2.74)

[a] ORs for being a missing value across levels of the variable

[b] Values presented for this row are median (IQR) and OR represents the change in odds of being a missing value for a 1 cigarette per day increase among smokers only

Table 7-7: Counts of missing data in each variable separated by the number of missing variables

	Number of missing variables, N (%) [a]								
Missing variable	1	2	3	4	5	6	7	8	Total
Total	4549	1446	437	839	254	121	97	439	8182
IQ at age 8	3561 (78.28)	1342 (92.81)	198 (45.31)	764 (91.06)	240 (94.49)	104 (85.95)	36 (37.11)	439 (100.00)	6684 (36.53)
Maternal smoking	0 (0.00)	0 (0.00)	0 (0.00)	42 (5.01)	119 (46.85)	90 (74.38)	97 (100.00)	439 (100.00)	787 (4.30)
Child sex	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Parity [b]	70 (1.54)	82 (5.67)	34 (7.78)	51 (6.08)	138 (54.33)	97 (80.17)	97 (100.00)	439 (100.00)	1008 (5.51)
Maternal age	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Maternal depression	212 (4.66)	283 (19.57)	199 (45.54)	203 (24.20)	219 (86.22)	120 (99.17)	97 (100.00)	439 (100.00)	1772 (9.69)
Maternal alcohol use	15 (0.33)	36 (2.49)	28 (6.41)	57 (6.79)	145 (57.09)	116 (95.87)	97 (100.00)	439 (100.00)	933 (5.10)
Maternal occupation	664 (14.60)	1090 (75.38)	426 (97.48)	797 (94.99)	148 (58.27)	119 (98.35)	97 (100.00)	439 (100.00)	3780 (20.66)
Maternal education	8 (0.18)	24 (1.66)	155 (35.47)	658 (78.43)	126 (49.61)	33 (27.27)	62 (63.92)	439 (100.00)	1505 (8.23)
Financial difficulties	19 (0.42)	35 (2.42)	271 (62.01)	784 (93.44)	135 (53.15)	47 (38.84)	96 (98.97)	439 (100.00)	1826 (9.98)

[a] Percentage for each column sums to 100 × the number of missing variables.

[b] Example table interpretation: 4549 individuals have data missing in 1 variable only, 70 of those have data missing in the parity variable. 1446 individuals have data missing in 2 variables, for 82 individuals, one of those variables is parity.

7.3.3 – Observational analysis

The results of the observational analyses for both binary and count exposure are presented in Table 7-8. Both complete case analysis and MI analysis found unadjusted associations of increased odds of ID among those exposed to maternal smoking during pregnancy. Adjustment for maternal characteristics and socioeconomic characteristics led to attenuation in the association. This attenuation was greater for socioeconomic characteristics, suggesting socioeconomic position plays a larger role in confounding the association.

The results for the count exposure mirrored those found for the binary exposure. In MI analyses offspring were 1.05 times as likely to have ID for each additional cigarette their mother smoked per day during pregnancy. This association attenuated once maternal and socioeconomic characteristics had been accounted for. Again, the attenuation following adjustment for socioeconomic characteristics was greater than for maternal characteristics.

Table 7-8: Results of the observational analyses using a binary exposure.

	OR (95% CI)	
Model	Complete case analysis	Multiple imputation analysis
Binary exposure [a]		
Unadjusted	1.43 (0.80-2.56)	1.85 (1.38-2.40)
Adjusted for maternal characteristics	1.30 (0.72-2.37)	1.35 (0.98-1.86)
Adjusted for socioeconomic characteristics	1.04 (0.57-1.89)	1.24 (0.90-1.69)
Adjusted for all confounders	1.05 (0.57-1.92)	1.10 (0.79-1.52)
Count exposure (cigarettes per day) [b]		
Unadjusted	1.03 (0.97-1.09)	1.05 (1.03-1.07)
Adjusted for maternal characteristics	1.02 (0.96-1.08)	1.03 (1.00-1.05)
Adjusted for socioeconomic characteristics	1.00 (0.94-1.06)	1.02 (0.99-1.05)
Adjusted for all confounders	1.00 (0.94-1.06)	1.01 (0.99-1.04)

[a] Binary exposure: N for complete case analysis = 5,307, N for multiple imputation analysis = 13,489

[b] Count exposure: N for complete case analysis = 5,298, N for multiple imputation analysis = 13,489

Table 7-9: Results of the sensitivity analyses for the different methods of imputing the outcome.

	Impute then dichotomise		Impute as binary		Passive imputation	
Model	OR (95% CI)	FMI	OR (95% CI)	FMI	OR (95% CI)	FMI
Binary exposure [a]						
Unadjusted	1.72 (1.23-2.40)	0.384	1.62 (1.06-2.47)	0.602	1.73 (1.24-2.42)	0.392
Adjusted for maternal characteristics	1.31 (0.92-1.89)	0.414	1.32 (0.85-2.04)	0.587	1.32 (0.93-1.88)	0.396
Adjusted for socioeconomic characteristics	1.07 (0.74-1.54)	0.407	1.02 (0.63-1.64)	0.639	1.09 (0.75-1.57)	0.423
Adjusted for all confounders	0.98 (0.67-1.43)	0.421	0.98 (0.61-1.57)	0.622	0.99 (0.68-1.44)	0.421

[a] N for all multiple imputation analysis = 13,489

The sensitivity analyses for the different methods of imputing the outcome are presented in Table 7-9. Imputing the outcome as continuous then dichotomising (as is done in the primary analysis) provided consistent estimates with passive imputation suggesting that incompatibility of the imputation and analysis model has not substantially influenced the conclusions of the primary analyses. The results of these sensitivity analyses, which used a simplified imputation model, were all slightly attenuated towards the null compared to the primary analyses. Imputing the outcome as binary directly lead to greater attenuation than for the other models and also wider confidence intervals. The wider confidence intervals seem to be attributable to greater between imputation variance as evidenced by the greater FMI than for the other two methods (see the Chapter 4 for an explanation of the FMI).

7.3.4 – Negative control analysis

In complete case analyses the effect estimates for both the binary and count forms of the exposure suggested that maternal smoking was associated with increased odds of ID in offspring while partner smoking was protective (see Table 7-10). The maternal estimate was approximately twice the size of the partner estimate for the unadjusted and all adjusted models for the binary exposure. The ratio of maternal to paternal estimates was even greater for the count form of the exposure. These estimates, if true, would suggest a causal role of maternal smoking during pregnancy.

The MI analyses produced estimates that were in contrast to those of the complete case analyses. Both maternal and partner smoking during pregnancy were found to be associated with increased odds of ID, though both associations were attenuated following adjustment for

maternal and partner covariates. The effect estimates for maternal and partner smoking were approximately equal in all models. This instead suggests a non-causal role of maternal smoking in pregnancy on offspring ID and suggests that the results found using complete case analysis were the result of selection bias due to missing data.

Table 7-10: Results of the negative control analyses.

	Complete case analysis		Multiple imputation analysis	
	Maternal OR	Paternal OR	Maternal OR	Paternal OR
Model	(95% CI)	(95% CI)	(95% CI)	(95% CI)
Binary exposure [a]				
Unadjusted - complete case data	1.71	0.82	1.59	1.47
	(0.69-4.24)	(0.36-1.90)	(1.15-2.18)	(1.10-1.98)
Adjusted for maternal characteristics	1.53	0.73	1.20	1.24
	(0.61-3.82)	(0.32-1.70)	(0.85-1.68)	(0.92-1.68)
Adjusted for socioeconomic characteristics	1.17	0.58	1.08	1.07
	(0.47-2.94)	(0.24-1.36)	(0.78-1.50)	(0.79-1.44)
Fully adjusted for confounders	1.19	0.55	0.99	1.03
	(0.47-2.98)	(0.23-1.30)	(0.71-1.39)	(0.76-1.39)
Count exposure (cigarettes per day) [b]				
Unadjusted - complete case data	2.21	0.82	1.03	1.03
	(1.01-4.81)	(0.45-1.49)	(1.01-1.06)	(1.01-1.04)
Adjusted for maternal characteristics	1.96	0.77	1.01	1.02
	(0.88-4.39)	(0.42-1.40)	(0.99-1.04)	(1.00-1.03)
Adjusted for socioeconomic characteristics	1.67	0.61	1.01	1.01
	(0.74-3.76)	(0.32-1.14)	(0.98-1.04)	(0.99-1.02)
Fully adjusted for confounders	1.61	0.61	1.00	1.00
	(0.71-3.67)	(0.33-1.15)	(0.98-1.03)	(0.99-1.02)

All models are mutually adjusted for maternal and paternal exposure

[a] Binary exposure: N for complete case analysis = 3393, N for multiple imputation analysis = 13489

[b] Count exposure: N for complete case analysis = 3343, N for multiple imputation analysis = 13489

7.3.5 – Mendelian randomisation analysis

No association between SNP and ID met genome wide significance. The average number of observations available in models was 4218.6 (SD = 209.6; range = 2,562-4,346).

The SNP rs76460663, which was found to be associated with smoking initiation in the GSCAN study, was removed from further analysis as the ALSPAC GWAS produced an effect estimate for the association with ID that was over 2400% bigger than the next largest effect estimate and was measured with substantial error (rs76460663 log odds ratio=15.3, SE=640.8). The SNP was looked up on the National Human Genome Research Institute's GWAS catalogue (URL: https://www.ebi.ac.uk/gwas/variants/rs76460663) which showed that it has only been identified as a top hit in the GSCAN study, and not in any other GWAS.

A forest plot of MR ratio estimates of the SNP-outcome association to the SNP-exposure association, sorted on effect size, for the associations of ID with SNPs for smoking initiation and with SNPs for number of cigarettes smoked per day is presented in Figure 7-2. Every eighth SNP is presented for smoking initiation due to the number of SNPs tested. The plot shows substantial overlap in confidence intervals for all SNPs tested, suggesting that there was little heterogeneity in effects. This is supported by the p-value for Cochran's Q statistic [314] being greater than 0.1 which was also the case for the collation of results for each combination of smoking behaviour and outcome definition (smoking initiation, p=0.897; cigarettes per day, p=0.539).

Combined statistics across all SNPs for the influence of each smoking behaviour on ID are presented in Table 7-11 (see Figure 7-3 for plots of individual SNP-outcome associations against SNP-exposure associations with fitted lines for each MR method). The IVW estimate showed little or no evidence of an effect of smoking initiation by the mother on offspring ID. Each of the sensitivity analyses (MR-Egger, weighted median and weighted mode) were consistent with the interpretation of the IVW method. For number of cigarettes smoked per day, the IVW estimate and all sensitivity analyses also showed no evidence for an effect of maternal smoking behaviour on offspring risk of ID. The confidence intervals of these estimates were very wide showing that error in the models was high.

Table 7-11: MR primary and sensitivity analysis estimates

		Smoking initiation		Cigarettes per day	
Test	Parameter	Log OR (95% CI)	p-value	Log OR (95% CI)	p-value
IVW	beta	-0.24 (-1.43 - 0.95)	0.691	2.55 (-0.75 - 5.84)	0.126
MR-Egger	beta	-3.46 (-8.90 - 1.98)	0.211	2.36 (-6.50 - 11.22)	0.594
MR-Egger	alpha	0.06 (-0.04 - 0.17)	0.233	0.00 (-0.18 - 0.19)	0.963
Weighted median	beta	-0.44 (-2.17 - 1.34)	0.636	2.31 (-2.37 - 7.43)	0.356
Weighted mode	beta	-1.49 (-7.80 - 4.82)	0.643	2.07 (-5.10 - 9.25)	0.574

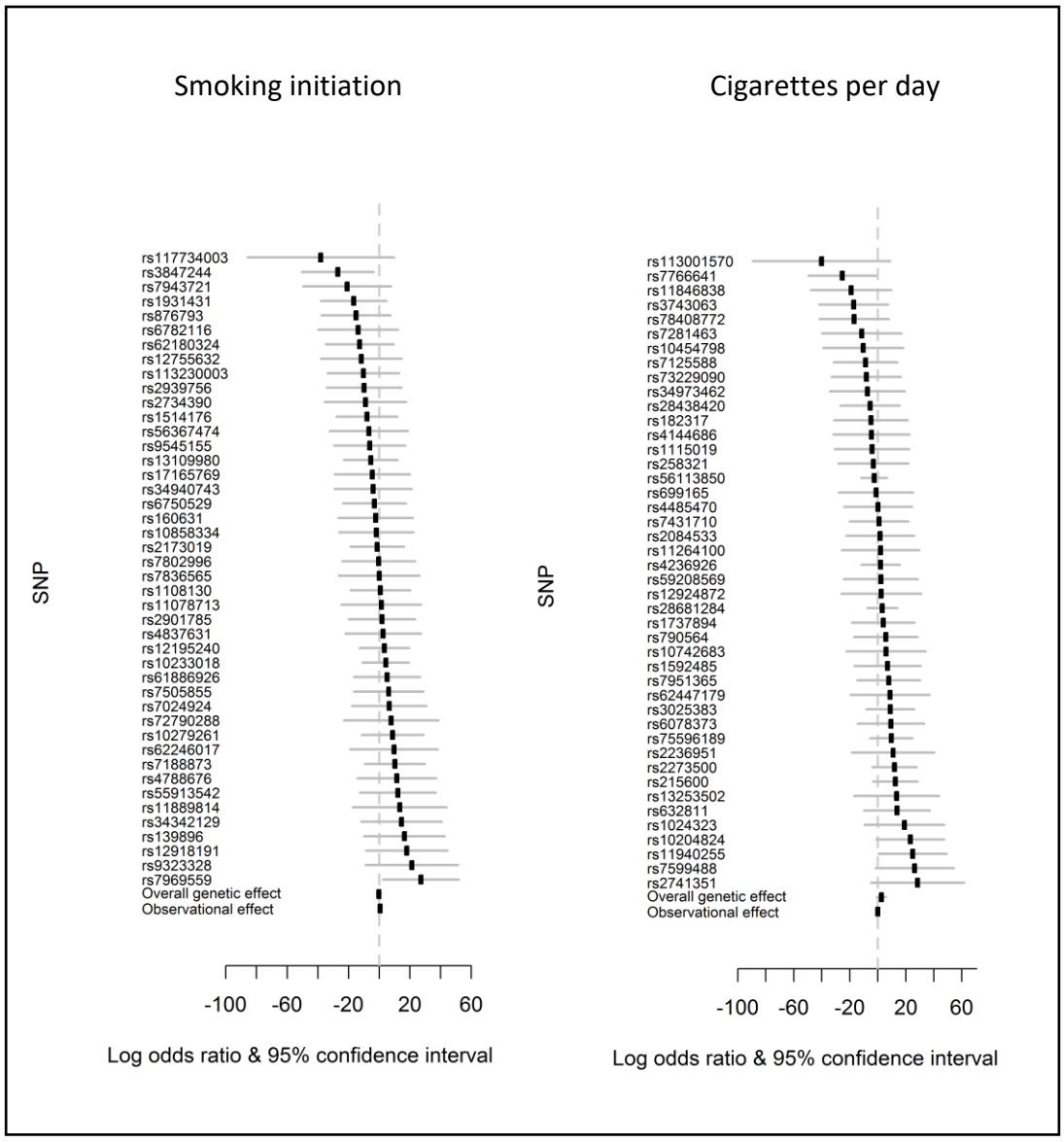

Figure 7-2: Two-sample MR estimates based on the ratio method for individual SNPs. Plots are for smoking initiation (left) and cigarettes per day (right) for the outcome possible ID.

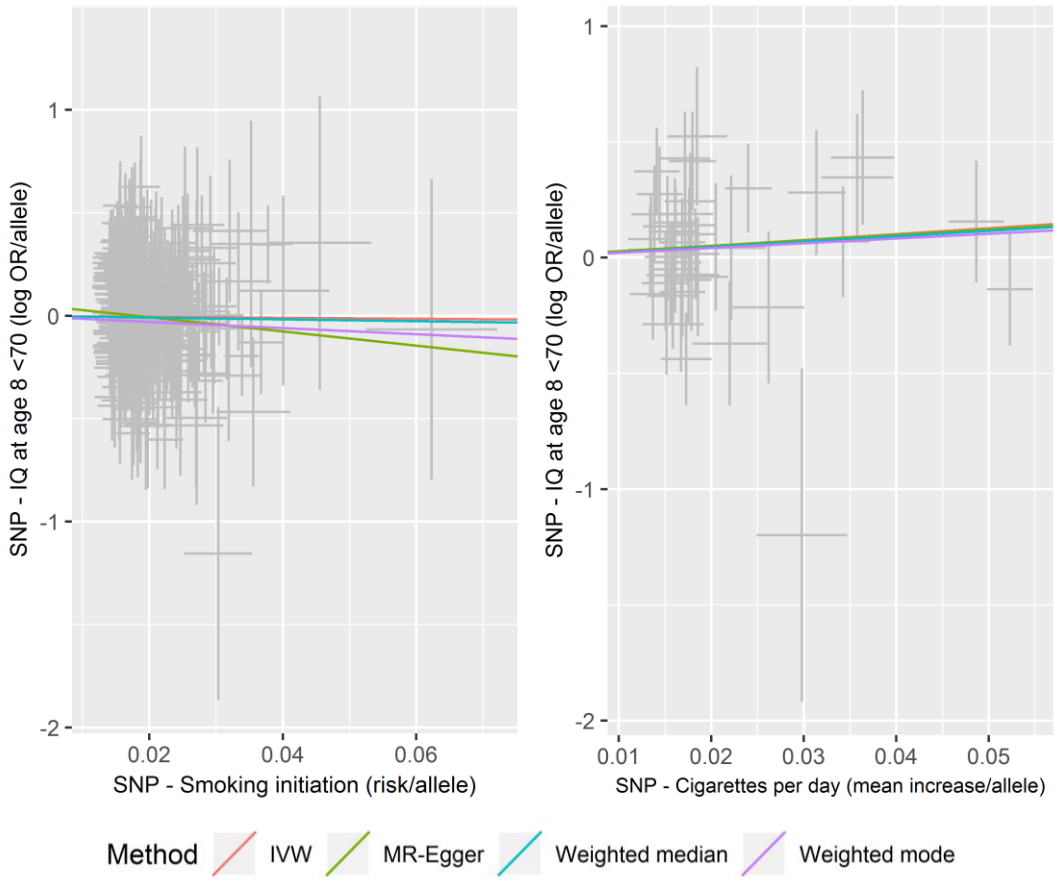

Figure 7-3: Plots of the SNP-outcome association against SNP-exposure association for each smoking behaviour (left: smoking initiation; right: number of cigarettes smoked per day) and outcome. Fitted lines show the estimates of the different MR methods that collate effects across all SNPs.

7.4 – Discussion

Both observational and negative control designs suggested that any association between maternal smoking during pregnancy and offspring ID is driven by confounding. The observational analyses found unadjusted associations that were largely attenuated by adjustment for socioeconomic factors and to a lesser extent by maternal characteristics including alcohol use and depression. Once selection bias from missing data had been accounted for by MI, the negative control design suggested that maternal and paternal effect estimates were equivalent in all unadjusted and adjusted models, providing evidence against a causal role of maternal smoking in pregnancy in the development of ID. These findings are consistent with previous literature that shows that maternal smoking during pregnancy is associated with offspring ID (e.g. the studies by Drews et al. [90] and Mann et al. [96]) but that once bias from confounding has been adequately accounted for this association is attenuated to a null effect (as evidenced by Braun et al. [98] and Lundberg et al. [99]).

Evidence from MR models was less clear. All confidence intervals suggested that the effect estimates were consistent with a null effect. However, the estimate provided by analysis that combined across genetic variants suggested an increase in the risk of ID (as measured by IQ at age 8 less than 70) with each cigarette smoked per day. This was supported by sensitivity analyses that relaxed different aspects of the MR assumptions and analyses that showed more extreme effects among the offspring of mothers who actually smoked during pregnancy. Given the findings of the other analysis methods in this study, the somewhat ambiguous findings of the MR models are likely to reflect a null effect.

The current study highlights that while causal inference methods are able to account for biases in a way that observational analyses can not, they are still susceptible to their own biases that can result in misleading conclusions. In the negative control analyses, complete case analysis found a greater risk of ID for the maternal coefficient than the partner coefficient. The partner coefficient seemed to suggest a protective effect of partner smoking during pregnancy. This pattern of results is likely the result of selection bias from missing data. Partner exposure and covariate data was missing more often than maternal data which may suggest that the causes of missing data, or at least the strength of association, may differ between parents which in turn could lead to differing directions of bias for the two coefficients. Here MI was important for removing bias from missing data and leading to a conclusion that maternal and partner smoking coefficients were equivalent, providing evidence against a causal role of smoking in pregnancy on the development of ID.

7.4.1 – Strengths and limitations

This study benefited from multiple approaches being applied in the same dataset which can aid in triangulation where results are in agreement with each other. Here, comparison across methodologies enabled inference to the most likely conclusion where one of the three did not provide clear evidence for a single conclusion.

The ALSPAC dataset provided a rare opportunity to explore cross generational associations between maternal genetic variants and offspring outcomes. No GWAS of maternal genetics and offspring ID outcomes was available to be used in two-sample MR and therefore a targeted GWAS had to be conducted. The relatively small sample size of ALSPAC was a severe limitation for this method, however. As genetic variants for complex traits (such as smoking behaviour or ID) tend to have small individual effects, sample sizes for GWAS need to be into the hundreds of thousands with several thousand available cases in order to be adequately powered. The lack of power lead to extremely wide confidence intervals that prevented clear

interpretation of whether consistent effect estimates across primary and sensitivity analyses truly reflected a causal effect or were due to chance.

Other, one-sample methods of MR could have been employed using a polygenic risk score for smoking behaviours in the mother. These methods would also have suffered from power issues and would likely have been more susceptible to weak instrument bias than using the GSCAN GWAS which contained data from 1 million individuals. Methodology for one-sample MR has not been developed to the extent that two-sample MR has, and fewer sensitivity analyses are available to explore whether the MR assumptions hold. A further issue of the approach taken in this study is that the ALSPAC study is contained within the consortia of cohorts comprising GSCAN [168]. The exposure and outcome genetic summary statistics were therefore not produced from two entirely independent samples and therefore may suffer bias such as winner's curse whereby the strongest association in a sample of associations is usually overestimated [173, 315]. ALSPAC made up only a small fraction of the GSCAN sample, which may mean that any such biases would be small in size.

A key strength of the study was the ability to use auxiliary information from the widespread data collected by ALSPAC in order to supplement missing information in MI analyses. Data were most often missing in the outcome variable and socioeconomic variables. Here it was assumed that those with lower IQ and greater socioeconomic challenge would be more likely to have missing data. This assumption was supported by the missing data assessment, though it should be noted that this assessment itself may be biased by the absence of values not observed and may also reflect confounded associations, not true causal effects of variables on missing data. The MAR assumption was made more plausible for the outcome variable by using auxiliary variables with limited missing data that were initially intended for use in the multi-sourced indicator of ID. The MAR assumption was also made more plausible for missing socioeconomic variables through the use of home ownership and marital status as auxiliaries. Provided that the MAR assumption has been met, the bias from missing data will be reduced even in the presence of substantial quantities of missing data [157]. This assumption is not verifiable however, and it should also be noted that when data is not MAR based on the analysis model variables and all auxiliaries, then increasing proportions of missing data do lead to increasing quantities of bias (see Chapter 4, Figure 4-2).

The chosen method of implementing MI may have issues also. The choice to impute the outcome as a continuous variable then dichotomise post imputation was made based on a single simulation study that suggested that bias is introduced when the outcome is imputed as a binary variable directly [311]. This is in opposition to the position of the majority of authors working on the implementation of MI who state that the imputation model should be compatible with the analysis model. It is possible that imputation of the outcome is a special case though I can find

no argument as to why this would be so. The decision to use this approach was also made out of necessity, as attempts to impute the outcome as a binary variable lead to issues of model convergence, a problem that has also been reported in simulations focusing on the dichotomisation of continuous covariates [312]. Other modelling strategies were investigated, including the use of substantive model compatible fully conditional specification (SCMFCS) which allows imputation only from models that are compatible with the analysis model (also referred to as the substantive model by the authors) [316]. This method has been shown not to exhibit the bias of the post imputation derivation method while also not experiencing the numerical issues of the binary variable imputation method. SMCFCS does not allow derivations to be applied to the outcome variable, or for multiple forms of the analysis model to be stated (i.e. with and without dichotomisation). This approach was therefore not suitable for the present analysis and the post imputation dichotomisation approach was taken for the outcome.

Although I have aimed to demonstrate a complementary approach to answering the research question using data from a longitudinal cohort, the results and conclusions of this chapter as a whole need to be viewed with caution. One of the largest limitations of this study was that the outcome, ID, was a dichotomised variable derived from IQ scores. Statisticians have long argued against the use of such variables due to the loss of information from the original to the derived variable. From a more clinical perspective, diagnosis of ID requires more information than simply using a cut off based on an IQ score and relies heavily on the functional ability and independence of the individual (see Chapter 1, Section 1.1.1).

The IDI project diagnosis variable could have been used as an alternative outcome definition for all analyses. On the surface, this variable had less missing data (assumed none) and so would not be influenced by selection bias. However, bias from missing data in this variable may be replaced with differential information bias. Diagnosis should be thought of as the measurement of the underlying concept of ID that is observed with error (referred to as Y^* in Chapter 2). The potential exists for diagnosis to be influenced by confounding factors if those from more affluent backgrounds are more likely to have parents who push primary care clinicians for a referral for diagnosis or they attend schools which have greater resources to initiate investigations of ID in a student.

A better method of ID definition would be to use multiple sources of information, each with their own susceptibility for bias, to create a combined variable that assesses intelligence and functional ability, as was attempted at the start of this chapter. Delays in obtaining data and the incompleteness of the obtained data (due to not having section 251 approval) prevented this approach being used. Further work, post PhD, will be undertaken to explore the development of such a variable once data become available. Further exploration of the ALPSAC resource will

be required in this work to improve the assessment of functional ability that is currently lacking in the sources of information for ID obtained so far.

7.5 – Chapter conclusions and summary

The results of analyses in this chapter suggest that the association between maternal smoking during pregnancy and offspring risk of ID can be accounted for by confounding structures and selection bias from missing data. Confidence in any conclusion is limited, however, by the quality of the outcome definition used. Further work is needed to develop the study presented here once data from the Section 251 sample becomes available, which should improve the validity of the outcome measure used.

An important consideration to take forward from this chapter is that causal inference methods are heavily influenced by their own sources of bias that can lead to dramatically different conclusions (i.e. the difference between the complete case and MI analyses of the negative control design). Care needs to be taken to address these sources of bias and not simply accept the results of causal inference technique on the statement that the method was designed to address the biases of observational designs.

Chapter 8 Discussion

I began this thesis by defining intellectual disability (ID), stating the issues surrounding its measurement, establishing the presence of an association between maternal smoking during pregnancy and offspring risk of ID and highlighting that the current evidence suggests that this association may be driven by confounding factors (Chapter 1). In Chapter 2 I established how a causal estimate can be obtained from an observational analyses, the methods to be used in the theses and the biases that can arise. In Chapters 3 and 4 I used simulation studies to further investigate bias in epidemiological analysis methods, first assessing how assortative mating may bias the negative control design and then showing that multiple imputation (MI) analysis is valid even at high proportions of missing data, provided that data are missing at random (MAR). In the empirical assessment of the causal nature of the association (see Table 8-1 for a summary of findings), Chapter 5 and Chapter 6 employed the use of sibling designs to account for shared genetic and environmental confounding in families in Danish and Swedish registry data while in Chapter 7 I used the negative control design and Mendelian randomisation to attempt to account for unmeasured confounding in the Avon Longitudinal Study of Parents and Children (ALSPAC).

In this chapter I discuss and triangulate across the evidence produced for the overarching question of whether maternal smoking during pregnancy causally influences offspring risk of intellectual disability. I explore the strengths and limitations of the approaches used and attempt to highlight further work needed in the area.

Table 8-1: Summary of empirical chapters

Chapter	Dataset used	Methods used	Summary of findings
5	Danish registers	-Traditional epidemiological models (logistic regression) - Sibling comparisons - Positive control analysis (ID vs. birthweight)	- Increased odds of ID among mothers who smoked in traditional models (adjusted OR [95% CI] = 1.35 [1.28-1.42]) - Sibling analysis showed null within-family effect of smoking on offspring ID (OR [95% CI] = 0.93 [0.79-1.09]) - Within-family effect showed increased odds of low birthweight among offspring of children who smoked (OR [95% CI] = 1.06 [1.00-1.13])
6	Swedish registers	-Traditional epidemiological models (logistic regression) - Sibling comparisons - Cross context comparison (smoking vs. snus) - Positive control analysis (ID vs. born SGA)	- Increased odds of ID among mothers who smoked in traditional models (adjusted OR [95% CI] = 1.24 [1.16-1.33]) - Sibling analysis showed null within-family effect of smoking on offspring ID (adjusted OR [95% CI] = 0.92 [0.75-1.14]) - Within-family effect showed increased odds of offspring born SGA among children who's mother smoked in pregnancy (adjusted OR [95% CI] = 2.43 [1.27-2.62]) - Null within-family effect of snus use in pregnancy on offspring ID (adjusted OR [95% CI] = 0.88 [0.60-1.31]) and SGA (adjusted OR [95% CI] = 1.07 [0.84-1.36])
7	ALSPAC	-Traditional epidemiological models (logistic regression) - Negative control (maternal vs. partner smoking) - Mendelian randomisation - Multiple imputation	- No association between smoking in pregnancy and offspring ID after accounting for missing data and adjusting for confounders (OR [95% CI] = 1.10 [0.79-1.52]) - Equivalent maternal ID (OR [95% CI] =0.99 [0.71-1.39]) and partner (OR [95% CI] =1.03 [0.76-1.39]) associations with offspring ID - No association found using MR analysis of SNPs for smoking initiation or cigarettes smoked per day

8.1 – Triangulation of evidence across empirical chapters

Using data from the Danish registers I found increased odds of ID among mothers who smoked during pregnancy in unadjusted conventional models and in conventional models that adjusted for potential confounders. In sibling comparison models no evidence for a within-family effect was found, providing evidence that, holding fixed the shared environmental and genetic factors between siblings, individual level exposure to smoking during pregnancy did not influence the odds of offspring ID. These models provided evidence of a between-family effect, showing increased odds of ID in the families in which mother's tended to smoke during a greater proportion of their pregnancies, holding fixed the individual exposure to smoking during pregnancy. This pattern of results suggested that the evidence from observational/conventional models can be attributed to the determinants of smoking that differ between families as opposed to a causal effect of smoking during pregnancy.

Results of analyses on data from Swedish registers supported the findings of analyses on Danish data, with comparable effect estimates across all models. This is despite differing prevalence of maternal smoking in pregnancy between the two cohorts (18.6% in Denmark vs. 8.8% in Sweden), which may also reflect different confounding structures. The Swedish study further showed a similar pattern of results when snus as opposed to smoking was the method of nicotine exposure, suggesting that the pattern of results is not sensitive to the method of nicotine administration.

Investigations using both Danish and Swedish data were supported by a positive control analysis, replacing the outcome of ID with an outcome of fetal growth restriction. This was used to demonstrate the validity of the within-between modelling approach for an outcome in which there already exists good evidence for a causal role of smoking during pregnancy. In each positive control analysis a within-family and between-family effect of smoking during pregnancy was found supporting a causal role of smoking during pregnancy on offspring fetal growth restriction, but also highlighting that some of the observational association can be accounted for by family-level and not individual-level determinants of smoking during pregnancy.

Differences in the effect sizes of the positive control analyses were found between the Danish and Swedish data, however. The two studies used different definitions of the outcome which may account for this difference. The Danish study defined fetal growth restriction as birthweight less than 2500g which was approximately equivalent to 2.5 standard deviations (SD) below the unconditional population mean of the sample. The Swedish study instead defined fetal growth restriction as 2 SD below the mean birthweight for a given gender and gestational age. The

difference in estimates between the two are likely the result of different cut offs and the Swedish outcome being conditioned upon gender and gestational age.

Findings in ALSPAC also showed increased odds of offspring ID among those exposed to maternal smoking in pregnancy in unadjusted regression models. These associations did not survive adjustment for maternal and socioeconomic characteristics. The largest difference between the ALSPAC cohort and the Danish and Swedish cohorts were the size of the cohorts and the definition of ID. The lower power of the smaller ALSPAC cohort may have led to an inability to detect effects in adjusted models. Due to the inability to obtain all linked data for the ALSPAC cohort, I decided to define ID based on IQ as opposed to on medical diagnoses, as was the case for the Danish and Swedish studies. The underlying construct of the outcome may therefore differ between the studies, with ALSPAC potentially containing greater measurement error in this variable, which could have led to the lack of evidence for increased odds of ID among children of maternal smokers in adjusted models in ALSPAC. Alternatively, the measures adjusted for in ALSPAC models may have better captured confounding than the measures used in the Danish and Swedish registers. ALSPAC used detailed questionnaires with data collected for the purpose of research. In comparison, the measures from registry data are routinely collected data that were repurposed for research and so may be less able to capture information on confounding.

The results of negative control analyses in ALSPAC supported the results of the sibling comparisons in previous chapters once bias from missing data had been accounted for. Similarly sized unadjusted and adjusted effect estimates were found for maternal and partner coefficients which provided evidence against a causal role of maternal smoking during pregnancy in the development of offspring ID. This is based on the assumption of equivalent bias for maternal and partner effect estimates and that partner smoking should have no, or at least a weaker effect than maternal smoking given that maternal smoking should deliver a greater dosage of nicotine and combustible components to the fetus than partner smoking. The first of these assumptions was potentially untenable, as evidenced by the descriptive statistics provided in the chapter. Attempts were made to account for this by adjusting for both parents confounder variables to control for non-shared confounding.

Increasing odds of ID with increased dose of exposure would be expected if smoking during pregnancy were to have a causal influence. Dosage analyses in both the Danish and ALSPAC studies suggested that increased dosage of maternal smoking during pregnancy did not causally influence risk of offspring ID. The Danish study found no evidence of a non-null within-family effect of increasing number of cigarettes smoked per day. In the ALSPAC study a null adjusted observational effect estimate was found alongside similar effect sizes for maternal and partner number of cigarettes smoked per day.

The timing of exposure might also be expected to influence the risk of ID if an association were causal. Timing analyses were conducted in the Danish and Swedish studies. The Danish work investigated the effect of stopping or continuing smoking after the first trimester as compared to not smoking at all. Here a greater effect was found for smoking late into pregnancy than quitting in the first trimester in both unadjusted and adjusted models. These analyses did not account for shared unmeasured confounders however. The Swedish study compared smoking before pregnancy only, quitting during pregnancy and use late into pregnancy as compared to no use at all. In this study an increasing effect size was found for increasing length of use in pregnancy, however, this attenuated to a null effect in models that accounted for shared unmeasured confounders (the conditional logistic regression models). Based on the evidence of the better quality model from the Swedish study, this suggests that the timing of exposure to smoking during pregnancy does not influence the offspring's risk of ID.

MR analyses in Chapter 7 also did not support a causal role of maternal smoking in pregnancy on offspring risk of ID. Here the quality of evidence was low. The ALSPAC sample size was too low to detect SNP-outcome associations with adequate precision meaning that true causal effects may be hidden by large estimate standard errors. Further, the GWAS used for the SNP-exposure associations was not specific to pregnancy meaning that the lack of evidence for a non-null association could reflect no influence of increased smoking behaviours across the mother's life course, but does not rule out an influence of smoking specific to the pregnancy period.

Overall, the combined evidence across all analyses presented in this thesis do not support a causal role of smoking during pregnancy on the offspring's risk of ID. Once confounding has been adequately accounted for by analysis design the association attenuated to reveal a null effect.

8.2 – Comparison with the current literature

Results reported in this thesis do not support the results of the recent meta-analysis that suggested a small increase in the risk of ID among children of mothers who smoked during pregnancy [97]. Instead, the results support the findings of Braun et al. [98] and Lundberg et al. [99], that associations using conventional epidemiological analyses do not adequately account for the influence of unmeasured confounding and therefore reflect biased estimates. I have further provided evidence against an influence of timing of exposure as may be suggested by the findings of Hirvonen et al. [104].

A number of studies have used sibling designs to investigate the consequences of smoking during pregnancy [65, 317, 318]. These studies provided evidence for causal effects of smoking during pregnancy on birthweight, preterm birth and being born small for gestational age, supporting the results of positive control analyses in Chapters 5 and 6. Of importance, these studies did not find evidence for a causal influence of smoking upon cognitive measures such as academic achievement and general cognitive ability or neurodevelopmental outcomes such as conduct problems. Authors of studies using other causal inference methods such as MR and the negative control design have come to similar conclusions, that the available evidence suggests that smoking during pregnancy is associated with pregnancy outcomes but possibly not more complex later life outcomes such as cognition [319], depression [319, 320] and BMI [101, 319].

With regard to ID specifically, it has been suggested by Reichenberg and colleagues that that the factors influencing the mild end of ID spectrum are the same as those that influence the normal distribution of IQ, while severe ID may influenced by a distinct set of risk factors [30]. It is possible that if Reichenberg's theory holds then severe ID could be caused by rare de novo mutations, which smoking during pregnancy has been suggested as risk factor for [321, 322]. This theory was examined briefly in Chapter 5 by looking at the association of maternal smoking during pregnancy with different severities of ID. Smoking was associated with milder but not more severe ID, though analyses were unadjusted and a large proportion of the diagnoses had an unspecified severity. In all cohorts used in this thesis, individuals with genetic disorders related to ID were excluded from analyses, preventing assessment of associations between maternal smoking during pregnancy and severe offspring ID resulting from de novo mutations. This therefore remains an area of research that could be explored further.

8.3 – Strengths and limitations

Specific strengths and weaknesses of each study design have been described and addressed in the chapters in which analyses took place. Here I present the strengths and weaknesses of the thesis as a whole.

A key strength of this thesis was the use of different analysis methods to investigate the research question, each of which suffered from biases in different ways, as highlighted in Chapter 2. Further I used three different cohorts originating from three different countries each with their own confounding structures. The data from these cohorts was obtained in different ways and for different purposes. The registry data of Denmark and Sweden was routinely collected data while

the ALSPAC data was collected for the purpose of research. This would lead to different selection pressures for continuing contribution of data for participants contained within each cohort. For example, collection of outcome data would occur provided an individual was in contact with health services in the Danish and Swedish cohorts, whereas ascertainment of the outcome via response to questionnaires and clinics may be reduced in ALSPAC participants among some subgroups such as those who have ID. In spite of all of these differences, consistent findings were obtained across the different analysis methods in each of the cohorts. This raises the confidence in my conclusion that smoking does not causally influence offspring risk of intellectual disability.

Another key strength of this thesis is that biases were not just identified but were investigated in detail using simulation studies. Work from this thesis has contributed to the scientific literature regarding the nature of bias surrounding assortative mating and its influence on the results of the negative control design of a prenatal exposure (Chapter 3) and on the influence of the proportion of missing data on the bias of complete cases analyses and improperly specified multiple imputation models (Chapter 4). These studies provided the opportunity to explore the biases first hand and use the results to guide analysis strategies in later empirical work (implemented in Chapter 7). Questions were left unanswered in these simulation studies, for example how might assortative mating influence bias from non-shared confounding in the negative control design. Despite not having evidence on the potential consequences of this, the work made me aware of the issue and was used to inform methods used to account for it based on the knowledge already developed. Regarding this specific example, I chose to control for both maternal and paternal confounders in the negative control design.

The thesis has been a learning process, and as such not all analyses have been conducted perfectly due to developing knowledge combined with time pressures. For example in Chapter 2 I highlighted that effect estimates of within-between models of sibling studies are biased when non-shared confounders are not accounted for [180]. While it is not possible to account for unmeasured non-shared confounding, measured non-shared confounders could have been adjusted for in models if only two siblings had been included per family. A trade-off existed between using the largest sample size possible and obtaining an estimate as close to the true value of the within-family effect (i.e. an unbiased estimate). In Danish and Swedish analyses I opted for the largest sample size possible although lower sample sizes may have been sufficient. Using data from overseas cohorts came with time limitations, particularly with the implementation of the General Data Protection Regulation part way through writing (implemented May 2018) which limited access to personal data across the European Union and European Economic Area [323]. My understanding of the importance of non-shared confounding did not develop until after I had completed analyses and was no longer able to

access the Danish and Swedish datasets. Given the opportunity, I would consider revising the analyses to use only two siblings, thereby prioritising reduction of bias over sample size. Selecting which measured non-shared confounders to adjust for would require further thought as many of the potential candidates, such as parental income, education and psychiatric history are likely to remain constant due to the often relatively short time between pregnancies. It is also important to note that further considerations would have to be made regarding which siblings to select so as to not introduce new bias from patterns of selection. It is likely that as my understanding of epidemiology and bias continues to develop, I will look back at the work contained in this thesis and conclude that many of the decisions made could be improved.

Timing was also an issue in the use of linked data in the UK based ALSPAC cohort. The process to obtain the linked health records was long and data were not accessible until very late into the project. Even when linked data were accessed, individuals who were more likely to have ID (those unable to provide explicit consent) did not have health records data available. This meant that the selection biases that the approach of creating a multiply sourced variable aimed to address, still afflicted the study. When creating the multi-sourced variable using the available data it seemed that the measure would not be valid on the basis that the selected sources were identifying ID in individuals with IQ higher than the cut off of 70 used in diagnostic measures, though here it may have been better to focus on indicators of functional ability that underly the definition of ID. There was insufficient time remaining after health records access to fully explore developing this variable. A pragmatic decision had to be taken for the form of the outcome variable so that the overarching research question could be investigated. This led to the multiple imputation approach using some of the sources of information for ID as auxiliary variables to inform about reasons for missing data and reduce bias. With greater time the outcome variable for the ALSPAC study could be developed further to improve the validity of results in the chapter. My intention is to carry this forward in the next stages of my research career, though when the necessary data will be available is unclear.

Survival of the fetus through pregnancy and to an age where ID can be diagnosed is a potential source of bias that was not addressed by any of the methods implemented in this thesis. This issue is referred to as “immortal time bias” by some [324] and as “survival bias” by others [325, 326]. In analyses, sometimes explicitly and sometimes by default, I have selected those that have survived to an age where they can be assessed for ID, thereby conditioning on survival. As smoking has been suggested to be a risk factor for spontaneous abortion [327, 328] this leads to the potential for collider bias if the outcome is also associated with survival via an indirect pathway, for example by a common cause of survival and ID such as genetic variants or socioeconomics [329]. It has been suggested that controlling for such common causes of fetal/child survival and the outcome of interest can reduce such survival bias [325]. Some of

these may have been accounted for by controlling for measured confounders or by using sibling comparisons. However, not all of these common causes will have been measured and exclusion of siblings who did not survive to term from the calculation of family-averaged exposure may have limited control for shared confounders. The consequences of fetal/child survival is beginning to be discussed in the literature but further work needs to be undertaken to consider the consequences to estimates obtained from causal inference methods such as the sibling design, negative control design and Mendelian randomisation.

8.4 – Future research

The overarching evidence produced from the studies presented in this thesis, combined with that of the wider literature suggests that associations between smoking during pregnancy and offspring intellectual disability do not reflect causal effects. They instead reflect residual confounding not accounted for by analysis methods. Determining the factors that drive this residual confounding would inform further research as to what additional variables need to be adjusted for in analyses and may also further our understanding of the aetiology of ID.

Determining the extent to which this residual confounding is environmental or genetic in nature, or the result of important interactions between the two, may be a useful starting point. Sibling studies are not able to provide evidence of this nature, however quasi-experimental methods that exploit family structure to vary the degree to which individuals are genetically similar to one another could be used [330]. For example the association of smoking in pregnancy and ID can be compared between full (50% genetics shared on average) and half siblings (25% shared) and between cousins whose mothers are full siblings (12.5% genetics shared between the cousins; maternal full cousins) or half siblings (6.25% shared; maternal half cousins). If the association decreases in strength as the genetic similarity between family members increases then this implicates a role of genetics in driving the confounding. Such a design could feasibly implemented in the Danish and Swedish registries. D'Onofrio and colleagues noted that such a finding would not prove a role of genetics, due to the increasingly different environments between family relations with increasingly different genetics, though would provide evidence in favour of this conclusion [65]. In contrast, evidence for environmental causes of the confounding would be provided by consistent association strength between analyses of decreasingly genetically related individuals. Such analyses would be a natural follow-up to the work presented in Chapters 5 and 6.

It should be noted that different strategies were employed for defining families in the Danish (Chapter 5) and Swedish (Chapter 6) sibling comparison studies. For the Danish cohort only full siblings were included in families, while for the Swedish cohort both full siblings and half siblings who shared the same mother were included in a family unit. The within-family effect will reflect slightly greater control for shared genetics in the Danish than Swedish study as a result of the different definitions of family structure. Similar estimates for the within-family effect were found between the two chapters which may suggest that there is little evidence of genetic confounding. The definitions of family structure did overlap, however, meaning that any differences between the two would be expected to be very small, and larger genetic differences such as restricting to half siblings, maternal full cousins and maternal half cousins in separate analyses would be required to investigate the genetic nature of residual confounding thoroughly.

Mendelian randomisation (MR) techniques are under utilised in investigations of prenatal exposures and offspring cognitive and neurodevelopmental outcomes, largely due to the absence of large scale intergenerational genome wide association studies (GWAS) for these outcomes. This absence meant that in Chapter 7 SNP-outcome associations had to be modelled using ALSPAC data. The resulting summary statistics had low power which, in turn, severely reduced the power of MR analyses so that null results could not meaningfully be distinguished between a lack of evidence for a causal effect or an inability to detect an effect. Adequately powered MR analyses that could provide useful evidence of the causal nature of prenatal exposure and offspring cognitive outcomes would likely require creation of a consortia across cohorts that had comparable measures of the outcome and also had available data on maternal genetics. Such consortia have been created for other outcomes such as birthweight [331] and could therefore feasibly be extended to cognitive outcomes. Issues are likely to arise where the outcome has been measured using different methods and at different time points across component cohorts of the consortium.

An important step for further research is to create more consistent definitions of ID across studies. The regional differences in terminology (see Section 1.1.1 for a description) and different criteria used between studies may lead to heterogeneity of the concept of ID that could slow progress of research into its aetiology. Many studies, including the definition used in Chapter 7 of this thesis, have not adequately captured information on functional ability. This information is important in order to accurately reflect that not all those with an IQ less than 70 are severely negatively impacted. Investigation of the aetiology of functional impairments may in fact be more important than investigations into IQ given that these impairments are more likely to negatively impact the individuals life experience and provide societal cost as a result of support needs.

8.5 – Conclusion

In conclusion, maternal smoking in pregnancy is associated with increased odds of ID but, using triangulation across multiple analytical methods, this association does not appear to be causal in nature. Effect estimates of the association produced using traditional methods were likely biased by residual confounding. The exact nature of the confounding is unknown but further investigation may reveal important determinants of ID and guide decisions on what information should be recorded when designing future cohorts.

While the evidence produced does not suggest that smoking during pregnancy causes ID, this finding does not suggest that smoking in pregnancy is safe. Smoking in pregnancy should still be avoided as evidence from positive control analysis found associations with fetal growth restriction that did appear to be causal. Fetal growth restriction reflects global changes as a consequence of exposure that may in turn point to more specific negative effects for the child. Continued effort needs to be made to reduce the prevalence of smoking during pregnancy which remains high in some countries.

References

1. European Perinatal Health Report. *Health and care of pregnant women and babies in Europe in 2010*. 2013 09/08/2018]; Available from: http://www.europeristat.com/reports/european-perinatal-health-report-2010.html.
2. Abel E.L., *Smoking during pregnancy: a review of effects on growth and development of offspring.* Hum Biol, 1980. **52**(4): p. 593-625.
3. Evans W.N. and Ringel J.S., *Can higher cigarette taxes improve birth outcomes?* Journal of Public Economics, 1999. **72**(1): p. 135-154.
4. Hamilton B.H., *Estimating treatment effects in randomized clinical trials with non-compliance: the impact of maternal smoking on birthweight.* Health Econ, 2001. **10**(5): p. 399-410.
5. Kramer M.S., *Determinants of low birth weight: methodological assessment and meta-analysis.* Bull World Health Organ, 1987. **65**(5): p. 663-737.
6. Permutt T. and Hebel J.R.J.B., *Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight.* 1989: p. 619-622.
7. Sexton M. and Hebel J.R.J.J., *A clinical trial of change in maternal smoking and its effect on birth weight.* 1984. **251**(7): p. 911-915.
8. Smith G.D., *Assessing intrauterine influences on offspring health outcomes: can epidemiological studies yield robust findings?* Basic Clin Pharmacol Toxicol, 2008. **102**(2): p. 245-56.
9. Tyrrell J., Huikari V., Christie J.T., Cavadino A., Bakker R., Brion M.J., et al., *Genetic variation in the 15q25 nicotinic acetylcholine receptor gene cluster (CHRNA5-CHRNA3-CHRNB4) interacts with maternal self-reported smoking status during pregnancy to influence birth weight.* Human Molecular Genetics, 2012. **21**(24): p. 5344-58.
10. Castles A., Adams E.K., Melvin C.L., Kelsch C., and Boulton M.L., *Effects of smoking during pregnancy - Five meta-analyses.* American Journal of Preventive Medicine, 1999. **16**(3): p. 208-215.
11. Anderson T.M., Ferres J.M.L., Ren S.Y., Moon R.Y., Goldstein R.D., Ramirez J.M., et al., *Maternal Smoking Before and During Pregnancy and the Risk of Sudden Unexpected Infant Death.* Pediatrics, 2019. **143**(4).
12. World Health Organization. *ICD Classifications*. 2018 30/05/2018]; Available from: http://www.who.int/classifications/icd/en/.

13. The Baily Thomas Charitable Fund. *What the Baily Thomas Charitable Fund means by "learning disabilities"*. 2020 [cited 2020 10 Aug]; Available from: http://bailythomas.org.uk/what-baily-thomas-charitable-fund-means-learning-disabilities.
14. American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*. 4 ed. 1994, Arlington, VA: Author.
15. American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*. 5 ed. 2013, Arlington, VA: Author.
16. Cooper S.-A., Henderson A., Jacobs M., and Smiley E. *What are Learning Disabilities? How Common Are Learning Disabilities?* 2016 [cited 2020 31 July]; Available from: https://www.sldo.ac.uk/media/1610/what-are-learning-disabilities-how-common-are-learning-disabilities.pdf.
17. Welton J., *Implementing the 1981 education ACT.* Higher education, 1983. **12**(5): p. 597-607.
18. O'Brien G., *Defining learning disability: what place does intelligence testing have now?* Developmental Medicine and Child Neurology, 2001. **43**(8): p. 570.
19. Barnett W.S., *Definition and classification of mental retardation: A reply to Zigler, Balla, and Hodapp.* American journal of mental deficiency, 1986. **91**(2): p. 111-16.
20. Maulik P.K., Mascarenhas M.N., Mathers C.D., Dua T., and Saxena S., *Prevalence of intellectual disability: a meta-analysis of population-based studies.* Res Dev Disabil, 2011. **32**(2): p. 419-36.
21. Idring S., Rai D., Dal H., Dalman C., Sturm H., Zander E., et al., *Autism spectrum disorders in the Stockholm Youth Cohort: design, prevalence and validity.* PLoS One, 2012. **7**(7): p. e41280.
22. Golding J., Ellis G., Gregory S., Birmingham K., Iles-Caven Y., Rai D., et al., *Grand-maternal smoking in pregnancy and grandchild's autistic traits and diagnosed autism.* Scientific reports, 2017. **7**: p. 46179.
23. Emerson E., Einfeld S., and Stancliffe R.J., *The mental health of young children with intellectual disabilities or borderline intellectual functioning.* Social Psychiatry and Psychiatric Epidemiology, 2010. **45**(5): p. 579-87.
24. Hassiotis A., Strydom A., Hall I., Ali A., Lawrence-Smith G., Meltzer H., et al., *Psychiatric morbidity and social functioning among adults with borderline intelligence living in private households.* Journal of Intellect Disability Research, 2008. **52**(Pt 2): p. 95-106.
25. Hosking F.J., Carey I.M., Shah S.M., Harris T., DeWilde S., Beighton C., et al., *Mortality Among Adults With Intellectual Disability in England: Comparisons With the General Population.* American Journal of Public Health, 2016. **106**(8): p. 1483-1490.

26. Lennox N.G. and Kerr M.P., *Primary health care and people with an intellectual disability: the evidence base.* Journal of Intellectual Disability Research, 1997. **41 (Pt 5)**: p. 365-72.
27. Michael J., *Healthcare for all: report of the independent inquiry into access to healthcare for people with learning disabilities*. 2008, Department of Health: www.dh.gov.uk.
28. Whitfield M., Langan J., and Russell O., *Assessing general practitioners' care of adult patients with learning disability: case-control study.* Quality in Health Care, 1996. **5**(1): p. 31-5.
29. Wilson D.N. and Haire A., *Health care screening for people with mental handicap living in the community.* British Medical Journal, 1990. **301**(6765): p. 1379-81.
30. Reichenberg A., Cederlof M., McMillan A., Trzaskowski M., Kapra O., Fruchter E., et al., *Discontinuity in the genetic and environmental causes of the intellectual disability spectrum.* Proc Natl Acad Sci U S A, 2016. **113**(4): p. 1098-103.
31. Leach R.M. and Treacher D.F., *ABC of oxygen - Oxygen transport - 2. Tissue hypoxia.* British Medical Journal, 1998. **317**(7169): p. 1370-1373.
32. Andersen M.R., Uldbjerg N., Stender S., Sandager P., and Aalkjraer C., *Maternal smoking and impaired endothelium-dependent nitric oxide-mediated relaxation of uterine small arteries in vitro.* American Journal of Obstetrics and Gynecology, 2011. **204**(2).
33. Philipp K., Pateisky N., and Endler M., *Effects of Smoking on Uteroplacental Blood-Flow*. Gynecologic and Obstetric Investigation, 1984. **17**(4): p. 179-182.
34. Bureau M.A., Shapcott D., Berthiaume Y., Monette J., Blouin D., Blanchard P., et al., *Maternal Cigarette-Smoking and Fetal Oxygen-Transport - a Study of P50, 2,3-Diphosphoglycerate, Total Hemoglobin, Hematocrit, and Type-F Hemoglobin in Fetal Blood.* Pediatrics, 1983. **72**(1): p. 22-26.
35. Luck W., Nau H., Hansen R., and Steldinger R., *Extent of Nicotine and Cotinine Transfer to the Human-Fetus, Placenta and Amniotic-Fluid of Smoking Mothers.* Developmental Pharmacology and Therapeutics, 1985. **8**(6): p. 384-395.
36. Jauniaux E., Gulbis B., Acharya G., Thiry P., and Rodeck C., *Maternal tobacco exposure and cotinine levels in fetal fluids in the first half of pregnancy.* Obstetrics and Gynecology, 1999. **93**(1): p. 25-29.
37. Dwyer J.B., Broide R.S., and Leslie F.M., *Nicotine and brain development.* Birth Defects Research Part C: Embryo Today: Reviews, 2008. **84**(1): p. 30-44.
38. Slotkin T.A., *Fetal nicotine or cocaine exposure: Which one is worse?* Journal of Pharmacology and Experimental Therapeutics, 1998. **285**(3): p. 931-945.

39. Chan Y.L., Oliver B.G., and Chen H., *What lessons have we learnt about the impact of maternal cigarette smoking from animal models?* Clinical and Experimental Pharmacology and Physiology, 2019.
40. Krishnaswamy A. and Cooper E., *Reactive oxygen species inactivate neuronal nicotinic acetylcholine receptors through a highly conserved cysteine near the intracellular mouth of the channel: implications for diseases that involve oxidative stress.* Journal of Physiology-London, 2012. **590**(1): p. 39-47.
41. Dwyer J.B., McQuown S.C., and Leslie F.M., *The dynamic effects of nicotine on the developing brain.* Pharmacology & Therapeutics, 2009. **122**(2): p. 125-139.
42. Ekblad M., Korkeila J., and Lehtonen L., *Smoking during pregnancy affects foetal brain development.* Acta paediatrica, 2015. **104**(1): p. 12-8.
43. Bublitz M.H. and Stroud L.R., *Maternal Smoking During Pregnancy and Offspring Brain Structure and Function: Review and Agenda for Future Research.* Nicotine & Tobacco Research, 2012. **14**(4): p. 388-397.
44. Ekblad M., Korkeila J., Parkkola R., Lapinleimu H., Haataja L., Lehtonen L., et al., *Maternal Smoking during Pregnancy and Regional Brain Volumes in Preterm Infants.* Journal of Pediatrics, 2010. **156**(2): p. 185-U40.
45. Roza S.J., Verburg B.O., Jaddoe V.W., Hofman A., Mackenbach J.P., Steegers E.A., et al., *Effects of maternal smoking in pregnancy on prenatal brain development. The Generation R Study.* European Journal of Neuroscience, 2007. **25**(3): p. 611-7.
46. Toro R., Leonard G., Lerner J.V., Lerner R.M., Perron M., Pike G.B., et al., *Prenatal exposure to maternal cigarette smoking and the adolescent cerebral cortex.* Neuropsychopharmacology, 2008. **33**(5): p. 1019-1027.
47. Jacobsen L.K., Picciotto M.R., Heath C.J., Frost S.J., Tsou K.A., Dwan R.A., et al., *Prenatal and adolescent exposure to tobacco smoke modulates the development of white matter microstructure.* Journal of Neuroscience, 2007. **27**(49): p. 13491-13498.
48. Paus T., Nawazkhan I., Leonard G., Perron M., Pike G.B., Pitiot A., et al., *Corpus callosum in adolescent offspring exposed prenatally to maternal cigarette smoking.* Neuroimage, 2008. **40**(2): p. 435-441.
49. Biffen S.C., Warton C.M.R., Lindinger N.M., Randall S.R., Lewis C.E., Molteno C.D., et al., *Reductions in Corpus Callosum Volume Partially Mediate Effects of Prenatal Alcohol Exposure on IQ.* Frontiers in Neuroanatomy, 2017. **11**: p. 132.
50. Rivkin M.J., Davis P.E., Lemaster J.L., Cabral H.J., Warfield S.K., Mulkern R.V., et al., *Volumetric MRI study of brain in children with intrauterine exposure to cocaine, alcohol, tobacco, and marijuana.* Pediatrics, 2008. **121**(4): p. 741-750.

51. Peck J.D., Neas B., Robledo C., Saffer E., Beebe L., and Wild R.A., *Intrauterine tobacco exposure may alter auditory brainstem responses in newborns.* Acta Obstetricia Et Gynecologica Scandinavica, 2010. **89**(4): p. 592-596.
52. Kable J.A., Coles C.D., Lynch M.E., and Carroll J., *The impact of maternal smoking on fast auditory brainstem responses.* Neurotoxicology and Teratology, 2009. **31**(4): p. 216-224.
53. Bennett D.S., Mohamed F.B., Carmody D.P., Bendersky M., Patel S., Khorrami M., et al., *Response inhibition among early adolescents prenatally exposed to tobacco: An fMRI study.* Neurotoxicology and Teratology, 2009. **31**(5): p. 283-290.
54. Jacobsen L.K., Slotkin T.A., Mencl W.E., Frost S.J., and Pugh K.R., *Gender-specific effects of prenatal and adolescent exposure to tobacco smoke on auditory and visual attention.* Neuropsychopharmacology, 2007. **32**(12): p. 2453-2464.
55. Jacobsen L.K., Slotkin T.A., Westerveld M., Mencl W.E., and Pugh K.R., *Visuospatial memory deficits emerging during nicotine withdrawal in adolescents with prenatal exposure to active maternal smoking.* Neuropsychopharmacology, 2006. **31**(7): p. 1550-1561.
56. Martin-Loeches M., Munoz-Ruata J., Martinez-Lebrusant L., and Gomez-Jarabo G., *Electrophysiology and intelligence: the electrophysiology of intellectual functions in intellectual disability.* J Intellect Disabil Res, 2001. **45**(Pt 1): p. 63-75.
57. Munoz-Ruata J., Caro-Martinez E., Martinez Perez L., and Borja M., *Visual perception and frontal lobe in intellectual disabilities: a study with evoked potentials and neuropsychology.* J Intellect Disabil Res, 2010. **54**(12): p. 1116-29.
58. Pavlowsky A., Chelly J., and Billuart P., *Emerging major synaptic signaling pathways involved in intellectual disability.* Mol Psychiatry, 2012. **17**(7): p. 682-93.
59. Clifford A., Lang L., and Chen R., *Effects of maternal cigarette smoking during pregnancy on cognitive parameters of children and young adults: a literature review.* Neurotoxicology and teratology, 2012. **34**(6): p. 560-70.
60. Polanska K., Jurewicz J., and Hanke W., *Smoking and Alcohol Drinking during Pregnancy as the Risk Factors for Poor Child Neurodevelopment - a Review of Epidemiological Studies.* International Journal of Occupational Medicine and Environmental Health, 2015. **28**(3): p. 419-443.
61. Gilman S.E., Gardener H., and Buka S.L., *Maternal smoking during pregnancy and children's cognitive and physical development: a causal risk factor?* Am J Epidemiol, 2008. **168**(5): p. 522-31.
62. O'Callaghan F.V., Al Mamun A., O'Callaghan M., Alati R., Williams G.M., and Najman J.M., *Is smoking in pregnancy an independent predictor of academic*

difficulties at 14years of age? A birth cohort study. Early Hum Dev, 2010. **86**(2): p. 71-6.

63. Martin R.P., Dombrowski S.C., Mullis C., Wisenbaker J., and Huttunen M.O., *Smoking during pregnancy: association with childhood temperament, behavior, and academic performance.* J Pediatr Psychol, 2006. **31**(5): p. 490-500.
64. Lambe M., Hultman C., Torrang A., Maccabe J., and Cnattingius S., *Maternal smoking during pregnancy and school performance at age 15.* Epidemiology, 2006. **17**(5): p. 524-30.
65. D'Onofrio B.M., Singh A.L., Iliadou A., Lambe M., Hultman C.M., Neiderhiser J.M., et al., *A quasi-experimental study of maternal smoking during pregnancy and offspring academic achievement.* Child Development, 2010. **81**(1): p. 80-100.
66. Cornelius M.D., De Genna N.M., Leech S.L., Willford J.A., Goldschmidt L., and Day N.L., *Effects of prenatal cigarette smoke exposure on neurobehavioral outcomes in 10-year-old children of adolescent mothers.* Neurotoxicol Teratol, 2011. **33**(1): p. 137-44.
67. Roza S.J., Verhulst F.C., Jaddoe V.W., Steegers E.A., Mackenbach J.P., Hofman A., et al., *Maternal smoking during pregnancy and child behaviour problems: the Generation R Study.* Int J Epidemiol, 2009. **38**(3): p. 680-9.
68. Zhu J.L., Olsen J., Liew Z., Li J., Niclasen J., and Obel C., *Parental smoking during pregnancy and ADHD in children: the Danish national birth cohort.* Pediatrics, 2014. **134**(2): p. e382-8.
69. Huang L., Wang Y., Zhang L., Zheng Z., Zhu T., Qu Y., et al., *Maternal Smoking and Attention-Deficit/Hyperactivity Disorder in Offspring: A Meta-analysis.* Pediatrics, 2018. **141**(1).
70. Batty G.D., Der G., and Deary I.J., *Effect of maternal smoking during pregnancy on offspring's cognitive ability: empirical evidence for complete confounding in the US national longitudinal survey of youth.* Pediatrics, 2006. **118**(3): p. 943-50.
71. Fried P.A. and Watkinson B., *Visuoperceptual functioning differs in 9- to 12-year olds prenatally exposed to cigarettes and marihuana.* Neurotoxicol Teratol, 2000. **22**(1): p. 11-20.
72. Fried P.A., Watkinson B., and Gray R., *Differential effects on cognitive functioning in 13- to 16-year-olds prenatally exposed to cigarettes and marihuana.* Neurotoxicol Teratol, 2003. **25**(4): p. 427-36.
73. Huijbregts S.C.J., Seguin J.R., Zelazo P.D., Parent S., Japel C., and Tremblay R.E., *Interrelations Between Maternal Smoking During Pregnancy, Birth Weight and Sociodemographic Factors in the Prediction of Early Cognitive Abilities.* Infant and child development, 2006. **15**(6): p. 593-606.

74. Lawlor D.A., Najman J.M., Batty G.D., O'Callaghan M.J., Williams G.M., and Bor W., *Early life predictors of childhood intelligence: findings from the Mater-University study of pregnancy and its outcomes.* Paediatric and perinatal epidemiology, 2006. **20**(2): p. 148-62.
75. Mortensen E.L., Michaelsen K.F., Sanders S.A., and Reinisch J.M., *A dose-response relationship between maternal smoking during late pregnancy and adult intelligence in male offspring.* Paediatric and perinatal epidemiology, 2005. **19**(1): p. 4-11.
76. Olds D.L., Henderson C.R., Jr., and Tatelbaum R., *Intellectual impairment in children of women who smoke cigarettes during pregnancy.* Pediatrics, 1994. **93**(2): p. 221-7.
77. Cornelius M.D., Ryan C.M., Day N.L., Goldschmidt L., and Willford J.A., *Prenatal tobacco effects on neuropsychological outcomes among preadolescents.* J Dev Behav Pediatr, 2001. **22**(4): p. 217-25.
78. Julvez J., Ribas-Fito N., Torrent M., Forns M., Garcia-Esteban R., and Sunyer J., *Maternal smoking habits and cognitive development of children at age 4 years in a population-based birth cohort.* Int J Epidemiol, 2007. **36**(4): p. 825-32.
79. Kafouri S., Leonard G., Perron M., Richer L., Seguin J.R., Veillette S., et al., *Maternal cigarette smoking during pregnancy and cognitive performance in adolescence.* Int J Epidemiol, 2009. **38**(1): p. 158-72.
80. Huijbregts S.C., Warren A.J., de Sonneville L.M., and Swaab-Barneveld H., *Hot and cool forms of inhibitory control and externalizing behavior in children of mothers who smoked during pregnancy: an exploratory study.* J Abnorm Child Psychol, 2008. **36**(3): p. 323-33.
81. Mezzacappa E., Buckner J.C., and Earls F., *Prenatal cigarette exposure and infant learning stimulation as predictors of cognitive control in childhood.* Dev Sci, 2011. **14**(4): p. 881-91.
82. Noland J.S., Singer L.T., Short E.J., Minnes S., Arendt R.E., Kirchner H.L., et al., *Prenatal drug exposure and selective attention in preschoolers.* Neurotoxicol Teratol, 2005. **27**(3): p. 429-38.
83. Jung Y., Lee A.M., McKee S.A., and Picciotto M.R., *Maternal smoking and autism spectrum disorder: meta-analysis with population smoking metrics as moderators.* Scientific Reports, 2017. **7**(1): p. 4315.
84. Rosen B.N., Lee B.K., Lee N.L., Yang Y., and Burstyn I., *Maternal Smoking and Autism Spectrum Disorder: A Meta-analysis.* Journal of Autism and Developmental Disorders, 2015. **45**(6): p. 1689-98.
85. Tang S., Wang Y., Gong X., and Wang G., *A Meta-Analysis of Maternal Smoking during Pregnancy and Autism Spectrum Disorder Risk in Offspring.* Int J Environ Res Public Health, 2015. **12**(9): p. 10418-31.

86. Wang C., Geng H., Liu W., and Zhang G., *Prenatal, perinatal, and postnatal factors associated with autism: A meta-analysis.* Medicine (Baltimore), 2017. **96**(18): p. e6696.
87. Kalkbrenner A.E., Meier S.M., Madley-Dowd P., Ladd-Acosta C., Fallin M.D., Parner E., et al., *Familial confounding of the association between maternal smoking in pregnancy and autism spectrum disorder in offspring.* Autism Res, 2020. **13**(1): p. 134-144.
88. Kafouri S., Leonard G., Perron M., Richer L., Seguin J.R., Veillette S., et al., *Maternal cigarette smoking during pregnancy and cognitive performance in adolescence.* International journal of epidemiology, 2009. **38**(1): p. 158-72.
89. Ramsay H., Barnett J.H., Murray G.K., Maki P., Hurtig T., Nordstrom T., et al., *Smoking in pregnancy, adolescent mental health and cognitive performance in young adult offspring: results from a matched sample within a Finnish cohort.* BMC psychiatry, 2016. **16**(1): p. 430.
90. Drews C.D., Murphy C.C., Yeargin-Allsopp M., and Decoufle P., *The relationship between idiopathic mental retardation and maternal smoking during pregnancy.* Pediatrics, 1996. **97**(4): p. 547-53.
91. Roeleveld N., Vingerhoets E., Zielhuis G.A., and Gabreels F., *Mental-Retardation Associated with Parental Smoking and Alcohol-Consumption before, during, and after Pregnancy.* Preventive Medicine, 1992. **21**(1): p. 110-119.
92. Yearginallsopp M., Murphy C.C., Oakley G.P., and Sikes R.K., *A Multiple-Source Method for Studying the Prevalence of Developmental-Disabilities in Children - the Metropolitan Atlanta Developmental-Disabilities Study.* Pediatrics, 1992. **89**(4): p. 624-630.
93. Berkson J., *Limitations of the Application of Fourfold Table Analysis to Hospital Data.* Biometrics Bulletin, 1946. **2**(3): p. 47-53.
94. Snoep J.D., Morabia A., Hernandez-Diaz S., Hernan M.A., and Vandenbroucke J.P., *Commentary: A structural approach to Berkson's fallacy and a guide to a history of opinions about it.* International Journal of Epidemiology, 2014. **43**(2): p. 515-521.
95. Mann J.R., Mcdermott S., Barnes T.L., Hardin J., Bao H.K., and Zhou L., *Trichomoniasis in Pregnancy and Mental Retardation in Children.* Annals of Epidemiology, 2009. **19**(12): p. 891-899.
96. Mann J.R., Pan C., Rao G.A., McDermott S., and Hardin J.W., *Children Born to Diabetic Mothers May be More Likely to Have Intellectual Disability.* Maternal and Child Health Journal, 2013. **17**(5): p. 928-932.
97. Huang J.C., Zhu T.T., Qu Y., and Mu D.Z., *Prenatal, Perinatal and Neonatal Risk Factors for Intellectual Disability: A Systemic Review and Meta-Analysis.* Plos One, 2016. **11**(4).

98. Braun J.M., Daniels J.L., Kalkbrenner A., Zimmerman J., and Nicholas J.S., *The effect of maternal smoking during pregnancy on intellectual disabilities among 8-year-old children.* Paediatric and Perinatal Epidemiolgy, 2009. **23**(5): p. 482-91.
99. Lundberg F., Cnattingius S., D'Onofrio B., Altman D., Lambe M., Hultman C., et al., *Maternal smoking during pregnancy and intellectual performance in young adult Swedish male offspring.* Paediatric and Perinatal Epidemiology, 2010. **24**(1): p. 79-87.
100. Lipsitch M., Tchetgen E., and Cohen T., *Negative controls: a tool for detecting confounding and bias in observational studies.* Epidemiology, 2010. **21**(3): p. 383-8.
101. Davey Smith G., *Negative control exposures in epidemiologic studies.* Epidemiology, 2012. **23**(2): p. 350-1; author reply 351-2.
102. Begg M.D. and Parides M.K., *Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data.* Statistics in Medicine, 2003. **22**(16): p. 2591-602.
103. Carlin J.B., Gurrin L.C., Sterne J.A., Morley R., and Dwyer T., *Regression models for twin studies: a critical review.* International Journal of Epidemiology, 2005. **34**(5): p. 1089-99.
104. Hirvonen M., Ojala R., Korhonen P., Haataja P., Eriksson K., Rantanen K., et al., *Intellectual disability in children aged less than seven years born moderately and late preterm compared with very preterm and term-born children - a nationwide birth cohort study.* Journal of intellectual disability research, 2017. **61**(11): p. 1034-1054.
105. Bouyssi-Kobar M., du Plessis A.J., McCarter R., Brossard-Racine M., Murnick J., Tinkleman L., et al., *Third Trimester Brain Growth in Preterm Infants Compared With In Utero Healthy Fetuses.* Pediatrics, 2016. **138**(5).
106. Andescavage N.N., du Plessis A., McCarter R., Serag A., Evangelou I., Vezina G., et al., *Complex Trajectories of Brain Development in the Healthy Human Fetus.* Cereb Cortex, 2017. **27**(11): p. 5274-5283.
107. Westreich D. and Greenland S., *The table 2 fallacy: presenting and interpreting confounder and modifier coefficients.* Am J Epidemiol, 2013. **177**(4): p. 292-8.
108. Schisterman E.F., Cole S.R., and Platt R.W., *Overadjustment bias and unnecessary adjustment in epidemiologic studies.* Epidemiology, 2009. **20**(4): p. 488-95.
109. Hiscock R., Bauld L., Amos A., Fidler J.A., and Munafo M., *Socioeconomic status and smoking: a review.* Ann N Y Acad Sci, 2012. **1248**: p. 107-23.
110. Lu Y., Tong S., and Oldenburg B., *Determinants of smoking and cessation during and after pregnancy.* Health Promotional International, 2001. **16**(4): p. 355-65.
111. Hair N.L., Hanson J.L., Wolfe B.L., and Pollak S.D., *Association of Child Poverty, Brain Development, and Academic Achievement.* JAMA Pediatrics, 2015. **169**(9): p. 822-9.

112. Eriksen H.L., Kesmodel U.S., Underbjerg M., Kilburn T.R., Bertrand J., and Mortensen E.L., *Predictors of intelligence at the age of 5: family, pregnancy and birth characteristics, postnatal influences, and postnatal growth.* PLoS One, 2013. **8**(11): p. e79200.
113. Hackman D.A., Gallop R., Evans G.W., and Farah M.J., *Socioeconomic status and executive function: developmental trajectories and mediation.* Dev Sci, 2015. **18**(5): p. 686-702.
114. Sellers R., Warne N., Rice F., Langley K., Maughan B., Pickles A., et al., *Using a cross-cohort comparison design to test the role of maternal smoking in pregnancy in child mental health and learning: evidence from two UK cohorts born four decades apart.* International Journal of Epidemiology, 2020.
115. Prior M., *Is there an increase in the prevalence of autism spectrum disorders?* J Paediatr Child Health, 2003. **39**(2): p. 81-2.
116. Polanczyk G.V., Willcutt E.G., Salum G.A., Kieling C., and Rohde L.A., *ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis.* International Journal of Epidemiology, 2014. **43**(2): p. 434-42.
117. Elsabbagh M., Divan G., Koh Y.J., Kim Y.S., Kauchali S., Marcin C., et al., *Global prevalence of autism and other pervasive developmental disorders.* Autism Res, 2012. **5**(3): p. 160-79.
118. Lasser K., Boyd J.W., Woolhandler S., Himmelstein D.U., McCormick D., and Bor D.H., *Smoking and mental illness: A population-based prevalence study.* JAMA, 2000. **284**(20): p. 2606-10.
119. Goodwin R.D., Keyes K., and Simuro N., *Mental disorders and nicotine dependence among pregnant women in the United States.* Obstetrics and Gynecology, 2007. **109**(4): p. 875-83.
120. Pedersen W. and von Soest T., *Smoking, nicotine dependence and mental health among young adults: a 13-year population-based longitudinal study.* Addiction, 2009. **104**(1): p. 129-137.
121. Le Cook B., Wayne G.F., Kafali E.N., Liu Z.M., Shu C., and Flores M., *Trends in Smoking Among Adults With Mental Illness and Association Between Mental Health Treatment and Smoking Cessation.* Jama-Journal of the American Medical Association, 2014. **311**(2): p. 172-182.
122. Hill W.D., Harris S.E., and Deary I.J., *What genome-wide association studies reveal about the association between intelligence and mental health.* Current Opinion in Psychology, 2019. **27**: p. 25-30.

123. Gutierrez-Galve L., Stein A., Hanington L., Heron J., and Ramchandani P., *Paternal depression in the postnatal period and child development: mediators and moderators.* Pediatrics, 2015. **135**(2): p. e339-47.
124. Abdullahi I., Leonard H., Cherian S., Mutch R., Glasson E.J., de Klerk N., et al., *The Risk of Neurodevelopmental Disabilities in Children of Immigrant and Refugee Parents: Current Knowledge and Directions for Future Research.* Review Journal of Autism and Developmental Disorders, 2018. **5**(1): p. 29-42.
125. Melchior M., Chollet A., Glangeaud-Freudenthal N., Saurel-Cubizolles M.J., Dufourg M.N., van der Waerden J., et al., *Tobacco and alcohol use in pregnancy in France: the role of migrant status: the nationally representative ELFE study.* Addictive Behaviors, 2015. **51**: p. 65-71.
126. Troe E.J., Raat H., Jaddoe V.W., Hofman A., Steegers E.A., Verhulst F.C., et al., *Smoking during pregnancy in ethnic populations: the Generation R study.* Nicotine Tob Res, 2008. **10**(8): p. 1373-84.
127. McGrath J.J., Petersen L., Agerbo E., Mors O., Mortensen P.B., and Pedersen C.B., *A comprehensive assessment of parental age and psychiatric disorders.* JAMA Psychiatry, 2014. **71**(3): p. 301-9.
128. Merikangas A.K., Calkins M.E., Bilker W.B., Moore T.M., Gur R.C., and Gur R.E., *Parental Age and Offspring Psychopathology in the Philadelphia Neurodevelopmental Cohort.* Journal of the American Academy of Child and Adolescent Psychiatry, 2017. **56**(5): p. 391-400.
129. Belmont L. and Marolla F.A., *Birth order, family size, and intelligence.* Science, 1973. **182**(4117): p. 1096-101.
130. Belmont L., Stein Z.A., and Susser M.W., *Comparison of associations of birth order with intelligence test score and height.* Nature, 1975. **255**(5503): p. 54-6.
131. Smith R., *Employee earnings in the UK: 2019. Measures of employee earnings, using data from the Annual Survey for Hours and Earnings (ASHE).* 2019, Office for National Statistics: ons.gov.uk.
132. Kahn R.S., Certain L., and Whitaker R.C., *A reexamination of smoking before, during, and after pregnancy.* American Journal of Public Health, 2002. **92**(11): p. 1801-1808.
133. Jones K.L. and Smith D.W., *Recognition of the fetal alcohol syndrome in early infancy.* Lancet (London, England), 1973. **302**(7836): p. 999-1001.
134. Weyrauch D., Schwartz M., Hart B., Klug M.G., and Burd L., *Comorbid Mental Disorders in Fetal Alcohol Spectrum Disorders: A Systematic Review.* Journal of developmental and behavioral pediatrics : JDBP, 2017. **38**(4): p. 283-291.
135. Lawlor D.A., Tilling K., and Davey Smith G., *Triangulation in aetiological epidemiology.* Int J Epidemiol, 2016. **45**(6): p. 1866-1886.

136. Hernán M. and Robins J., *Causal Inference: What If*. November 10, 2019 ed. 2020, Boca Raton: Chapman & Hall/CRC.
137. Rubin D.B., *Randomization Analysis of Experimental-Data - the Fisher Randomization Test - Comment.* Journal of the American Statistical Association, 1980. **75**(371): p. 591-593.
138. Cox D.R., *Planning of experiments*. 1958, New York, NY: John Wiley and Sons.
139. Greenland S., Robins J.M., and Pearl J., *Confounding and collapsibility in causal inference.* Statistical Science, 1999: p. 29-46.
140. Greenland S., *Interpretation and choice of effect measures in epidemiologic analyses.* Am J Epidemiol, 1987. **125**(5): p. 761-8.
141. Pearl J., *Causality : models, reasoning, and inference*. 2000, Cambridge, U.K. ; New York: Cambridge University Press. xvi, 384 p.
142. Flanders W.D. and Eldridge R.C., *Summary of relationships between exchangeability, biasing paths and bias.* Eur J Epidemiol, 2015. **30**(10): p. 1089-99.
143. Hernán M.A. and Cole S.R., *Invited commentary: causal diagrams and measurement bias.* American journal of epidemiology, 2009. **170**(8): p. 959-962.
144. Shahar E., *Causal diagrams for encoding and evaluation of information bias.* Journal of Evaluation in Clinical Practice, 2009. **15**(3): p. 436-440.
145. Rubin D.B., *Inference and Missing Data.* Biometrika, 1976. **63**(3): p. 581-590.
146. Little R.J., D'Agostino R., Cohen M.L., Dickersin K., Emerson S.S., Farrar J.T., et al., *The prevention and treatment of missing data in clinical trials.* N Engl J Med, 2012. **367**(14): p. 1355-60.
147. Bartlett J.W., Harel O., and Carpenter J.R., *Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression.* Am J Epidemiol, 2015. **182**(8): p. 730-6.
148. Carpenter J. and Kenward M., *Multiple imputation and its application*. 2012: John Wiley & Sons.
149. White I.R. and Carlin J.B., *Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values.* Statistics in medicine, 2010. **29**(28): p. 2920-31.
150. Hughes R.A., Heron J., Sterne J.A.C., and Tilling K., *Accounting for missing data in statistical analyses: multiple imputation is not always the answer.* Int J Epidemiol, 2019.
151. Hunter S.C. *Missing data within participants - how much is OK?* 2013 [cited 2018 January 29]; Available from: https://www.researchgate.net/post/Missing_data_within_participants-how_much_is_OK.

152. purplesocks. *How much missing data is too much? Multiple Imputation (MICE) & R*. 2015 [cited 2018 January 29]; Available from: https://stats.stackexchange.com/questions/149140/how-much-missing-data-is-too-much-multiple-imputation-mice-r.
153. shuvayan. *What should be the allowed percentage of Missing Values?* 2015 [cited 2018 January 29]; Available from: https://discuss.analyticsvidhya.com/t/what-should-be-the-allowed-percentage-of-missing-values/2456.
154. Zingora T. *What proportion of missing data is too big for multiple imputation in longitudinal data?* 2016 [cited 2018 January 29]; Available from: https://www.researchgate.net/post/What_proportion_of_missing_data_is_too_big_for_multiple_imputation_in_longitudinal_data.
155. Seaman S.R. and White I.R., *Review of inverse probability weighting for dealing with missing data.* Stat Methods Med Res, 2013. **22**(3): p. 278-95.
156. White I.R., Royston P., and Wood A.M., *Multiple imputation using chained equations: Issues and guidance for practice.* Statistics in Medicine, 2011. **30**(4): p. 377-399.
157. Madley-Dowd P., Hughes R., Tilling K., and Heron J., *The proportion of missing data should not be used to guide decisions on multiple imputation.* J Clin Epidemiol, 2019.
158. Wolke D., Waylen A., Samara M., Steer C., Goodman R., Ford T., et al., *Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders.* Br J Psychiatry, 2009. **195**(3): p. 249-56.
159. Wood M.E., Lapane K.L., van Gelder M., Rai D., and Nordeng H.M.E., *Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control.* Pharmacoepidemiol Drug Saf, 2018. **27**(2): p. 140-147.
160. Gage S.H., Munafo M.R., and Davey Smith G., *Causal Inference in Developmental Origins of Health and Disease (DOHaD) Research.* Annu Rev Psychol, 2016. **67**: p. 567-85.
161. Keyes K.M., Smith G.D., and Susser E., *Commentary: Smoking in pregnancy and offspring health: early insights into family-based and 'negative control' studies?* Int J Epidemiol, 2014. **43**(5): p. 1381-8.
162. Taylor O., Keatley D.A., and Clarke D.D., *A Behavior Sequence Analysis of Perceptions of Alcohol-Related Violence Surrounding Drinking Establishments.* Journal of interpersonal violence, 2017: p. 886260517702490.
163. Richmond R.C., Sharp G.C., Ward M.E., Fraser A., Lyttleton O., McArdle W.L., et al., *DNA Methylation and BMI: Investigating Identified Methylation Sites at HIF3A in a Causal Framework.* Diabetes, 2016. **65**(5): p. 1231-44.

164. Sanderson E., Macdonald-Wallis C., and Davey Smith G., *Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration.* Int J Epidemiol, 2018. **47**(2): p. 587-596.
165. Davey Smith G. and Ebrahim S., *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* International Journal of Epidemiology, 2003. **32**(1): p. 1-22.
166. Davey Smith G. and Hemani G., *Mendelian randomization: genetic anchors for causal inference in epidemiological studies.* Human Molecular Genetics, 2014. **23**(R1): p. R89-98.
167. Loukola A., Hallfors J., Korhonen T., and Kaprio J., *Genetics and smoking.* Curr Addict Rep, 2014. **1**(1): p. 75-82.
168. Liu M., Jiang Y., Wedow R., Li Y., Brazel D.M., Chen F., et al., *Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use.* Nat Genet, 2019. **51**(2): p. 237-244.
169. Morris T.T., Davies N.M., Hemani G., and Smith G.D., *Population phenomena inflate genetic associations of complex social traits.* Science Advances, 2020. **6**(16): p. eaay0328.
170. Haworth S., Mitchell R., Corbin L., Wade K.H., Dudding T., Budu-Aggrey A., et al., *Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis.* Nat Commun, 2019. **10**(1): p. 333.
171. Davies N.M., Holmes M.V., and Davey Smith G., *Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians.* British Medical Journal, 2018. **362**: p. k601.
172. Gage S.H., Bowden J., Davey Smith G., and Munafo M.R., *Investigating causality in associations between education and smoking: a two-sample Mendelian randomization study.* Int J Epidemiol, 2018. **47**(4): p. 1131-1140.
173. Burgess S., Butterworth A., and Thompson S.G., *Mendelian randomization analysis with multiple genetic variants using summarized data.* Genet Epidemiol, 2013. **37**(7): p. 658-65.
174. Hartwig F.P., Davies N.M., Hemani G., and Davey Smith G., *Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique.* Int J Epidemiol, 2016. **45**(6): p. 1717-1726.
175. Bowden J., Del Greco M.F., Minelli C., Davey Smith G., Sheehan N.A., and Thompson J.R., *Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic.* Int J Epidemiol, 2016. **45**(6): p. 1961-1974.

176. Bowden J., Davey Smith G., and Burgess S., *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* Int J Epidemiol, 2015. **44**(2): p. 512-25.
177. Bowden J., Davey Smith G., Haycock P.C., and Burgess S., *Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator.* Genet Epidemiol, 2016. **40**(4): p. 304-14.
178. Kang H., Zhang A., Cai T.T., and Small D.S., *Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization.* Journal of the American statistical Association, 2016. **111**(513): p. 132-144.
179. Hartwig F.P., Davey Smith G., and Bowden J., *Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption.* Int J Epidemiol, 2017. **46**(6): p. 1985-1998.
180. Sjölander A., Frisell T., and Öberg S., *Causal Interpretation of Between-Within Models for Twin Research.* Epidemiologic Methods, 2012. **1**(1).
181. Frisell T., Oberg S., Kuja-Halkola R., and Sjolander A., *Sibling comparison designs: bias from non-shared confounders and measurement error.* Epidemiology, 2012. **23**(5): p. 713-20.
182. Sjolander A., Frisell T., Kuja-Halkola R., Oberg S., and Zetterqvist J., *Carryover Effects in Sibling Comparison Designs.* Epidemiology, 2016. **27**(6): p. 852-858.
183. Juarez S.P. and Merlo J., *Revisiting the effect of maternal smoking during pregnancy on offspring birthweight: a quasi-experimental sibling analysis in Sweden.* PLoS One, 2013. **8**(4): p. e61734.
184. Foulds J., Ramstrom L., Burke M., and Fagerstrom K., *Effect of smokeless tobacco (snus) on smoking and public health in Sweden.* Tob Control, 2003. **12**(4): p. 349-59.
185. Gilljam H. and Galanti M.R., *Role of snus (oral moist snuff) in smoking cessation and smoking reduction in Sweden.* Addiction, 2003. **98**(9): p. 1183-9.
186. Lund K.E., McNeill A., and Scheffels J., *The use of snus for quitting smoking compared with medicinal products.* Nicotine Tob Res, 2010. **12**(8): p. 817-22.
187. Ramstrom L.M. and Foulds J., *Role of snus in initiation and cessation of tobacco smoking in Sweden.* Tob Control, 2006. **15**(3): p. 210-4.
188. Holm H., Jarvis M.J., Russell M.A., and Feyerabend C., *Nicotine intake and dependence in Swedish snuff takers.* Psychopharmacology (Berl), 1992. **108**(4): p. 507-11.
189. Merrell D.J., *Measurement of Sexual Isolation and Selective Mating.* Evolution, 1950. **4**(4): p. 326-331.

190. Agrawal A., Heath A.C., Grant J.D., Pergadia M.L., Statham D.J., Bucholz K.K., et al., *Assortative mating for cigarette smoking and for alcohol consumption in female Australian twins and their spouses.* Behav Genet, 2006. **36**(4): p. 553-66.
191. Reynolds C.A., Barlow T., and Pedersen N.L., *Alcohol, tobacco and caffeine use: Spouse similarity processes.* Behavior Genetics, 2006. **36**(2): p. 201-215.
192. Sutton G.C., *Assortative marriage for smoking habits.* Ann Hum Biol, 1980. **7**(5): p. 449-56.
193. Tooley E.M. and Borrelli B., *Characteristics of Cigarette Smoking in Individuals in Smoking Concordant and Smoking Discordant Couples.* Couple Family Psychol, 2017. **6**(2): p. 106-116.
194. Ask H., Rognmo K., Torvik F.A., Roysamb E., and Tambs K., *Non-Random Mating and Convergence Over Time for Alcohol Consumption, Smoking, and Exercise: The Nord-Trondelag Health Study.* Behavior Genetics, 2012. **42**(3): p. 354-365.
195. Grant J.D., Heath A.C., Bucholz K.K., Madden P.A.F., Agrawal A., Statham D.J., et al., *Spousal concordance for alcohol dependence: Evidence for assortative mating or spousal interaction effects?* Alcoholism-Clinical and Experimental Research, 2007. **31**(5): p. 717-728.
196. Hall R.L., Hesselbrock V.M., and Stabenau J.R., *Familial Distribution of Alcohol-Use .1. Assortative Mating in the Parents of Alcoholics.* Behavior Genetics, 1983. **13**(4): p. 361-372.
197. Hall R.L., Hesselbrock V.M., and Stabenau J.R., *Familial Distribution of Alcohol-Use .2. Assortative Mating of Alcoholic Probands.* Behavior Genetics, 1983. **13**(4): p. 373-382.
198. Howe L.J., Lawson D.J., Davies N.M., St. Pourcain B., Lewis S.J., Davey Smith G., et al., *Alcohol consumption and mate choice in UK Biobank: comparing observational and Mendelian randomization estimates.* 2019: p. 418269.
199. Mcleod J.D., *Spouse Concordance for Alcohol Dependence and Heavy Drinking - Evidence from a Community Sample.* Alcoholism-Clinical and Experimental Research, 1993. **17**(6): p. 1146-1155.
200. Jeffery R.W. and Rick A.M., *Cross-sectional and longitudinal associations between body mass index and marriage-related factors.* Obesity Research, 2002. **10**(8): p. 809-815.
201. Oreffice S. and Quintana-Domeque C., *Anthropometry and socioeconomics among couples: evidence in the United States.* Econ Hum Biol, 2010. **8**(3): p. 373-84.
202. Silventoinen K., Kaprio J., Lahelma E., Viken R.J., and Rose R.J., *Assortative mating by body height and BMI: Finnish twins and their spouses.* Am J Hum Biol, 2003. **15**(5): p. 620-7.

203. Buss D.M., *Human Mate Selection: Opposites are sometimes said to attract, but in fact we are likely to marry someone who is similar to us in almost every variable.* American Scientist, 1985. **73**(1): p. 47-51.
204. Nordsletten A.E., Larsson H., Crowley J.J., Almqvist C., Lichtenstein P., and Mataix-Cols D., *Patterns of Nonrandom Mating Within and Across 11 Major Psychiatric Disorders.* Jama Psychiatry, 2016. **73**(4): p. 354-361.
205. de Jong S., Diniz M.J.A., Rodrigues A.C.S., Gadelha A., Santoro M., Ota V., et al., *Polygenic Risk Scores Reveal Patterns of Assortative Mating And Anticipation In A Large Pedigree Affected With Mood Disorders.* 2017. **27**: p. S483.
206. Mare R.D., *5 Decades of Educational Assortative Mating.* American Sociological Review, 1991. **56**(1): p. 15-32.
207. Vandenburg S.G., *Assortative mating, or who marries whom?* Behav Genet, 1972. **2**(2): p. 127-57.
208. Watson D., Klohnen E.C., Casillas A., Simms E.N., Haig J., and Berry D.S., *Match makers and deal breakers: analyses of assortative mating in newlywed couples.* J Pers, 2004. **72**(5): p. 1029-68.
209. Whyte S. and Torgler B., *Things change with age: Educational assortment in online dating.* Personality and Individual Differences, 2017. **109**: p. 5-11.
210. Rolan-Alvarez E. and Caballero M., *Estimating sexual selection and sexual isolation effects from mating frequencies.* Evolution, 2000. **54**(1): p. 30-36.
211. Rolan-Alvarez E., Carvajal-Rodriguez A., de Coo A., Cortes B., Estevez D., Ferreira M., et al., *The scale-of-choice effect and how estimates of assortative mating in the wild can be biased due to heterogeneous samples.* Evolution, 2015. **69**(7): p. 1845-1857.
212. White I.R., *simsum: Analyses of simulation studies including Monte Carlo error.* Stata Journal, 2010. **10**(3): p. 369-385.
213. Langley K., Heron J., Smith G.D., and Thapar A., *Maternal and paternal smoking during pregnancy and risk of ADHD symptoms in offspring: testing for intrauterine effects.* Am J Epidemiol, 2012. **176**(3): p. 261-8.
214. Cohen J.M., Wood M.E., Hernández-Díaz S., Ystrom E., and Nordeng H., *Paternal antidepressant use as a negative control for maternal use: assessing familial confounding on gestational length and anxiety traits in offspring.* International Journal of Epidemiology, 2019. **48**(5): p. 1665-1672.
215. Brew B.K. and Gong T., *Modelling paternal exposure as a negative control.* International Journal of Epidemiology, 2020.
216. Wood M.E., Cohen J.M., Ystrom E., Nordeng H.M.E., and Hernandez-Diaz S., *Response to: Modelling Paternal Exposure as a Negative Control.* International Journal of Epidemiology, 2020.

217. Brew B.K., Gong T., Williams D.M., Larsson H., and Almqvist C., *Using fathers as a negative control exposure to test the Developmental Origins of Health and Disease Hypothesis: A case study on maternal distress and offspring asthma using Swedish register data.* Scandinavian journal of public health, 2017. **45**(17_suppl): p. 36-40.
218. Madley-Dowd P., Hughes R., Tilling K., and Heron J., *The proportion of missing data should not be used to guide decisions on multiple imputation.* J Clin Epidemiol, 2019. **110**: p. 63-73.
219. Graham J.W., *Missing Data Analysis: Making It Work in the Real World.* Annual Review of Psychology, 2009. **60**: p. 549-576.
220. Donders A.R., van der Heijden G.J., Stijnen T., and Moons K.G., *Review: a gentle introduction to imputation of missing values.* J Clin Epidemiol, 2006. **59**(10): p. 1087-91.
221. Zhang Y., Florez I.D., Colunga Lozano L.E., Aloweni F.A.B., Kennedy S.A., Li A., et al., *A systematic survey on reporting and methods for handling missing participant data for continuous outcomes in randomized controlled trials.* J Clin Epidemiol, 2017. **88**: p. 57-66.
222. Rubin D.B., *Multiple imputation for nonresponse in surveys*. 1987, New York: Wiley.
223. Rubin D.B. and Little R.J., *Statistical analysis with missing data*. 2 ed. Hoboken, NJ: J Wiley & Sons. 2002.
224. Rubin D.B., *Multiple imputation after 18+ years.* Journal of the American Statistical Association, 1996. **91**(434): p. 473-489.
225. Moons K.G.M., Donders R.A.R.T., Stijnen T., and Harrell F.E., *Using the outcome for imputation of missing predictor values was preferred.* Journal of Clinical Epidemiology, 2006. **59**(10): p. 1092-1101.
226. Sterne J.A.C., White I.R., Carlin J.B., Spratt M., Royston P., Kenward M.G., et al., *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.* Bmj-British Medical Journal, 2009. **339**.
227. Tilling K., Williamson E.J., Spratt M., Sterne J.A., and Carpenter J.R., *Appropriate inclusion of interactions was needed to avoid bias in multiple imputation.* J Clin Epidemiol, 2016. **80**: p. 107-115.
228. Schafer J.L., *Multiple imputation: a primer.* Statistical methods in medical research, 1999. **8**(1): p. 3-15.
229. Alice M. *Imputing Missing Data with R; MICE package*. 2015 January 16, 2018 [cited 2018 January 29]; Available from: https://datascienceplus.com/imputing-missing-data-with-r-mice-package/.
230. Dong Y. and Peng C.Y., *Principled missing data methods for researchers.* Springerplus, 2013. **2**(1): p. 222.

231. Jakobsen J.C., Gluud C., Wetterslev J., and Winkel P., *When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts.* BMC medical research methodology, 2017. **17**(1): p. 162.
232. McNeish D., *Missing data methods for arbitrary missingness with small samples.* Journal of Applied Statistics, 2017. **44**(1): p. 24-39.
233. Clavel J., Merceron G., and Escarguel G., *Missing Data Estimation in Morphometrics: How Much is Too Much?* Systematic Biology, 2014. **63**(2): p. 203-218.
234. Mishra S. and Khare D., *On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: a simulation study.* Journal of Medical Statistics and Informatics, 2014. **2**(1): p. 9.
235. Lee J.H. and Huber Jr. J., *Multiple imputation with large proportions of missing data: How much is too much?*, in *United Kingdom Stata Users' Group Meetings 2011*. 2011, Stata Users Group.
236. Lee K.J. and Carlin J.B., *Recovery of information from multiple imputation: a simulation study.* Emerg Themes Epidemiol, 2012. **9**(1): p. 3.
237. Hardt J., Herke M., Brian T., and Laubach W., *Multiple imputation of missing data: a simulation study on a binary response.* Open Journal of Statistics, 2013. **3**(05): p. 370.
238. Wagner J., *The fraction of missing information as a tool for monitoring the quality of survey data.* Public Opinion Quarterly, 2010: p. nfq007.
239. Dempster A.P., Laird N.M., and Rubin D.B., *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the royal statistical society. Series B (methodological), 1977: p. 1-38.
240. StataCorp, *Stata Statistical Software: Release 14*. 2015, TX: Stata Press: College Station.
241. Bodner T.E., *What improves with increased missing data imputations?* Structural Equation Modeling, 2008. **15**(4): p. 651-675.
242. Boyd A., Golding J., Macleod J., Lawlor D.A., Fraser A., Henderson J., et al., *Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children.* International journal of epidemiology, 2013. **42**(1): p. 111-27.
243. Fraser A., Macdonald-Wallis C., Tilling K., Boyd A., Golding J., Davey Smith G., et al., *Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort.* Int J Epidemiol, 2013. **42**(1): p. 97-110.
244. Donald H.S., Giti C., and Colin S., *Concurrent Validity of the Wechsler Abbreviated Scale of Intelligence (WASI) with a Sample of Canadian Children.* Canadian Journal of School Psychology, 2000. **16**(1): p. 87-94.
245. Wechsler D., Golombok S., and Rust J., *WISC-III UK*. Sidcup, Kent: The Psychological Corporation, 1992.

246. Bishop D.V., *Development of the Children's Communication Checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children.* The Journal of Child Psychology and Psychiatry and Allied Disciplines, 1998. **39**(6): p. 879-891.
247. Buuren S.v. and Oudshoorn C., *Multivariate imputation by chained equations: MICE V1. 0 user's manual*. 2000, TNO.
248. Allison P.D., *Multiple imputation for missing data: A cautionary tale.* J Sociological Methods, 2000. **28**(3): p. 301-309.
249. Collins L.M., Schafer J.L., and Kam C.M., *A comparison of inclusive and restrictive strategies in modern missing data procedures.* Psychol Methods, 2001. **6**(4): p. 330-51.
250. Hardt J., Herke M., and Leonhart R., *Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research.* BMC Med Res Methodol, 2012. **12**: p. 184.
251. Jones M., Mishra G.D., and Dobson A., *Analytical results in longitudinal studies depended on target of inference and assumed mechanism of attrition.* J Clin Epidemiol, 2015. **68**(10): p. 1165-75.
252. Carpenter J.R., Kenward M.G., and White I.R., *Sensitivity analysis after multiple imputation under missing at random: a weighting approach.* Stat Methods Med Res, 2007. **16**(3): p. 259-75.
253. Munafo M.R. and Davey Smith G., *Robust research needs many lines of evidence.* Nature, 2018. **553**(7689): p. 399-401.
254. Schmidt M., Pedersen L., and Sorensen H.T., *The Danish Civil Registration System as a tool in epidemiology*. European Journal of Epidemiology, 2014. **29**(8): p. 541-9.
255. Bliddal M., Broe A., Pottegard A., Olsen J., and Langhoff-Roos J., *The Danish Medical Birth Register.* European Journal of Epidemiology, 2018. **33**(1): p. 27-36.
256. Mors O., Perto G.P., and Mortensen P.B., *The Danish Psychiatric Central Research Register.* Scandinavian Journal of Public Health, 2011. **39**(7 Suppl): p. 54-7.
257. Schmidt M., Schmidt S.A., Sandegaard J.L., Ehrenstein V., Pedersen L., and Sorensen H.T., *The Danish National Patient Registry: a review of content, data quality, and research potential.* Clinical Epidemiology, 2015. **7**: p. 449-90.
258. Jensen V.M. and Rasmussen A.W., *Danish Education Registers.* Scandinavian Journal of Public Health, 2011. **39**(7 Suppl): p. 91-4.
259. Baadsgaard M. and Quitzau J., *Danish registers on personal income and transfer payments.* Scandinavian Journal of Public Health, 2011. **39**(7 Suppl): p. 103-5.
260. Hannon E., Schendel D., Ladd-Acosta C., Grove J., Hansen C.S., Andrews S.V., et al., *Elevated polygenic burden for autism is associated with differential DNA methylation at birth.* Genome Medicine, 2018. **10**.

261. Pedersen C.B., Bybjerg-Grauholm J., Pedersen M.G., Grove J., Agerbo E., Baekvad-Hansen M., et al., *The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders.* Molecular Psychiatry, 2018. **23**(1): p. 6-14.
262. Statistics Denmark. *PERINDKIALT_13 - Total personal income*. 2018 [cited 2018 09 October]; Available from: https://www.dst.dk/da/Statistik/dokumentation/Times/personindkomst/perindkialt-13.
263. Pedersen C.B., Mors O., Bertelsen A., Waltoft B.L., Agerbo E., McGrath J.J., et al., *A comprehensive nationwide study of the incidence rate and lifetime risk for treated mental disorders.* JAMA Psychiatry, 2014. **71**(5): p. 573-81.
264. R Core Team, *R: A language and environment for statistical computing*. 2017, R Foundation for Statistical Computing.
265. Carey V.J., Lumley T., Ripley B., and Moler C., *Package 'gee': Generalized Estimation Equation Solver*. 2015, The Comprehensive R Archive Network: cran.r-project.org.
266. Therneau T.M., Lumley T., Atkinson E., and Crowson C., *Package 'survival: Survival Analysis*. 2015, The Comprehensive R Archive Network: cran.r-project.org.
267. Therneau T.M. and Grambsch P.M., *Modeling survival data : extending the Cox model*. Statistics for biology and health. 2000, New York: Springer. xiii, 350 p.
268. George L., Granath F., Johansson A.L.V., and Cnattingius S., *Self-reported nicotine exposure and plasma levels of cotinine in early and late pregnancy.* Acta Obstetricia Et Gynecologica Scandinavica, 2006. **85**(11): p. 1331-1337.
269. Lawlor D., Richmond R., Warrington N., McMahon G., Davey Smith G., Bowden J., et al., *Using Mendelian randomization to determine causal effects of maternal pregnancy (intrauterine) exposures on offspring outcomes: Sources of bias and methods for assessing them.* Wellcome open research, 2017. **2**: p. 11.
270. Smith G.D. and Ebrahim S., *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* Int J Epidemiol, 2003. **32**(1): p. 1-22.
271. Demontis D., Walters R.K., Martin J., Mattheisen M., Als T.D., Agerbo E., et al., *Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder.* Nature Genetics, 2019. **51**(1): p. 63-75.
272. Leppert B., Havdahl A., Riglin L., Jones H.J., Zheng J., Davey Smith G., et al., *Association of Maternal Neurodevelopmental Risk Alleles With Early-Life Exposures.* JAMA Psychiatry, 2019.

273. Burgess S. and Thompson S.G., *Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects.* American Journal of Epidemiology, 2015. **181**(4): p. 251-60.
274. Baba S., Wikstrom A.K., Stephansson O., and Cnattingius S., *Influence of smoking and snuff cessation on risk of preterm birth.* Eur J Epidemiol, 2012. **27**(4): p. 297-304.
275. Wikstrom A.K., Cnattingius S., Galanti M.R., Kieler H., and Stephansson O., *Effect of Swedish snuff (snus) on preterm birth.* BJOG, 2010. **117**(8): p. 1005-10.
276. Baba S., Wikstrom A.K., Stephansson O., and Cnattingius S., *Changes in snuff and smoking habits in Swedish pregnant women and risk for small for gestational age births.* BJOG, 2013. **120**(4): p. 456-62.
277. Baba S., Wikstrom A.K., Stephansson O., and Cnattingius S., *Influence of snuff and smoking habits in early pregnancy on risks for stillbirth and early neonatal mortality.* Nicotine Tob Res, 2014. **16**(1): p. 78-83.
278. Wikstrom A.K., Cnattingius S., and Stephansson O., *Maternal use of Swedish snuff (snus) and risk of stillbirth.* Epidemiology, 2010. **21**(6): p. 772-8.
279. England L.J., Levine R.J., Mills J.L., Klebanoff M.A., Yu K.F., and Cnattingius S., *Adverse pregnancy outcomes in snuff users.* American Journal of Obstetrics and Gynecology, 2003. **189**(4): p. 939-943.
280. The National Board of Health and Welfare. *The Swedish Medical Birth Register - A Summary of Content and Quality*. 2003 [cited 2019 19/03]; Available from: https://www.socialstyrelsen.se/publikationer2003/2003-112-3.
281. Ludvigsson J.F., Andersson E., Ekbom A., Feychting M., Kim J.L., Reuterwall C., et al., *External review and validation of the Swedish national inpatient register.* BMC Public Health, 2011. **11**: p. 450.
282. Statistics Sweden, *Longitudinal integration database for health insurance and labour market studies (LISA by Swedish acronym)*. 2012, English.
283. Ekbom A., *The Swedish Multi-generation Register.* Methods Mol Biol, 2011. **675**: p. 215-20.
284. Marsal K., Persson P.H., Larsen T., Lilja H., Selbing A., and Sultan B., *Intrauterine growth curves based on ultrasonically estimated foetal weights.* Acta Paediatrica, 1996. **85**(7): p. 843-848.
285. Keyes K.M., Smith G.D., and Susser E., *On Sibling Designs.* Epidemiology, 2013. **24**(3): p. 473-474.
286. Northstone K., Lewcock M., Groom A., Boyd A., Macleod J., Timpson N., et al., *The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019.* Wellcome Open Res, 2019. **4**: p. 51.

287. Williams E., Thomas K., Sidebotham H., and Emond A., *Prevalence and characteristics of autistic spectrum disorders in the ALSPAC cohort.* Dev Med Child Neurol, 2008. **50**(9): p. 672-7.
288. NHS Health Research Authority. *Confidentiality Advisory Group*. 2017 [cited 2020 22 July]; Available from: https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/.
289. NHS Digital. *Appendix 1: Section 251 of the National Health Service Act 2006*. 2019 [cited 2020 28/04]; Available from: https://digital.nhs.uk/services/data-access-request-service-dars/how-the-national-data-opt-out-affects-data-released-by-nhs-digital/national-data-opt-out-guidance-for-researchers/appendix-1-section-251-of-the-national-health-service-act-2006.
290. University of Bristol (ALSPAC). *Linkage to routine health and social records*. [cited 2020 22 July]; Available from: http://www.bristol.ac.uk/alspac/researchers/our-data/linkage/.
291. NHS Health Research Authority. *Integrated Research Application System*. 2018 [cited 2020 22 July]; Available from: https://www.hra.nhs.uk/about-us/committees-and-services/integrated-research-application-system/.
292. Department for Education and Skills. *Data Collection by Type of Special Educational Need (DfES/0536/2003)*. 2003 [cited 2019 12 June]; Available from: http://www.education.gov.uk/publications/eOrderingDownload/DFES-0536-2003.rtf.
293. Chisholm J., *The Read clinical classification.* BMJ: British Medical Journal, 1990. **300**(6732): p. 1092.
294. NHS Digital. *Read Codes*. 2018 [cited 2020 28/04]; Available from: https://digital.nhs.uk/services/terminology-and-classifications/read-codes.
295. Sheehan R., Hassiotis A., Walters K., Osborn D., Strydom A., and Horsfall L., *Mental illness, challenging behaviour, and psychotropic drug prescribing in people with intellectual disability: UK population based cohort study.* BMJ, 2015. **351**: p. h4326.
296. Digital N. *Hospital Episode Statistics (HES)*. 2019 [cited 2020 15 June]; Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics.
297. Boyd A., *Understanding Hospital Episode Statistics (HES)*. 2017, CLOSER: London, UK.
298. NHS Digital. *Mental Health Services Data Set*. 2020 [cited 2020 28/04]; Available from: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/mental-health-services-data-set.

299. Wing J., Beevor A., Curtis R., Park S., Hadden J., and Burns A., *Health of the Nation Outcome Scales (HoNOS): research and development.* The British Journal of Psychiatry, 1998. **172**(1): p. 11-18.
300. Cornish R.P., Macleod J., Carpenter J.R., and Tilling K., *Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: a simulation study.* Emerg Themes Epidemiol, 2017. **14**: p. 14.
301. Passaro K.T., Noss J., Savitz D.A., and Little R.E., *Agreement between self and partner reports of paternal drinking and smoking. The ALSPAC Study Team. Avon Longitudinal Study of Pregnancy and Childhood.* Int J Epidemiol, 1997. **26**(2): p. 315-20.
302. Cox J.L., Chapman G., Murray D., and Jones P., *Validation of the Edinburgh Postnatal Depression Scale (EPDS) in non-postnatal women.* J Affect Disord, 1996. **39**(3): p. 185-9.
303. Office of population censuses and surveys and General register office for Scotland, *1991 Census: Definitions Great Britain*. 1992, HMSO: London.
304. Paternoster L., Evans D.M., Nohr E.A., Holst C., Gaborieau V., Brennan P., et al., *Genome-wide population-based association study of extremely overweight young adults--the GOYA study.* PLoS One, 2011. **6**(9): p. e24303.
305. The 1000 Genomes Project Consortium, Abecasis G.R., Auton A., Brooks L.D., DePristo M.A., Durbin R.M., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
306. R Core Team, *R: A language and environment for statistical computing*. 2019, R Foundation for Statistical Computing.
307. van Buuren S., *Multiple imputation of discrete and continuous data by fully conditional specification.* Stat Methods Med Res, 2007. **16**(3): p. 219-42.
308. van Buuren S. and Groothuis-Oudshoorn K., *Package 'mice': Multivariate Imputation by Chained Equations*. 2020, The Comprehensive R Archive Network: cran.r-project.org.
309. Meng X.-L., *Multiple-imputation inferences with uncongenial sources of input.* Statistical Science, 1994: p. 538-558.
310. Xie X. and Meng X.-L., *Dissecting multiple imputation from a multi-phase inference perspective: what happens when god's, imputer's and analyst's models are uncongenial?* Statistica Sinica, 2017: p. 1485-1545.
311. Floden L. and Bell M.L., *Imputation strategies when a continuous outcome is to be dichotomized for responder analysis: a simulation study.* BMC Med Res Methodol, 2019. **19**(1): p. 161.

312. Grobler A.C. and Lee K., *Multiple imputation in the presence of an incomplete binary variable created from an underlying continuous variable.* Biom J, 2020. **62**(2): p. 467-478.
313. Hemani G., Zheng J., Wade K.H., Laurin C., Elsworth B., Burgess S., et al., *MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations.* bioRxiv, 2016.
314. Cochran W.G., *The combination of estimates from different experiments.* Biometrics, 1954. **10**(1): p. 101-129.
315. Burgess S., Davies N.M., and Thompson S.G., *Bias due to participant overlap in two-sample Mendelian randomization.* Genet Epidemiol, 2016. **40**(7): p. 597-608.
316. Bartlett J.W., Seaman S.R., White I.R., Carpenter J.R., and Alzheimer's Disease Neuroimaging I., *Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model.* Stat Methods Med Res, 2015. **24**(4): p. 462-87.
317. D'Onofrio B.M., Van Hulle C.A., Waldman I.D., Rodgers J.L., Harden K.P., Rathouz P.J., et al., *Smoking during pregnancy and offspring externalizing problems: an exploration of genetic and environmental confounds.* Development and Psychopathology, 2008. **20**(1): p. 139-64.
318. Kuja-Halkola R., D'Onofrio B.M., Larsson H., and Lichtenstein P., *Maternal smoking during pregnancy and adverse outcomes in offspring: genetic and environmental sources of covariance.* Behavioural Genetics, 2014. **44**(5): p. 456-67.
319. Yang Q., Millard L.A.C., and Davey Smith G., *Proxy gene-by-environment Mendelian randomization study confirms a causal effect of maternal smoking on offspring birthweight, but little evidence of long-term influences on offspring health.* Int J Epidemiol, 2019.
320. Taylor A.E., Carslake D., de Mola C.L., Rydell M., Nilsen T.I.L., Bjorngaard J.H., et al., *Maternal Smoking in Pregnancy and Offspring Depression: a cross cohort and negative control study.* Sci Rep, 2017. **7**(1): p. 12579.
321. Laubenthal J., Zlobinskaya O., Poterlowicz K., Baumgartner A., Gdula M.R., Fthenou E., et al., *Cigarette smoke-induced transgenerational alterations in genome stability in cord blood of human F1 offspring.* FASEB J, 2012. **26**(10): p. 3946-56.
322. Liu X., Conner H., Kobayashi T., Kim H., Wen F., Abe S., et al., *Cigarette smoke extract induces DNA damage but not apoptosis in human bronchial epithelial cells.* Am J Respir Cell Mol Biol, 2005. **33**(2): p. 121-9.
323. European Parliament and Council of the European Union, *General Data Protection Regulation (GDPR)*. 2016, Intersoft Consulting: https://gdpr-info.eu/.
324. Suissa S., *Immortal Time Bias in Pharmacoepidemiology.* American Journal of Epidemiology, 2007. **167**(4): p. 492-499.

325. Liew Z., Olsen J., Cui X., Ritz B., and Arah O.A., *Bias from conditioning on live birth in pregnancy cohorts: an illustration based on neurodevelopment in children after prenatal exposure to organic pollutants.* Int J Epidemiol, 2015. **44**(1): p. 345-54.
326. Nobles J. and Hamoudi A., *Detecting the Effects of Early-Life Exposures: Why Fecundity Matters.* Population Research and Policy Review, 2019. **38**(6): p. 783-809.
327. Kline J., Stein Z.A., Susser M., and Warburton D., *Smoking: a risk factor for spontaneous abortion.* N Engl J Med, 1977. **297**(15): p. 793-6.
328. Weselak M., Arbuckle T.E., Walker M.C., and Krewski D., *The influence of the environment and other exogenous agents on spontaneous abortion risk.* J Toxicol Environ Health B Crit Rev, 2008. **11**(3-4): p. 221-41.
329. Wickham S., Anwar E., Barr B., Law C., and Taylor-Robinson D., *Poverty and child health in the UK: using evidence for action.* Arch Dis Child, 2016. **101**(8): p. 759-66.
330. D'Onofrio B.M., Lahey B.B., Turkheimer E., and Lichtenstein P., *Critical need for family-based, quasi-experimental designs in integrating genetic and social science research.* American Journal of Public Health, 2013. **103 Suppl 1**: p. S46-55.
331. Warrington N.M., Beaumont R.N., Horikoshi M., Day F.R., Helgeland O., Laurin C., et al., *Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors.* Nat Genet, 2019. **51**(5): p. 804-814.
332. Caraballo R.S., Giovino G.A., Pechacek T.F., and Mowery P.D., *Factors associated with discrepancies between self-reports on cigarette smoking and measured serum cotinine levels among persons aged 17 years or older: Third National Health and Nutrition Examination Survey, 1988-1994.* Am J Epidemiol, 2001. **153**(8): p. 807-14.
333. Klebanoff M.A., Levine R.J., Morris C.D., Hauth J.C., Sibai B.M., Ben Curet L., et al., *Accuracy of self-reported cigarette smoking among pregnant women in the 1990s.* Paediatr Perinat Epidemiol, 2001. **15**(2): p. 140-3.
334. Mattsson K., Kallen K., Rignell-Hydbom A., Lindh C.H., Jonsson B.A., Gustafsson P., et al., *Cotinine Validation of Self-Reported Smoking During Pregnancy in the Swedish Medical Birth Register.* Nicotine Tob Res, 2016. **18**(1): p. 79-83.

Appendices

Appendix A – Supplementary material to Chapter 3

A.1 - Calculation of the pair sexual isolation index

The pair sexual isolation index (I_{PSI}) [210, 211] is calculated as

$$I_{PSI} = \frac{(PSI_{aa} - PSI_{ab} - PSI_{ba} + PSI_{bb})}{(PSI_{aa} + PSI_{ab} + PSI_{ba} + PSI_{bb})},$$

where

$$PSI_{aa} = \frac{(aa)t}{(aa+ab)(aa+ba)},$$

$$PSI_{ab} = \frac{(ab)t}{(aa+ab)(ab+bb)},$$

$$PSI_{ba} = \frac{(ba)t}{(aa+ba)(ba+bb)},$$

$$PSI_{bb} = \frac{(bb)t}{(ba+bb)(ab+bb)},$$

and the frequency values aa, ab, ba and bb are taken from the cells in Table A.1-1. The frequencies and derivations used to calculate the I_{PSI} for each level of assortative mating in the simulation study are presented in Table A.1-2.

Table A.1-1: Frequency values used to calculate the I_{PSI} statistic.

		Maternal value		
		Non-smoker	Smoker	Total
Paternal value	Non-smoker	aa	ab	aa + ab
	Smoker	ba	bb	ba + bb
	Total	aa + ba	ab + bb	t

Table A.1-2: Values of each stage of the I_{PSI} statistic calculation for each level of assortative mating.

aa	ab	ba	bb	t	aa + ab	ba + bb	aa + ba	ab + bb	PSI_{aa}	PSI_{ab}	PSI_{ba}	PSI_{bb}	I_{PSI} numerator	I_{PSI} denominator	I_{PSI}
38	12	38	12	100	50	50	76	24	1	1	1	1	0	4	0
45.6	9.6	30.4	14.4	100	55.2	44.8	76	24	1.086957	0.724638	0.892857	1.339286	0.808747412	4.04373706	0.2
53.2	7.2	22.8	16.8	100	60.4	39.6	76	24	1.15894	0.496689	0.757576	1.767677	1.672352666	4.180881664	0.4
60.8	4.8	15.2	19.2	100	65.6	34.4	76	24	1.219512	0.304878	0.581395	2.325581	2.658820193	4.431366988	0.6
68.4	2.4	7.6	21.6	100	70.8	29.2	76	24	1.271186	0.141243	0.342466	3.082192	3.86966953	4.837086913	0.8

A.2 – Repetition of the simulation study using a binary outcome

A.2.1 – Methods

Exposure information was derived in the same way as for the continuous outcome. A binary outcome, Y, was then created with prevalence close to 10%. The outcome was designed to have an association with the maternal smoking value, but not the paternal smoking value. Designed maternal smoking odds ratio (DMOR) values between 0.5 and 3 were tested in 0.25 increments.

For each observation, i,

$$\alpha_i = \frac{exp(\beta_0 + \beta_{m\,true} M_i)}{exp(\beta_0 + \beta_{m\,true} M_i) + 1},$$

was derived such that $\beta_{m\,true}$ was equal to the log of the DMOR value and β_0 was a constant coefficient equal to -log(0.9/0.1). A random uniform variable, τ, between 0 and 1 was drawn for each observation. If $\tau_i < \alpha_i$ then the observation was defined as having the outcome (i.e. $Y_i = 1$) otherwise the observation did not have the outcome (i.e. $Y_i = 0$).

The same models were produced as for the continuous outcome but using logistic regression instead. Sample sizes of 1 000, and 10 000 were used as samples of 100 led to model convergence issues. The same performance statistics as for the continuous outcome were used with one exception. For the linear models the mean difference in maternal and paternal coefficients over simulations was taken. Instead, for the logistic models the ratio of the maternal OR to the paternal OR for exposure to smoking in pregnancy is presented. This statistic was produced by taking the mean of the difference in maternal and paternal coefficients over simulations on the log scale, as these were normally distributed, and then exponentiated to provide the ratio of ORs. As the paternal OR is equal to 1 the true value of the ratio of ORs will always be equal to the DMOR. 95% confidence intervals were also created using bootstrapping of the difference on the log scale. The bounds of the confidence interval were averaged over simulations and then exponentiated.

A.2.2 – Results

The bias of coefficient estimates against the quantity of assortative mating for logistic regression models are displayed in Figure A.2-1. As for the linear models the maternal coefficient is unbiased in both the maternal only model and the mutually adjusted model for all quantities of assortative mating (see part (i) of the figure). This is true for DMOR values that had positive and negative associations with the outcome. There is no bias for the paternal coefficient in the mutually adjusted model but there is increasing absolute bias for the paternal only model with increasing assortative mating (see part (ii) of the figure).

Figure A.2-2 shows the mean ratio of the ORs against the quantity of assortative mating for different sample and effect sizes. The results presented here closely resemble those of the linear models. The ratio of ORs for the maternal only and paternal only models tend towards the null (i.e. 1, indicating no difference between the two ORs) as assortative mating increases. The point estimate of the ratio of ORs obtained from the mutually adjusted model is not influenced by assortative mating, however, the confidence interval for the difference increases as a result the maternal and paternal exposure variables reflecting more similar information. This change to the width of the confidence interval is more noticeable at smaller sample sizes than larger sample sizes.

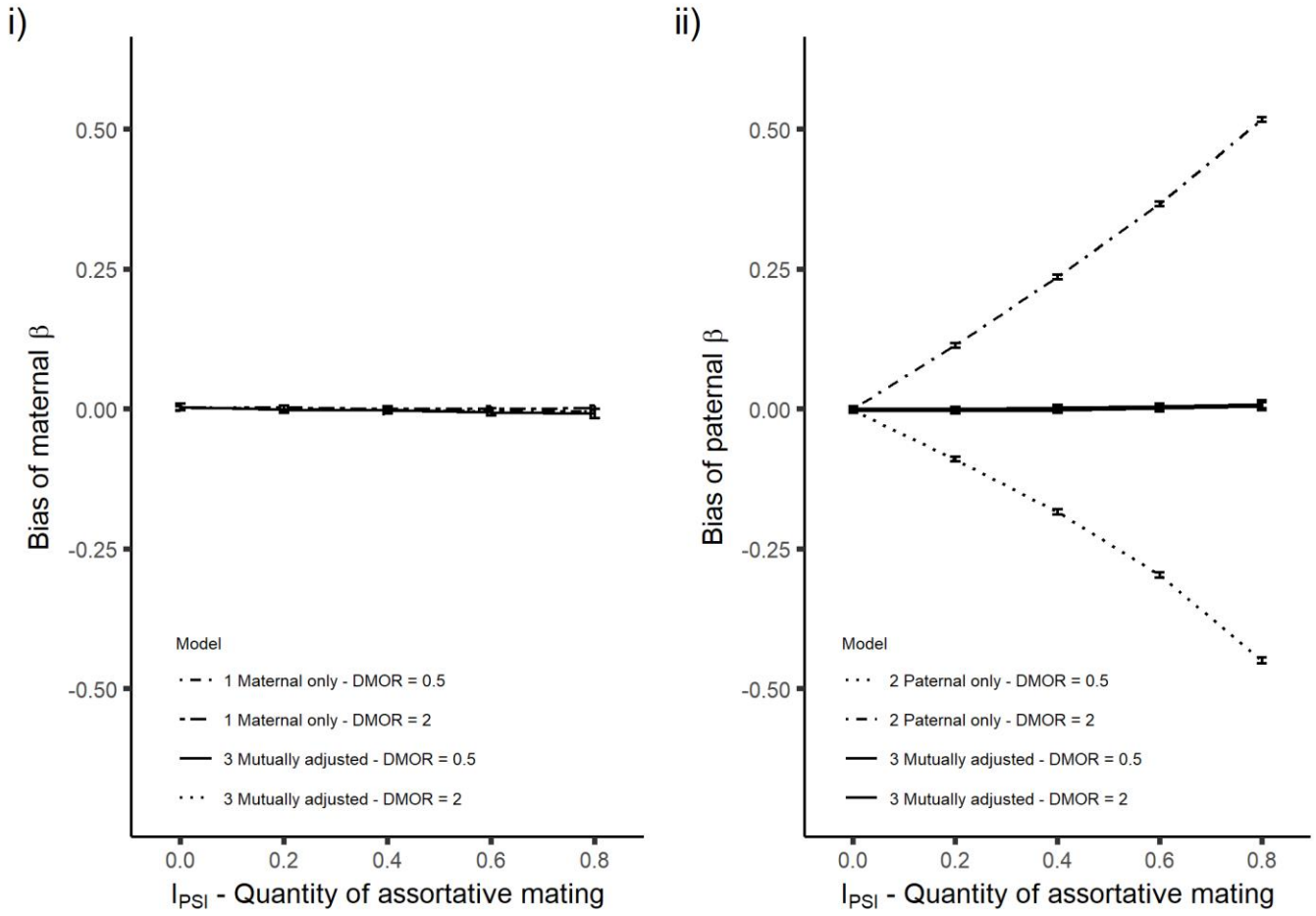

Figure A.2-1: Plots of bias against percentage assortative behaviour for binary outcome data for a) the maternal coefficient and b) the paternal coefficient. Error bars are 95% Monte Carlo confidence intervals across simulations. Sample size for data shown is 10,000. Note the large difference in Y-axis scale between the two plots.

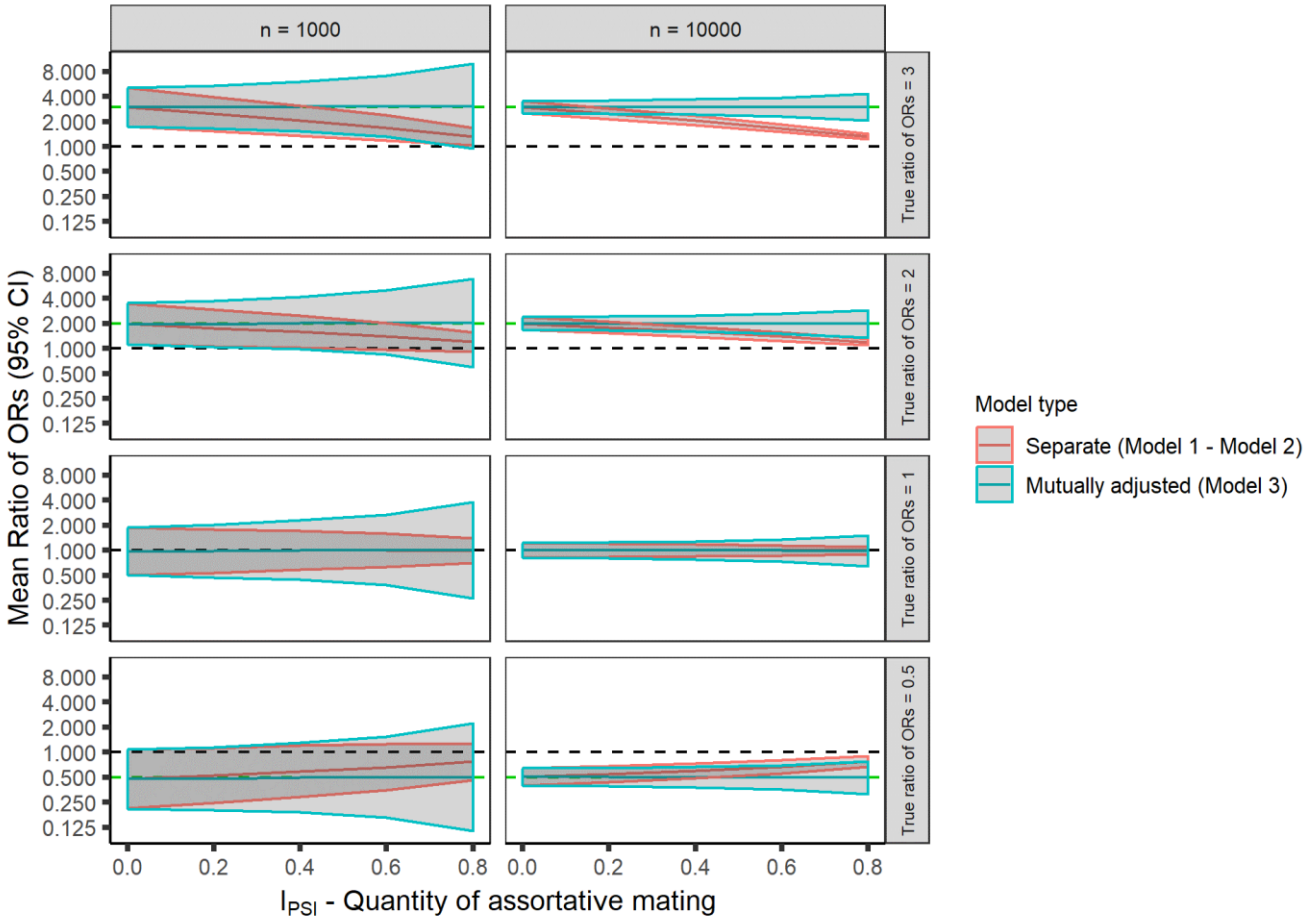

Figure A.2-2: Plot of the mean ratio of ORs across simulations of maternal and paternal β coefficients against the percentage of assortative mating for binary outcome data. 95% confidence bands are the mean lower and upper CI for the difference, produced using bootstrapping. Presented here is the difference between the coefficients of the maternal and paternal only models (red band) and the mutually adjusted model (blue band) for sample sizes of 1000 and 10 000.

A.3 – Supplementary figure

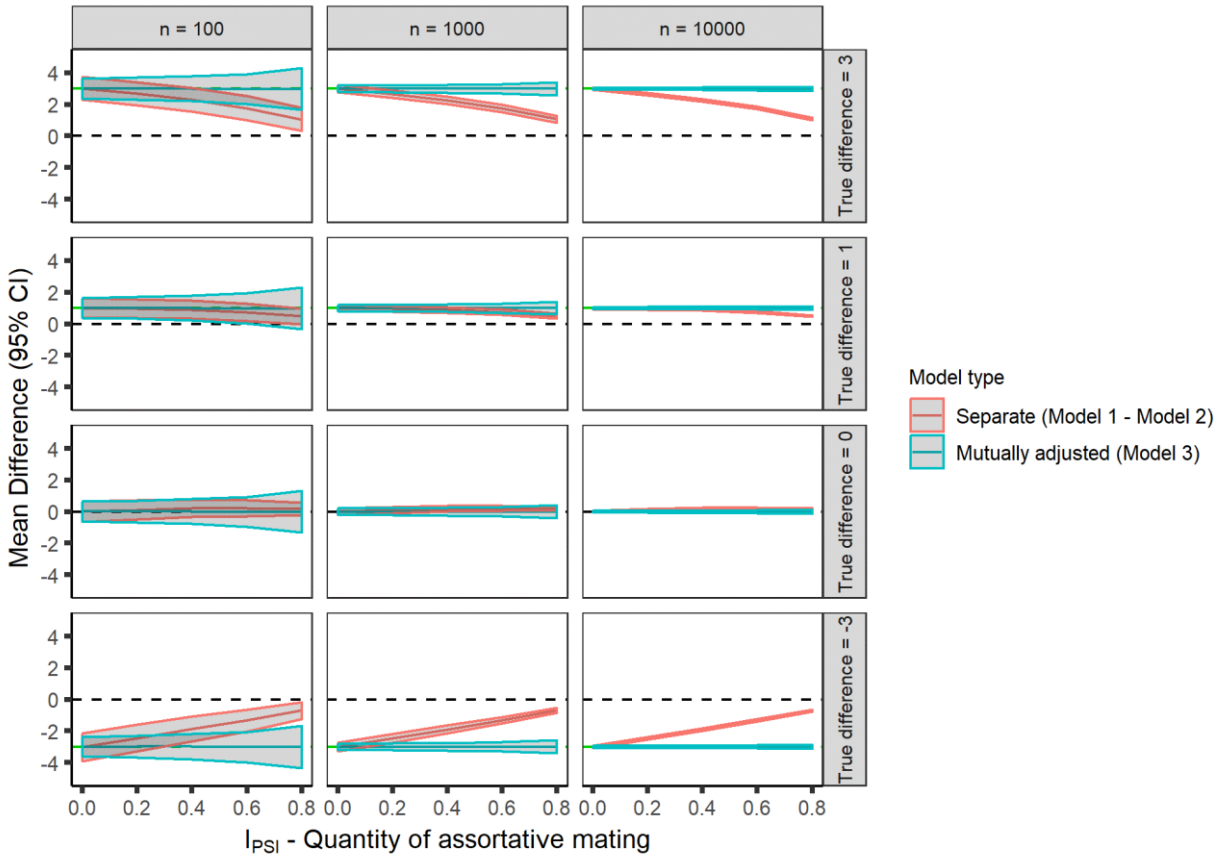

Figure A.3-1: Plot of the mean difference across simulations of maternal and paternal β coefficients against the percentage of assortative mating for data in which maternal and paternal smoking have independent effects. 95% confidence bands are the mean lower and upper CI for the difference, produced using bootstrapping. Presented here is the difference between the coefficients of the maternal and paternal only models (red band) and the mutually adjusted model (blue band) for sample sizes of 100, 1 000 and 10 000.

A.4 – Simulation study of a negative control design with assortative mating and error in the negative exposure

A.4.1 - Methods

Measurement error in the exposure or negative exposure has previously been shown to lead to biased effect estimates [164]. I investigated error in the negative exposure in the context of assortative mating. Frequently in a cohort study maternal report of paternal behaviours are used as a proxy for paternal self-report. Where this occurs, there will likely be an increase in the error for the negative exposure but not the exposure.

I repeated part 1 of the simulation study for continuous data only with the addition of error to the negative exposure/paternal smoking value. No error was added to the exposure/maternal smoking value. Again, I did not repeat the simulation study for the binary exposure due to the potential for non-collapsibility to influence the findings. Three types of error were considered: 1) Random error in the paternal smoking value, 2) over-reporting of assortative mating (where mothers report that their partner has the same exposure behaviour more often than is true) and 3) over-reporting of assortative mating among maternal smokers only. The quantities of error considered were 5%, 10% and 20%.

A.4.2 – Results

Figure E1 shows the bias in the paternal coefficient against percentage assortative mating for each error type and error quantity. None of the error types introduces influenced the mutually adjusted models. Randomly adding error to the paternal exposure (error type 1) reduced bias in the paternal only model at larger quantities of assortative mating compared to data without error. The more error the more bias was reduced. This is likely due to the error reducing the correlation between maternal and paternal smoking value.

Models with error type 2 or 3 showed bias that was greater than the models without error. Bias was observed even when there was no assortative mating for models with these types of error. Bias became closer in size to the models without error as the percentage of assortative mating increased.

A.4.3 – Discussion

The results showed that error in the negative exposure value, such as when maternal report of a paternal behaviour, can lead to bias in the coefficient estimate for the NCA by making the

negative exposure appear to be more or less like the exposure. Mutual adjustment appears to reduce bias occurring in such a way.

By considering error only in the negative exposure I have assumed that maternal report of maternal behaviour is measured with less error than maternal report of paternal behaviour. Prior work on the error rate of maternal report of maternal smoking behaviour has suggested that non-smokers are unlikely to falsely report active smoking [332] while 5% of those reporting that they are non-smokers have been suggested to be current smokers [333, 334]. In contrast, assessment of the agreement between self and maternal reports of paternal smoking in ALSPAC has shown a 5% discrepancy [301]. This does not suggest that the error rate of maternal self-report and maternal report of paternal smoking is equivalent as the 5% error rate for maternal self-report is among non-smokers only. It is also unclear whether those who self-report with error are the same individuals who would report with error for their partners.

Assortment of behaviours may influence the error in the negative exposure. Partners with similar interest in scientific research may be more likely to both engage with cohort studies. Dissimilar interest may lead to one parent engaging with a study while the other does not. Similarity of interest in scientific research may in turn be associated with similarity in behaviours which would be used as negative exposures. As a result, exposure discordant couples may be more likely to rely on maternal report of paternal behaviours. The above patterns of behaviour are speculative and further research is required to identify what patterns of error truly occur for maternal and paternal reports of smoking behaviour and what the determinants of these are. I have, however, provided some insight as to how error could influence conclusions from the negative control design in the presence of assortative mating.

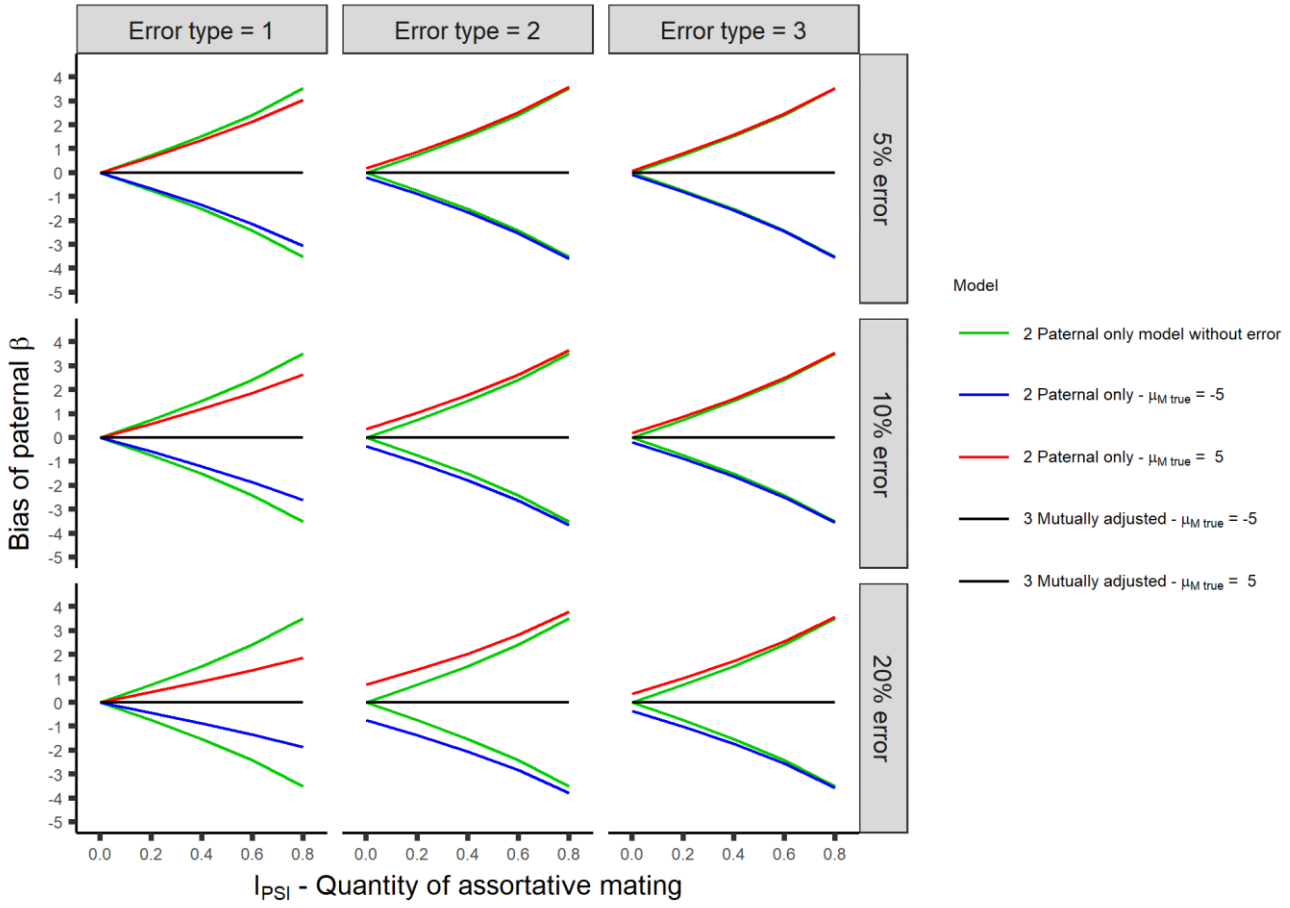

Figure A.4-1: Plot of the bias in the paternal β value against the quantity of assortative mating under three different error structures. Error type 1 is random error to the paternal smoking value, error type 2 is over-reporting of assortative mating (where mothers report that their partner has the same exposure behaviour more often than is true) and error type 3 is over-reporting of assortative mating among maternal smokers only.

Appendix B – Supplementary material to Chapter 4

B.1 – The fraction of missing information (FMI)

For an estimate $\hat{\beta}$, which could be a regression coefficient that has been averaged across m imputed datasets (i.e. $\hat{\beta} = (1/m)\sum_{k=1}^{m}\hat{\beta}_k$) the FMI is given by

$$FMI = \frac{\boldsymbol{B}}{(\boldsymbol{W} + \boldsymbol{B})}.$$

Here $\boldsymbol{W}$ is the within imputation variance of $\hat{\beta}$ and $\boldsymbol{B}$ is the between imputation variance of $\hat{\beta}$. These values are derived as follows. For $\hat{V}_k$, the squared standard error of $\hat{\beta}$ in the k^{th} imputed dataset, the within imputation variance across m imputations is given by

$$\boldsymbol{W} = (1/m)\sum_{k=1}^{m}\hat{V}_k.$$

In other words, $\boldsymbol{W}$ is calculated as the average squared standard error of $\hat{\beta}$. The between imputation variance is given by

$$\boldsymbol{B} = \left(\frac{1}{m-1}\right)\sum_{k=1}^{m}(\hat{\beta}_k - \hat{\beta})^2.$$

This is the square of the standard deviation of the estimated regression coefficient in each imputed dataset relative to the mean value across all imputed datasets.

The total variance, $var(\hat{\beta})$ is equal to $\boldsymbol{W} + \left(1 + \frac{1}{m}\right)\boldsymbol{B}$. Hence for large numbers of imputations m, the FMI is the fraction of the total variance that is attributable to between imputation variance.

Being a fraction of the total variance, the FMI can take values between 0 and 1. Low values (i.e. close to 0) indicate that much of the "missing" information is in fact captured by other, more completely observed, variables. In the absence of any missing data the FMI is 0 because the between imputation variance, $\boldsymbol{B}$, would be 0 since all imputation datasets would be identical. Of course, if we have no missing data we would not use multiple imputation and therefore would not estimate the FMI.

B.2 – Justification for the number of imputations used

Figure S1 displays the FMI of the $\hat{\beta}_1$ coefficient for 20%, 40%, 60% and 80% missing data for increasing numbers of imputations. 47 simulations, all displayed, were performed. An imputation model including Y, X, Z_1 and Z_3 was used along with the same analysis model used (Y regressed on X) as the main investigation. The datasets used were also the same as those used in the main investigation. Multiple imputation was performed using the Stata package *mi impute.* As Stata allows a maximum of 1000 imputations only the FMI had to be calculated manually instead of being obtained directly from Stata's *mi estimate*. The analysis model was performed on each imputed dataset using Stata's *regress* command. Following this the $\hat{\beta}_1$ coefficient for each imputed dataset was appended together and the FMI based on the first m imputations was calculated. FMI was calculated at every 5 imputations from 5-50 imputations, every 25 imputations from 25 to 500 imputations and every 100 imputations from 500 to 10000 imputations.

The FMI is highly variable at low numbers of imputations, only becoming stable for all proportions of missing data when 250 imputations are used. At 1000 imputations we therefore have greater confidence that the FMI estimates across simulations have become stable. The figure also seems to show that the FMI is more variable with greater proportions of missing data at lower numbers of imputations.

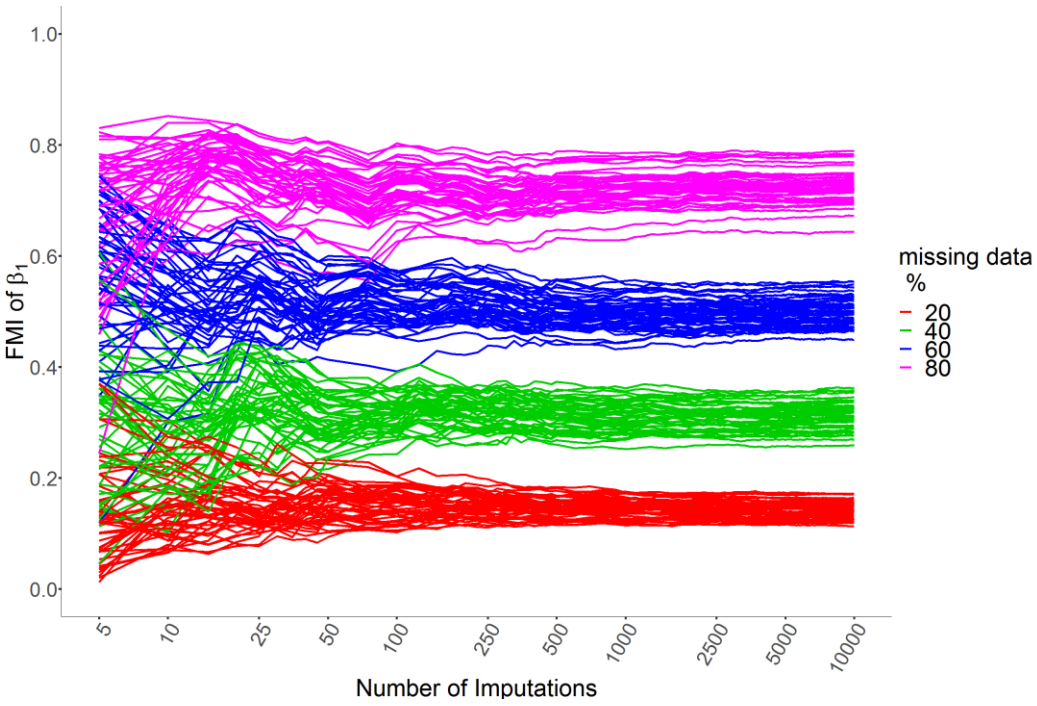

*Figure S1. Line graph of FMI for regression coefficient of **Y** on **X** against the number of imputations.*

B.3 – Calculation of performance statistics for the simulation study

i. Bias

Bias is calculated as the difference between the estimated and true values of β_j,

$$Bias = \hat{\beta}_j - \beta_j,$$

where the true values are given by

$$\beta_1 = \frac{Cov(Y,X)}{Var(X)},$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Here $\bar{y}$ and $\bar{x}$ are the arithmetic means of the variables *Y* (outcome) and *X* (exposure) respectively.

ii. Empirical standard error

For simulations $q = (1, \ldots, r)$, $\hat{\beta}_q$ equal to the estimate averaged across imputations in simulation q, and $\bar{\beta} = (1/r)\sum_{q=1}^{r} \hat{\beta}_q$, the empirical standard deviation is given by

$$\text{emp. s.e.} = \sqrt{\left(\frac{1}{r-1}\right)\sum_{q}^{r}\left(\hat{\beta}_q - \bar{\beta}\right)^2}.$$

Note the difference between empirical standard deviation and between imputation variance is that the former is calculated across simulations while the latter is calculated across imputations within a single simulation.

iii. Average estimated standard errors

For s_q, the standard error of $\hat{\beta}$ in the q^{th} simulation, the average estimated standard error is given by

$$model\ s.e. = \sqrt{\overline{s^2}} = \sqrt{\frac{1}{r}\sum\nolimits_{q=1}^{r} s_q^2}$$

We do not present the average standard errors. These are used to calculate the relative percentage error.

iv. Relative percentage error

The relative percentage error is the proportional error of the average estimated standard errors $\sqrt{\overline{s^2}}$, relative to the empirical standard deviation of $\hat{\beta}_j$.

$$\text{Relative \% Error} = \left(\frac{model\ s.e.}{emp.s.e.} - 1\right) \times 100$$

The empirical standard error and the average estimated standard error should be approximately equal. The relative percentage error is therefore a measure of bias in s_j with departures from 0 indicating bias. Relative percentage error was calculated using the *simsum* command in Stata.

v. Coverage probability of 95% CI

Letting $I(.)$ be the indicator function, the coverage of a 95% confidence interval is given by

$$C = \frac{1}{r}\sum_{q=1}^{r} I(\left|\hat{\beta}_q - \beta\right| < z_{\alpha/2}s_q),$$

where $z_{\alpha/2} = 1.96$ is the critical value of the normal distribution. The coverage probability indicates the percentage of simulations in which the true value of the coefficient is found within the span of the estimated 95% confidence interval. A value greater than 95% indicates over-coverage while a value less than 95% indicates under-coverage.

B.4 – Empirical SE of the MI exposure coefficient against FMI - Figure 1 separated by panels of percentage missing data

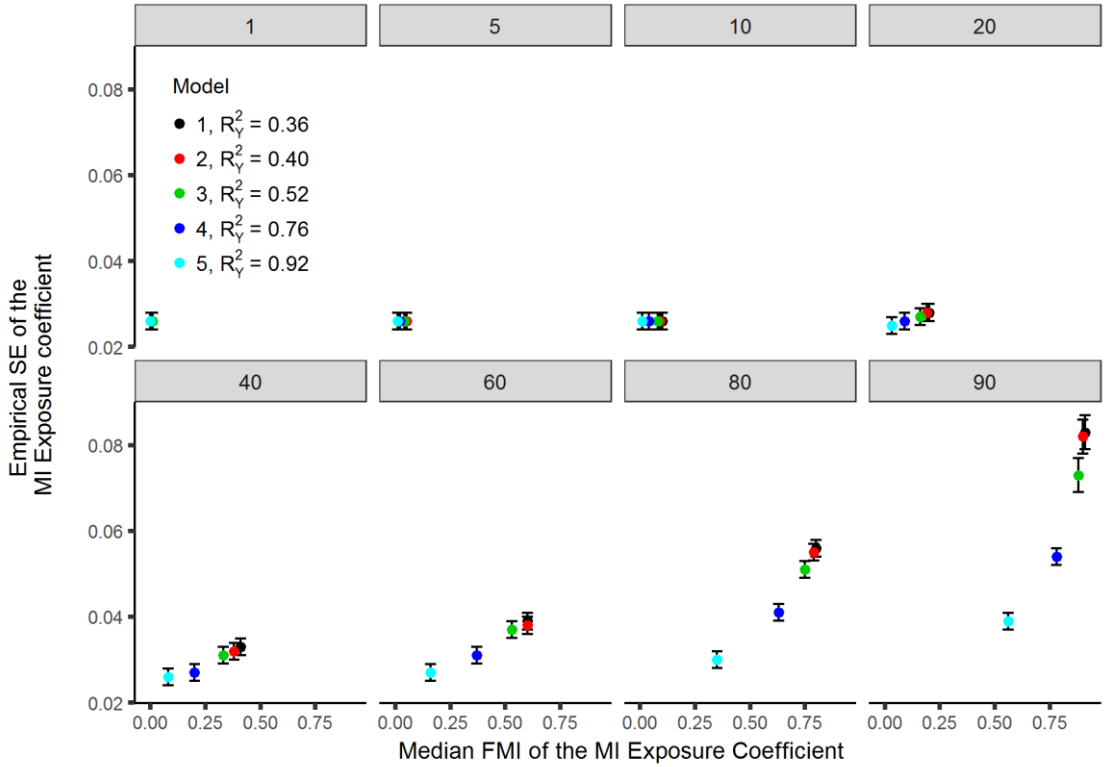

Figure B.4-1: Empirical SE of the MI exposure coefficient plotted against FMI for simulated MCAR data. Panels separate the data by percentage of missing data. Error bars are 95% confidence intervals based on Monte Carlo standard errors across simulations. FMI = fraction of missing information; SE = standard error.

B.5 – Performance statistics for the exposure coefficient in the simulation study

Table B.5-1: Performance Statistics for the MCAR Results of the Exposure Coefficient in the Simulation Study.

Statistic	% missing data	CCA		Model 1 No aux info, R^2=0.36		Model 2 R^2=0.40		Model 3 R^2=0.52		Model 4 R^2=0.76		Model 5 R^2=0.92	
FMI [a]	1			0.01	0.01, 0.01	0.01	0.01, 0.01	0.01	0.00, 0.01	0.00	0.00, 0.00	0.00	0.00, 0.00
	5			0.05	0.04, 0.06	0.05	0.04, 0.05	0.04	0.03, 0.04	0.02	0.02, 0.02	0.01	0.01, 0.01
	10			0.10	0.09, 0.11	0.09	0.08, 0.10	0.08	0.07, 0.08	0.04	0.04, 0.05	0.01	0.01, 0.02
	20			0.20	0.19, 0.21	0.19	0.18, 0.20	0.16	0.15, 0.17	0.09	0.08, 0.09	0.03	0.03, 0.03
	40			0.41	0.39, 0.42	0.38	0.37, 0.40	0.33	0.32, 0.35	0.20	0.19, 0.21	0.08	0.07, 0.08
	60			0.60	0.59, 0.62	0.60	0.58, 0.61	0.53	0.52, 0.55	0.37	0.35, 0.39	0.16	0.15, 0.17
	80			0.80	0.79, 0.81	0.79	0.78, 0.81	0.75	0.73, 0.76	0.63	0.60, 0.65	0.35	0.33, 0.38
	90			0.91	0.90, 0.91	0.90	0.89, 0.91	0.88	0.86, 0.89	0.78	0.76, 0.80	0.56	0.53, 0.60
Bias	1	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	5	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	10	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	20	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	40	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	60	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	80	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	-0.01, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	90	0.00	0.00, 0.01	0.00	0.00, 0.01	0.00	0.00, 0.01	0.00	0.00, 0.01	0.00	0.00, 0.00	0.00	0.00, 0.00
Empirical SE	1	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03
	5	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03
	10	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03
	20	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03
	40	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.02, 0.03
	60	0.04	0.04, 0.04	0.04	0.04, 0.04	0.04	0.04, 0.04	0.04	0.04, 0.04	0.03	0.03, 0.03	0.03	0.03, 0.03
	80	0.06	0.05, 0.06	0.06	0.05, 0.06	0.05	0.05, 0.06	0.05	0.05, 0.05	0.04	0.04, 0.04	0.03	0.03, 0.03
	90	0.08	0.08, 0.09	0.08	0.08, 0.09	0.08	0.08, 0.09	0.07	0.07, 0.08	0.05	0.05, 0.06	0.04	0.04, 0.04
Relative error	1	-1.34	-5.67, 3.00	-1.36	-5.69, 2.97	-1.19	-5.52, 3.15	-1.26	-5.59, 3.07	-1.10	-5.44, 3.24	-1.24	-5.58, 3.09
	5	-0.53	-4.90, 3.84	-0.52	-4.89, 3.85	-0.49	-4.86, 3.87	-0.10	-4.48, 4.29	-0.08	-4.46, 4.31	-1.14	-5.48, 3.19
	10	0.96	-3.47, 5.39	0.89	-3.54, 5.32	1.03	-3.41, 5.46	0.30	-4.10, 4.70	0.44	-3.97, 4.85	-0.62	-4.98, 3.74
	20	0.92	-3.51, 5.35	0.77	-3.65, 5.20	1.47	-2.98, 5.93	1.03	-3.41, 5.47	2.93	-1.58, 7.45	2.47	-2.02, 6.97
	40	0.63	-3.79, 5.05	1.09	-3.35, 5.53	1.46	-2.99, 5.92	1.04	-3.40, 5.48	3.25	-1.28, 7.78	2.94	-1.58, 7.46

Statistic	% missing data	CCA		Model 1 No aux info, R^2=0.36		Model 2 R^2=0.40		Model 3 R^2=0.52		Model 4 R^2=0.76		Model 5 R^2=0.92	
	60	2.37	-2.13, 6.87	2.81	-1.71, 7.33	4.80	0.19, 9.40	0.42	-3.99, 4.84	3.28	-1.25, 7.82	3.47	-1.07, 8.02
	80	1.80	-2.69, 6.29	1.11	-3.35, 5.57	2.56	-1.96, 7.08	-1.52	-5.85, 2.82	3.14	-1.40, 7.68	4.61	0.01, 9.20
	90	-2.06	-6.40, 2.28	-0.47	-4.88, 3.93	-0.71	-5.11, 3.68	-0.10	-4.52, 4.32	1.21	-3.26, 5.69	0.38	-4.05, 4.80
Coverage	1	94.70	93.31, 96.09	94.60	93.20, 96.00	94.60	93.20, 96.00	94.90	93.54, 96.26	94.70	93.31, 96.09	94.90	93.54, 96.26
	5	95.20	93.88, 96.52	95.30	93.99, 96.61	95.20	93.88, 96.52	95.40	94.10, 96.70	95.00	93.65, 96.35	94.70	93.31, 96.09
	10	95.50	94.22, 96.78	95.40	94.10, 96.70	95.50	94.22, 96.78	95.10	93.76, 96.44	95.30	93.99, 96.61	94.90	93.54, 96.26
	20	94.70	93.31, 96.09	94.80	93.42, 96.18	94.80	93.42, 96.18	94.40	92.97, 95.83	95.50	94.22, 96.78	95.50	94.22, 96.78
	40	94.90	93.54, 96.26	94.90	93.54, 96.26	94.90	93.54, 96.26	94.40	92.97, 95.83	95.20	93.88, 96.52	95.80	94.56, 97.04
	60	96.20	95.01, 97.39	96.20	95.01, 97.39	96.30	95.13, 97.47	95.00	93.65, 96.35	95.50	94.22, 96.78	95.30	93.99, 96.61
	80	95.00	93.65, 96.35	94.70	93.31, 96.09	95.70	94.44, 96.96	93.90	92.42, 95.38	95.50	94.22, 96.78	95.90	94.67, 97.13
	90	94.90	93.54, 96.26	95.30	93.99, 96.61	94.90	93.54, 96.26	94.30	92.86, 95.74	95.30	93.99, 96.61	95.10	93.76, 96.44

CCA – Complete case analysis; FMI- Fraction of Missing Information; SE – Standard error; R^2 – the squared coefficient of multiple correlation

[a] For FMI the estimate was the median across simulations and the interval represents the interquartile range. For all other statistics the mean across simulations was taken and the 95% confidence interval was calculated using Monte Carlo standard error.

Table B.5-2: Performance Statistics for the MAR Results of the Exposure Coefficient in the Simulation Study.

Statistic	% missing data	CCA		Model 1 No aux info, R^2=0.36		Model 2 R^2=0.40		Model 3 R^2=0.52		Model 4 R^2=0.76		Model 5 R^2=0.92	
FMI [a]	1			0.02	0.01, 0.02	0.02	0.01, 0.02	0.01	0.01, 0.02	0.01	0.00, 0.01	0.00	0.00, 0.00
	5			0.08	0.07, 0.09	0.08	0.07, 0.09	0.06	0.05, 0.07	0.03	0.03, 0.04	0.01	0.01, 0.01
	10			0.15	0.14, 0.17	0.15	0.13, 0.16	0.12	0.11, 0.14	0.06	0.06, 0.07	0.02	0.02, 0.03
	20			0.27	0.25, 0.29	0.26	0.24, 0.28	0.22	0.21, 0.24	0.13	0.11, 0.14	0.05	0.04, 0.05
	40			0.48	0.45, 0.49	0.46	0.44, 0.48	0.42	0.40, 0.44	0.26	0.25, 0.28	0.11	0.10, 0.12
	60			0.66	0.64, 0.68	0.64	0.62, 0.66	0.60	0.58, 0.62	0.43	0.41, 0.45	0.20	0.19, 0.22
	80			0.83	0.81, 0.84	0.82	0.80, 0.84	0.79	0.77, 0.81	0.65	0.63, 0.68	0.39	0.36, 0.43
	90			0.91	0.90, 0.92	0.91	0.90, 0.92	0.89	0.87, 0.90	0.80	0.78, 0.82	0.59	0.55, 0.63
Bias	1	0.00	-0.01, 0.00	0.00	-0.01, 0.00	0.00	-0.01, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	5	-0.01	-0.02, -0.01	-0.01	-0.02, -0.01	-0.01	-0.02, -0.01	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	10	-0.02	-0.03, -0.02	-0.02	-0.03, -0.02	-0.02	-0.03, -0.02	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	20	-0.04	-0.04, -0.04	-0.04	-0.04, -0.04	-0.04	-0.04, -0.04	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	40	-0.06	-0.06, -0.06	-0.06	-0.06, -0.06	-0.06	-0.06, -0.05	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	60	-0.06	-0.07, -0.06	-0.06	-0.07, -0.06	-0.06	-0.07, -0.06	0.00	0.00, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	80	-0.06	-0.07, -0.06	-0.06	-0.07, -0.06	-0.06	-0.07, -0.06	0.00	-0.01, 0.00	0.00	0.00, 0.00	0.00	0.00, 0.00
	90	-0.05	-0.06, -0.05	-0.05	-0.06, -0.05	-0.05	-0.06, -0.05	0.00	0.00, 0.00	0.00	0.00, 0.01	0.00	0.00, 0.00
Empirical SE	1	0.02	0.02, 0.03	0.02	0.02, 0.03	0.02	0.02, 0.03	0.02	0.02, 0.03	0.02	0.02, 0.03	0.02	0.02, 0.03
	5	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03
	10	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03
	20	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03
	40	0.03	0.03, 0.04	0.03	0.03, 0.04	0.03	0.03, 0.04	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03
	60	0.04	0.04, 0.04	0.04	0.04, 0.04	0.04	0.04, 0.04	0.04	0.04, 0.04	0.03	0.03, 0.04	0.03	0.03, 0.03
	80	0.06	0.06, 0.06	0.06	0.06, 0.06	0.06	0.06, 0.06	0.05	0.05, 0.06	0.04	0.04, 0.04	0.03	0.03, 0.03
	90	0.08	0.08, 0.09	0.08	0.08, 0.09	0.08	0.08, 0.08	0.08	0.07, 0.08	0.05	0.05, 0.06	0.04	0.04, 0.04
Relative error	1	2.45	-2.04, 6.95	2.42	-2.08, 6.91	2.65	-1.86, 7.15	2.45	-2.05, 6.95	2.39	-2.10, 6.89	2.33	-2.16, 6.82
	5	1.42	-3.03, 5.88	1.39	-3.06, 5.84	1.18	-3.26, 5.62	1.83	-2.64, 6.30	1.96	-2.52, 6.43	1.89	-2.59, 6.36
	10	1.20	-3.24, 5.65	1.22	-3.22, 5.67	1.64	-2.82, 6.11	1.43	-3.02, 5.89	1.77	-2.70, 6.23	1.79	-2.67, 6.26
	20	3.94	-0.63, 8.50	4.00	-0.57, 8.56	4.30	-0.28, 8.88	5.63	0.99, 10.27	4.47	-0.12, 9.05	2.64	-1.87, 7.14
	40	-1.16	-5.51, 3.18	-1.17	-5.51, 3.17	-1.38	-5.71, 2.95	0.51	-3.90, 4.93	-1.05	-5.39, 3.29	0.15	-4.25, 4.55
	60	2.40	-2.10, 6.91	2.65	-1.86, 7.17	2.22	-2.27, 6.72	3.91	-0.66, 8.48	-0.32	-4.70, 4.05	1.01	-3.43, 5.44
	80	2.11	-2.40, 6.61	2.46	-2.06, 6.98	1.73	-2.76, 6.22	4.91	0.28, 9.53	2.98	-1.55, 7.52	0.94	-3.50, 5.38
	90	1.38	-3.13, 5.88	2.73	-1.83, 7.30	2.03	-2.50, 6.56	2.07	-2.47, 6.60	6.09	1.39, 10.79	4.30	-0.30, 8.91

Statistic	% missing data	CCA		Model 1 No aux info, R^2=0.36		Model 2 R^2=0.40		Model 3 R^2=0.52		Model 4 R^2=0.76		Model 5 R^2=0.92	
Coverage	1	95.10	93.76, 96.44	95.10	93.76, 96.44	95.00	93.65, 96.35	95.00	93.65, 96.35	95.10	93.76, 96.44	95.30	93.99, 96.61
	5	91.40	89.66, 93.14	91.20	89.44, 92.96	91.50	89.77, 93.23	95.00	93.65, 96.35	95.00	93.65, 96.35	95.60	94.33, 96.87
	10	85.80	83.64, 87.96	86.00	83.85, 88.15	86.00	83.85, 88.15	95.70	94.44, 96.96	95.10	93.76, 96.44	95.90	94.67, 97.13
	20	73.30	70.56, 76.04	72.90	70.15, 75.65	72.60	69.84, 75.36	97.00	95.94, 98.06	96.10	94.90, 97.30	96.00	94.79, 97.21
	40	61.80	58.79, 64.81	61.40	58.38, 64.42	60.90	57.88, 63.92	94.30	92.86, 95.74	94.30	92.86, 95.74	95.20	93.88, 96.52
	60	70.40	67.57, 73.23	71.00	68.19, 73.81	67.70	64.80, 70.60	96.00	94.79, 97.21	94.30	92.86, 95.74	95.40	94.10, 96.70
	80	82.10	79.72, 84.48	82.20	79.83, 84.57	82.50	80.14, 84.86	96.60	95.48, 97.72	94.80	93.42, 96.18	94.90	93.54, 96.26
	90	90.20	88.36, 92.04	90.40	88.57, 92.23	89.40	87.49, 91.31	96.00	94.79, 97.21	96.40	95.25, 97.55	96.10	94.90, 97.30

CCA – Complete case analysis; FMI- Fraction of Missing Information; SE – Standard error; R^2 – the squared coefficient of multiple correlation

[a] For FMI the estimate was the median across simulations and the interval represents the interquartile range. For all other statistics the mean across simulations was taken and the 95% confidence interval was calculated using Monte Carlo standard error.

B.6 – FMI and efficiency gains were not sensitive to whether the auxiliary variable was included in the missingness mechanism

In imputation model 3 the auxiliary variable Z_1 can be replaced by Z_2 and still provide the same R_Y^2 value. However, the variable Z_1 is associated with the missing data mechanism while Z_2 is not. These two models can be used to check whether parameters are sensitive to the auxiliary variable being associated with the missing data mechanism. Table S6 shows that FMI and empirical SE do not differ between the two models. However, because Z1 is associated with missingness, model 3 (using Z_2 in place of Z_1) is biased where the original model 3 was not.

Table B.6-1. Performance Statistics for the MAR Results of the Exposure Coefficient of Model 3 replacing variable Z_1 with Z_2.

Statistic	% missing data	CCA		Model 3 R^2=0.52		Model 3 using Z2 R^2=0.52	
FMI [a]	1			0.01	0.01, 0.02	0.01	0.01, 0.02
	5			0.06	0.05, 0.07	0.06	0.05, 0.07
	10			0.12	0.11, 0.14	0.12	0.11, 0.13
	20			0.22	0.21, 0.24	0.22	0.20, 0.23
	40			0.42	0.40, 0.44	0.40	0.38, 0.42
	60			0.60	0.58, 0.62	0.59	0.57, 0.61
	80			0.79	0.77, 0.81	0.78	0.76, 0.80
	90			0.89	0.87, 0.90	0.89	0.87, 0.90
Bias	1	0.00	-0.01, 0.00	0.00	0.00, 0.00	0.00	-0.01, 0.00
	5	-0.01	-0.02, -0.01	0.00	0.00, 0.00	-0.01	-0.02, -0.01
	10	-0.02	-0.03, -0.02	0.00	0.00, 0.00	-0.02	-0.03, -0.02
	20	-0.04	-0.04, -0.04	0.00	0.00, 0.00	-0.04	-0.04, -0.04
	40	-0.06	-0.06, -0.06	0.00	0.00, 0.00	-0.06	-0.06, -0.05
	60	-0.06	-0.07, -0.06	0.00	0.00, 0.00	-0.06	-0.07, -0.06
	80	-0.06	-0.07, -0.06	0.00	-0.01, 0.00	-0.06	-0.07, -0.06
	90	-0.05	-0.06, -0.05	0.00	0.00, 0.00	-0.05	-0.05, -0.05
Empirical SE	1	0.02	0.02, 0.03	0.02	0.02, 0.03	0.02	0.02, 0.03
	5	0.03	0.02, 0.03	0.03	0.02, 0.03	0.03	0.02, 0.03
	10	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03
	20	0.03	0.03, 0.03	0.03	0.03, 0.03	0.03	0.03, 0.03
	40	0.03	0.03, 0.04	0.03	0.03, 0.03	0.03	0.03, 0.03
	60	0.04	0.04, 0.04	0.04	0.04, 0.04	0.04	0.04, 0.04
	80	0.06	0.06, 0.06	0.05	0.05, 0.06	0.05	0.05, 0.06
	90	0.08	0.08, 0.09	0.08	0.07, 0.08	0.07	0.07, 0.07
Relative error	1	2.45	-2.04, 6.95	2.45	-2.05, 6.95	2.15	-2.33, 6.64
	5	1.42	-3.03, 5.88	1.83	-2.64, 6.30	2.22	-2.26, 6.71
	10	1.20	-3.24, 5.65	1.43	-3.02, 5.89	1.60	-2.86, 6.06
	20	3.94	-0.63, 8.50	5.63	0.99, 10.27	3.06	-1.46, 7.59
	40	-1.16	-5.51, 3.18	0.51	-3.90, 4.93	-1.42	-5.75, 2.91
	60	2.40	-2.10, 6.91	3.91	-0.66, 8.48	1.09	-3.36, 5.53
	80	2.11	-2.40, 6.61	4.91	0.28, 9.53	1.32	-3.15, 5.79
	90	1.38	-3.13, 5.88	2.07	-2.47, 6.60	5.84	1.14, 10.54
Coverage	1	95.10	93.76, 96.44	95.00	93.65, 96.35	95.3	93.99, 96.61
	5	91.40	89.66, 93.14	95.00	93.65, 96.35	91.6	89.88, 93.32
	10	85.80	83.64, 87.96	95.70	94.44, 96.96	84.8	82.57, 87.03

Statistic	% missing data	CCA		Model 3 R^2=0.52		Model 3 using Z2 R^2=0.52	
	20	73.30	70.56, 76.04	97.00	95.94, 98.06	71.6	68.81, 74.39
	40	61.80	58.79, 64.81	94.30	92.86, 95.74	59.5	56.46, 62.54
	60	70.40	67.57, 73.23	96.00	94.79, 97.21	64.6	61.64, 67.56
	80	82.10	79.72, 84.48	96.60	95.48, 97.72	80.1	77.63, 82.57
	90	90.20	88.36, 92.04	96.00	94.79, 97.21	90.1	88.25, 91.95

CCA – Complete case analysis; FMI- Fraction of Missing Information; SE – Standard error; R^2 – the squared coefficient of multiple correlation

[a] For FMI the estimate was the median across simulations and the interval represents the interquartile range. For all other statistics the mean across simulations was taken and the 95% confidence interval was calculated using Monte Carlo standard error.

B.7 – Missing data in included and excluded sample for the applied example

Table B.7-1: Quantity of Missing Variables (Outcome, Exposure or Auxiliary) Separated by Inclusion in the Sample.

Number of missing variables (outcome, exposure or auxiliary)	N(%) Missing confounders/ excluded	No missing confounders/ included	Total
0 - No missing	42 (1.51)	1,607 (13.49)	1,649 (11.22)
1	142 (5.11)	1,408 (11.82)	1,550 (10.55)
2	201 (7.23)	1,292 (10.85)	1,493 (10.16)
3	461 (16.58)	1,817 (15.25)	2,278 (15.51)
4	482 (17.34)	1,771 (14.87)	2,253 (15.34)
5	249 (8.96)	1,174 (9.86)	1,423 (9.69)
6	257 (9.24)	1,118 (9.39)	1,375 (9.36)
7	551 (19.82)	1,724 (14.47)	2,275 (15.49)
8 - All missing	395 (14.21)	0 (0.00)	395 (2.69)
Total	2,780 (100.00)	11,911 (100.00)	14,691 (100.00)

B.8 – Missing data pattern in the applied example

Table B.8-1. Missing Data Pattern for the Exposure, Outcome and the Strongest Auxiliary for the Outcome

Pattern	Exposure	Outcome	Strongest auxiliary [a]	N(%)
1	Complete	Missing	Missing	4,803 (40.32)
2	Complete	Complete	Complete	3,974 (33.36)
3	Complete	Missing	Complete	2,698 (22.65)
4	Complete	Complete	Missing	496 (4.16)

[a] The strongest auxiliary for predicting the outcome was IQ at age 8

B.9 – Applied example exposure coefficient results for the unadjusted model

	Exposure			% reduction
Model	**Coefficient**	**SE**	**FMI**	**in SE**
CCA	-2.80	0.520		
A: Only confounders as auxiliary information	-3.10	0.506	0.708	2.80
B: IQ at age 8	-3.45	0.458	0.644	11.85
C: Intelligibility at age 9	-3.07	0.507	0.708	2.55
D: Maths assessment score	-3.47	0.471	0.658	9.52
E: Learning Disabilities	-3.42	0.500	0.695	3.89
F: Streaming for Maths and English	-3.46	0.473	0.661	8.97
G: IQ at age 8 and intelligibility at age 9	-3.44	0.468	0.657	10.03
H: IQ at age 8 and maths assessment score	-3.70	0.456	0.635	12.41
I: IQ, intel. and maths assessment	-3.70	0.458	0.637	12.00
J: IQ, intel., maths assessment and LD	-3.85	0.457	0.632	12.11
K: IQ, intel., maths assessment and streaming groups	-3.80	0.459	0.637	11.66
L: IQ, intel., maths assessment, LD and streaming groups	-3.84	0.460	0.636	11.59

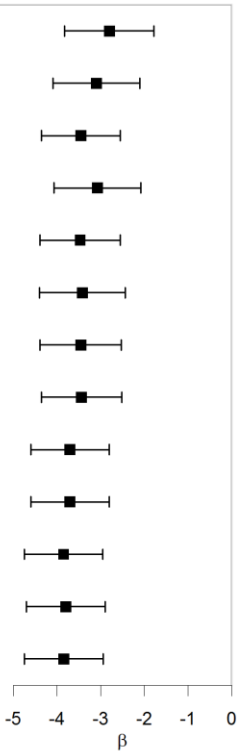

Figure S3. Estimate, standard error and FMI for the exposure coefficient in the applied example unadjusted analysis model. Reduction in SE is relative to CCA. CCA=complete case analysis; FMI= fraction of missing information; SE = standard error.

B.10 – Simulation study using a binary outcome

B.10.1 – Methods

Data model. We first simulated 1000 observations of exposure X and auxiliary variables $Z_1 - Z_3$ from independent standard normal distributions. Y was then simulated as a binary outcome variable using a logistic regression model as the linear combination of X and $Z_1 - Z_3$, each with a coefficient of log(2), and a constant of negative log(0.9455/0.0545).

Missingness was simulated in the same way as for our linear example. MCAR data was simulated by removing the first p observations such that $\frac{p}{n}$ gives the required proportion of missing data. MAR missingness was simulated under a logistic regression model using

$$\text{logit}(\lambda_i) = \alpha + \text{Z}_{1i} + X_i,$$

where α was manipulated for the different simulation settings to provide the required proportion of missing data on average across datasets.

Analysis model. For each simulation setting and imputation model the following logistic regression analysis model was used:

$$\text{logit}(E[Y]) = \beta_0 + \beta_1 x,$$

where β_0 and β_1 are the intercept and exposure coefficient respectively. We took the true value of β_1 to be the mean across all simulations of the coefficient obtained from analysis of the full dataset with no observations removed. We used this as the definition of the true value as opposed to the designed value, as was used in the linear model, in order to account for non-collapsibility of the OR for our exposure coefficient.

Imputation models. Five imputation models were considered for both MCAR and MAR data (see Table B.10-1). All models contained the variables included in the analysis model and used logistic regression to impute the missing outcome. Model 1 contained no auxiliary information. Model 2 and model 3 were identical in terms of quantity of auxiliary information but Model 3 contained all variables in the missingness mechanism while Model 2 did not. Models 4 and 5 contained increasing quantities of auxiliary information, achieved by increasing the number of Z variables included in the imputation model. To measure the quantity of auxiliary information we have used the sum of the designed odds ratios for Y, $\sum OR_Y$. The $\sum OR_Y$ is used in place of R^2_Y from the simulation study with a continuous outcome. For each imputation model 1000 imputations were run.

Table B.10-1. Description of the Imputation Models Used for Both MCAR and MAR Data.

Imputation model	Variables included	$\sum OR_Y$ [a]
1 (least auxiliary information)	*Y*, *X*	2
2	*Y*, *X*, Z_2	4
3	*Y*, *X*, Z_1	4
4	*Y*, *X*, $Z_{1\text{-}2}$	6
5 (most auxiliary information)	*Y*, *X*, $Z_{1\text{-}3}$	8

[a] $\sum OR_Y$ is the sum of the designed odds ratios in the logistic model used to produce Y for variables included in the imputation model.

Comparisons. We repeated the simulation study for 1%, 5%, 10%, 20%, 40% 60%, 80% and 90% missing data. For all scenarios, we generated 1000 independent simulated datasets We used the same performance statistics for the logistic model as we did for the linear model. All performance statistics were calculated on the log(OR) scale. Performance statistics should not be compared between our logistic and linear simulation studies due to differences between the imputation models used.

B.10.2 – Results

We present the results of the simulation study with a binary outcome in Table B.10-2 for MCAR data and Table B.10-3 for MAR data. In the MAR setting bias is reduced, compared to CCA, when all variables in the missingness mechanism are included in the imputation model. However, the quantity of bias is smaller than the size of the empirical SE. Bias can therefore be considered to be negligible in all of our simulation settings. Small improvements in efficiency over CCA were observed with increasing auxiliary information for both MCAR and MAR data.

We display a plot of the empirical SE against the FMI for MCAR data in Figure B.10-1 and MAR data in Figure B.10-2. As in our linear model example, the FMI was reduced with increasing auxiliary information. This reduction was greater for MAR data than for MCAR data.

Table B.10-2. Performance Statistics for the MCAR Results of the Exposure Coefficient in the Binary Outcome Simulation Study.

Statistic	% missing data	CCA		Model 1 No aux info, R^2=0.36		Model 2 R^2=0.40		Model 3 R^2=0.52		Model 4 R^2=0.76		Model 5 R^2=0.92	
FMI [a]	1			0.01	0.01, 0.01	0.01	0.01, 0.01	0.01	0.01, 0.01	0.01	0.01, 0.01	0.01	0.00, 0.01
	5			0.05	0.04, 0.06	0.05	0.04, 0.06	0.05	0.04, 0.05	0.04	0.04, 0.05	0.04	0.03, 0.05
	10			0.10	0.09, 0.11	0.09	0.08, 0.10	0.09	0.08, 0.10	0.09	0.08, 0.10	0.08	0.07, 0.10
	20			0.20	0.19, 0.22	0.19	0.18, 0.21	0.19	0.18, 0.21	0.18	0.17, 0.20	0.17	0.16, 0.19
	40			0.40	0.38, 0.42	0.39	0.37, 0.41	0.38	0.37, 0.41	0.38	0.36, 0.40	0.36	0.34, 0.38
	60			0.60	0.58, 0.61	0.59	0.57, 0.61	0.59	0.57, 0.61	0.58	0.56, 0.60	0.56	0.54, 0.59
	80			0.79	0.78, 0.81	0.78	0.77, 0.80	0.79	0.77, 0.80	0.79	0.77, 0.81	0.77	0.75, 0.79
	90			0.89	0.88, 0.90	0.90	0.88, 0.91	0.89	0.88, 0.91	0.89	0.87, 0.90	0.89	0.87, 0.90
Bias	1	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01
	5	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	0.00, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01
	10	0.00	0.00, 0.01	0.00	0.00, 0.01	0.00	0.00, 0.01	0.00	0.00, 0.01	0.00	0.00, 0.01	0.00	0.00, 0.01
	20	0.00	0.00, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	0.00, 0.01	0.00	-0.01, 0.01	0.00	0.00, 0.01
	40	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.01	0.00, 0.02	0.01	0.00, 0.02
	60	0.00	-0.01, 0.02	0.00	-0.01, 0.02	0.01	0.00, 0.02	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.02
	80	0.02	0.00, 0.04	0.02	0.01, 0.04	0.03	0.01, 0.04	0.01	-0.01, 0.02	0.01	-0.01, 0.02	0.00	-0.01, 0.02
	90	0.04	0.02, 0.07	0.07	0.04, 0.09	0.03	0.00, 0.05	0.05	0.02, 0.07	0.02	0.00, 0.04	0.02	0.00, 0.04
Empirical SE	1	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12
	5	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12
	10	0.12	0.11, 0.12	0.12	0.11, 0.12	0.12	0.11, 0.12	0.12	0.11, 0.12	0.12	0.11, 0.12	0.12	0.11, 0.12
	20	0.13	0.12, 0.13	0.13	0.12, 0.13	0.13	0.12, 0.13	0.13	0.12, 0.13	0.12	0.12, 0.13	0.12	0.12, 0.13
	40	0.14	0.14, 0.15	0.14	0.14, 0.15	0.14	0.14, 0.15	0.14	0.14, 0.15	0.14	0.14, 0.15	0.14	0.13, 0.15
	60	0.18	0.17, 0.19	0.18	0.17, 0.19	0.17	0.17, 0.18	0.18	0.17, 0.18	0.17	0.17, 0.18	0.17	0.16, 0.18
	80	0.27	0.26, 0.28	0.27	0.26, 0.28	0.27	0.25, 0.28	0.27	0.25, 0.28	0.26	0.25, 0.27	0.25	0.24, 0.26
	90	0.40	0.38, 0.42	0.40	0.38, 0.42	0.39	0.37, 0.40	0.39	0.37, 0.41	0.38	0.36, 0.39	0.36	0.35, 0.38
Relative error	1	0.58	-3.85, 5.01	0.57	-3.86, 4.99	0.64	-3.79, 5.07	0.64	-3.79, 5.07	0.76	-3.68, 5.19	0.69	-3.75, 5.12
	5	0.74	-3.70, 5.17	0.81	-3.63, 5.24	0.69	-3.75, 5.12	0.62	-3.81, 5.05	0.52	-3.91, 4.94	0.59	-3.84, 5.02
	10	1.19	-3.27, 5.64	1.15	-3.30, 5.61	1.01	-3.44, 5.46	0.78	-3.66, 5.22	0.84	-3.60, 5.29	0.82	-3.62, 5.26
	20	0.73	-3.71, 5.16	0.98	-3.47, 5.43	0.33	-4.09, 4.75	0.49	-3.94, 4.91	0.18	-4.23, 4.60	0.29	-4.13, 4.70
	40	1.55	-2.93, 6.03	1.52	-2.96, 6.00	1.36	-3.11, 5.83	0.46	-3.98, 4.89	1.00	-3.46, 5.45	0.12	-4.30, 4.54
	60	1.54	-2.95, 6.04	1.21	-3.26, 5.69	1.76	-2.74, 6.27	0.71	-3.74, 5.17	1.63	-2.87, 6.13	1.24	-3.25, 5.72
	80	-3.73	-8.03, 0.58	-4.01	-8.31, 0.28	-4.86	-9.12, -0.61	-4.73	-9.00, -0.47	-2.29	-6.66, 2.09	-3.70	-8.01, 0.62
	90	-0.77	-5.35, 3.82	-2.40	-6.92, 2.12	-0.42	-5.02, 4.18	-0.87	-5.44, 3.71	-2.51	-7.02, 2.00	-0.41	-5.05, 4.22

Statistic	% missing data	CCA		Model 1 No aux info, R^2=0.36		Model 2 R^2=0.40		Model 3 R^2=0.52		Model 4 R^2=0.76		Model 5 R^2=0.92	
Coverage	1	94.89	93.52, 96.26	94.89	93.52, 96.26	94.99	93.64, 96.34	94.89	93.52, 96.26	94.89	93.52, 96.26	94.99	93.64, 96.34
	5	94.69	93.30, 96.08	94.79	93.41, 96.17	94.99	93.64, 96.34	94.99	93.64, 96.34	95.19	93.86, 96.52	94.79	93.41, 96.17
	10	95.49	94.20, 96.78	95.49	94.20, 96.78	95.49	94.20, 96.78	95.19	93.86, 96.52	95.29	93.98, 96.60	95.19	93.86, 96.52
	20	95.59	94.32, 96.86	95.59	94.32, 96.86	95.29	93.98, 96.60	95.09	93.75, 96.43	95.19	93.86, 96.52	94.69	93.30, 96.08
	40	94.99	93.64, 96.34	94.99	93.64, 96.34	95.09	93.75, 96.43	95.09	93.75, 96.43	95.39	94.09, 96.69	94.89	93.52, 96.26
	60	95.49	94.20, 96.78	95.49	94.20, 96.78	95.29	93.98, 96.60	95.59	94.32, 96.86	94.69	93.30, 96.08	95.19	93.86, 96.52
	80	95.09	93.75, 96.43	95.39	94.09, 96.69	94.89	93.52, 96.26	94.19	92.74, 95.64	95.09	93.75, 96.43	95.19	93.86, 96.52
	90	96.49	95.35, 97.63	95.99	94.78, 97.21	95.49	94.20, 96.78	95.29	93.98, 96.60	95.09	93.75, 96.43	95.99	94.78, 97.21
% reduction in SE compared to CCA	1			-0.01%		0.10%		0.08%		0.23%		0.19%	
	5			0.06%		0.06%		0.01%		0.03%		0.26%	
	10			-0.01%		0.06%		-0.17%		0.15%		0.43%	
	20			0.05%		0.00%		0.08%		0.37%		1.07%	
	40			-0.28%		0.63%		0.09%		0.98%		2.14%	
	60			-0.18%		1.70%		0.87%		2.51%		4.36%	
	80			-0.17%		2.45%		2.31%		4.70%		6.90%	
	90			-0.60%		2.82%		2.27%		5.20%		8.60%	

CCA – Complete case analysis; FMI- Fraction of Missing Information; SE – Standard error; R^2 – the squared coefficient of multiple correlation

[a] For FMI the estimate was the median across simulations and the interval represents the interquartile range. For all other statistics the mean across simulations was taken and the 95% confidence interval was calculated using Monte Carlo standard error.

Table B.10-3. Performance Statistics for the MAR Results of the Exposure Coefficient in the Binary Outcome Simulation Study.

Statistic	% missing data	CCA		Model 1 No aux info, R^2=0.36		Model 2 R^2=0.40		Model 3 R^2=0.52		Model 4 R^2=0.76		Model 5 R^2=0.92	
FMI [a]	1			0.01	0.01, 0.02	0.01	0.01, 0.01	0.01	0.01, 0.01	0.01	0.00, 0.01	0.01	0.00, 0.01
	5			0.06	0.05, 0.06	0.05	0.05, 0.06	0.03	0.03, 0.04	0.03	0.03, 0.04	0.03	0.03, 0.04
	10			0.11	0.10, 0.12	0.11	0.10, 0.12	0.07	0.07, 0.08	0.07	0.06, 0.08	0.07	0.06, 0.08
	20			0.21	0.20, 0.23	0.21	0.19, 0.23	0.15	0.14, 0.17	0.15	0.13, 0.16	0.14	0.12, 0.16
	40			0.41	0.39, 0.43	0.40	0.38, 0.42	0.32	0.30, 0.35	0.31	0.29, 0.34	0.30	0.28, 0.33
	60			0.60	0.58, 0.62	0.59	0.57, 0.61	0.51	0.48, 0.54	0.51	0.47, 0.54	0.49	0.46, 0.53
	80			0.79	0.77, 0.81	0.79	0.76, 0.80	0.73	0.70, 0.76	0.72	0.69, 0.75	0.71	0.68, 0.75
	90			0.89	0.87, 0.90	0.88	0.87, 0.90	0.85	0.82, 0.87	0.84	0.81, 0.86	0.83	0.80, 0.86
Bias	1	0.00	-0.01, 0.00	0.00	-0.01, 0.00	0.00	-0.01, 0.00	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01
	5	-0.01	-0.02, -0.01	-0.01	-0.02, -0.01	-0.01	-0.02, -0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01	0.00	-0.01, 0.01
	10	-0.03	-0.04, -0.02	-0.03	-0.03, -0.02	-0.03	-0.03, -0.02	0.00	-0.01, 0.00	0.00	-0.01, 0.00	0.00	-0.01, 0.00
	20	-0.05	-0.05, -0.04	-0.05	-0.05, -0.04	-0.05	-0.05, -0.04	0.00	-0.01, 0.00	0.00	-0.01, 0.01	0.00	-0.01, 0.01
	40	-0.08	-0.09, -0.07	-0.08	-0.09, -0.07	-0.08	-0.09, -0.07	-0.01	-0.02, 0.00	-0.01	-0.02, 0.00	0.00	-0.01, 0.00
	60	-0.10	-0.11, -0.09	-0.10	-0.11, -0.09	-0.10	-0.11, -0.09	-0.01	-0.02, 0.00	-0.01	-0.02, 0.00	-0.01	-0.02, 0.00
	80	-0.11	-0.12, -0.10	-0.11	-0.13, -0.10	-0.11	-0.13, -0.10	-0.02	-0.03, 0.00	-0.01	-0.03, 0.00	-0.01	-0.02, 0.00
	90	-0.09	-0.11, -0.07	-0.10	-0.11, -0.08	-0.09	-0.11, -0.08	-0.01	-0.03, 0.01	0.00	-0.02, 0.02	0.00	-0.02, 0.01
Empirical SE	1	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12
	5	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12	0.11	0.11, 0.12
	10	0.12	0.11, 0.12	0.12	0.11, 0.12	0.12	0.11, 0.12	0.12	0.11, 0.12	0.12	0.11, 0.12	0.12	0.11, 0.12
	20	0.12	0.11, 0.13	0.12	0.12, 0.13	0.12	0.11, 0.13	0.12	0.11, 0.12	0.12	0.11, 0.12	0.12	0.11, 0.12
	40	0.13	0.13, 0.14	0.13	0.13, 0.14	0.13	0.13, 0.14	0.13	0.13, 0.14	0.13	0.13, 0.14	0.13	0.13, 0.14
	60	0.16	0.15, 0.17	0.16	0.15, 0.17	0.16	0.15, 0.17	0.16	0.16, 0.17	0.16	0.15, 0.17	0.16	0.15, 0.17
	80	0.22	0.21, 0.23	0.22	0.21, 0.23	0.22	0.21, 0.23	0.22	0.21, 0.23	0.22	0.21, 0.23	0.22	0.21, 0.23
	90	0.30	0.29, 0.32	0.30	0.29, 0.32	0.30	0.28, 0.31	0.31	0.29, 0.32	0.30	0.29, 0.32	0.29	0.28, 0.30
Relative error	1	0.73	-3.70, 5.16	0.67	-3.76, 5.10	0.71	-3.72, 5.14	0.81	-3.63, 5.25	0.77	-3.66, 5.21	0.76	-3.67, 5.20
	5	0.86	-3.57, 5.30	0.84	-3.60, 5.27	0.93	-3.51, 5.37	1.20	-3.26, 5.65	1.19	-3.27, 5.64	1.13	-3.32, 5.58
	10	0.49	-3.94, 4.91	0.56	-3.86, 4.99	0.41	-4.01, 4.83	0.77	-3.67, 5.20	0.54	-3.88, 4.97	0.42	-4.00, 4.85
	20	2.03	-2.46, 6.52	1.99	-2.50, 6.48	1.89	-2.59, 6.38	2.41	-2.10, 6.92	2.26	-2.25, 6.76	2.11	-2.39, 6.61
	40	1.56	-2.91, 6.04	1.91	-2.58, 6.40	2.44	-2.08, 6.95	1.91	-2.58, 6.40	2.32	-2.18, 6.83	1.97	-2.52, 6.47
	60	0.19	-4.23, 4.61	0.37	-4.06, 4.79	-0.26	-4.67, 4.14	-0.43	-4.82, 3.96	-1.11	-5.47, 3.26	-1.29	-5.65, 3.07
	80	-1.58	-5.94, 2.78	-0.94	-5.33, 3.46	-2.26	-6.60, 2.08	-1.28	-5.65, 3.10	-1.84	-6.20, 2.51	-3.00	-7.31, 1.31
	90	-1.45	-5.86, 2.97	-1.18	-5.61, 3.25	-1.57	-5.99, 2.85	-2.17	-6.56, 2.21	-2.45	-6.83, 1.93	-0.54	-5.01, 3.93

Statistic	% missing data	CCA		Model 1 No aux info, R^2=0.36		Model 2 R^2=0.40		Model 3 R^2=0.52		Model 4 R^2=0.76		Model 5 R^2=0.92	
Coverage	1	95.09	93.75, 96.43	95.09	93.75, 96.43	94.89	93.52, 96.26	95.09	93.75, 96.43	94.89	93.52, 96.26	94.89	93.52, 96.26
	5	94.99	93.64, 96.34	94.99	93.64, 96.34	94.99	93.64, 96.34	95.19	93.86, 96.52	95.29	93.98, 96.60	95.39	94.09, 96.69
	10	94.39	92.96, 95.82	94.49	93.07, 95.90	94.29	92.85, 95.73	95.09	93.75, 96.43	95.29	93.98, 96.60	95.29	93.98, 96.60
	20	93.89	92.40, 95.37	93.79	92.29, 95.29	93.89	92.40, 95.37	94.99	93.64, 96.34	95.49	94.20, 96.78	95.49	94.20, 96.78
	40	91.08	89.31, 92.85	90.68	88.88, 92.48	90.48	88.66, 92.30	95.39	94.09, 96.69	95.89	94.66, 97.12	95.69	94.43, 96.95
	60	88.88	86.93, 90.83	88.58	86.60, 90.55	87.88	85.85, 89.90	94.79	93.41, 96.17	94.79	93.41, 96.17	93.89	92.40, 95.37
	80	90.28	88.44, 92.12	89.88	88.01, 91.75	90.28	88.44, 92.12	94.39	92.96, 95.82	94.59	93.19, 95.99	94.09	92.62, 95.55
	90	93.59	92.07, 95.11	93.09	91.51, 94.66	92.99	91.40, 94.57	95.19	93.86, 96.52	94.49	93.07, 95.90	94.59	93.19, 95.99
% reduction in SE compared to CCA	1			-0.05%		0.00%		0.18%		0.15%		0.16%	
	5			-0.02%		0.17%		0.69%		0.72%		0.72%	
	10			0.01%		-0.07%		0.78%		0.72%		0.72%	
	20			-0.15%		0.01%		0.92%		1.03%		1.22%	
	40			0.09%		1.27%		0.63%		1.53%		1.72%	
	60			-0.10%		0.22%		-1.25%		-1.09%		-0.10%	
	80			0.39%		0.48%		-0.73%		-0.20%		0.26%	
	90			-0.07%		1.58%		-1.63%		0.01%		4.15%	
% reduction in bias compared to CCA	1			2.51%		3.72%		86.84%		90.22%		93.93%	
	5			-2.52%		-6.71%		88.27%		91.64%		93.16%	
	10			0.77%		2.52%		86.72%		85.34%		88.59%	
	20			-0.19%		0.42%		91.25%		95.04%		97.80%	
	40			-4.20%		-2.67%		89.05%		91.25%		93.78%	
	60			-2.47%		-3.38%		88.28%		90.88%		93.12%	
	80			-3.77%		-3.12%		83.97%		86.65%		89.63%	
	90			-3.11%		-1.05%		91.36%		96.95%		95.96%	

CCA – Complete case analysis; FMI- Fraction of Missing Information; SE – Standard error; R^2 – the squared coefficient of multiple correlation

[a] For FMI the estimate was the median across simulations and the interval represents the interquartile range. For all other statistics the mean across simulations was taken and the 95% confidence interval was calculated using Monte Carlo standard error.

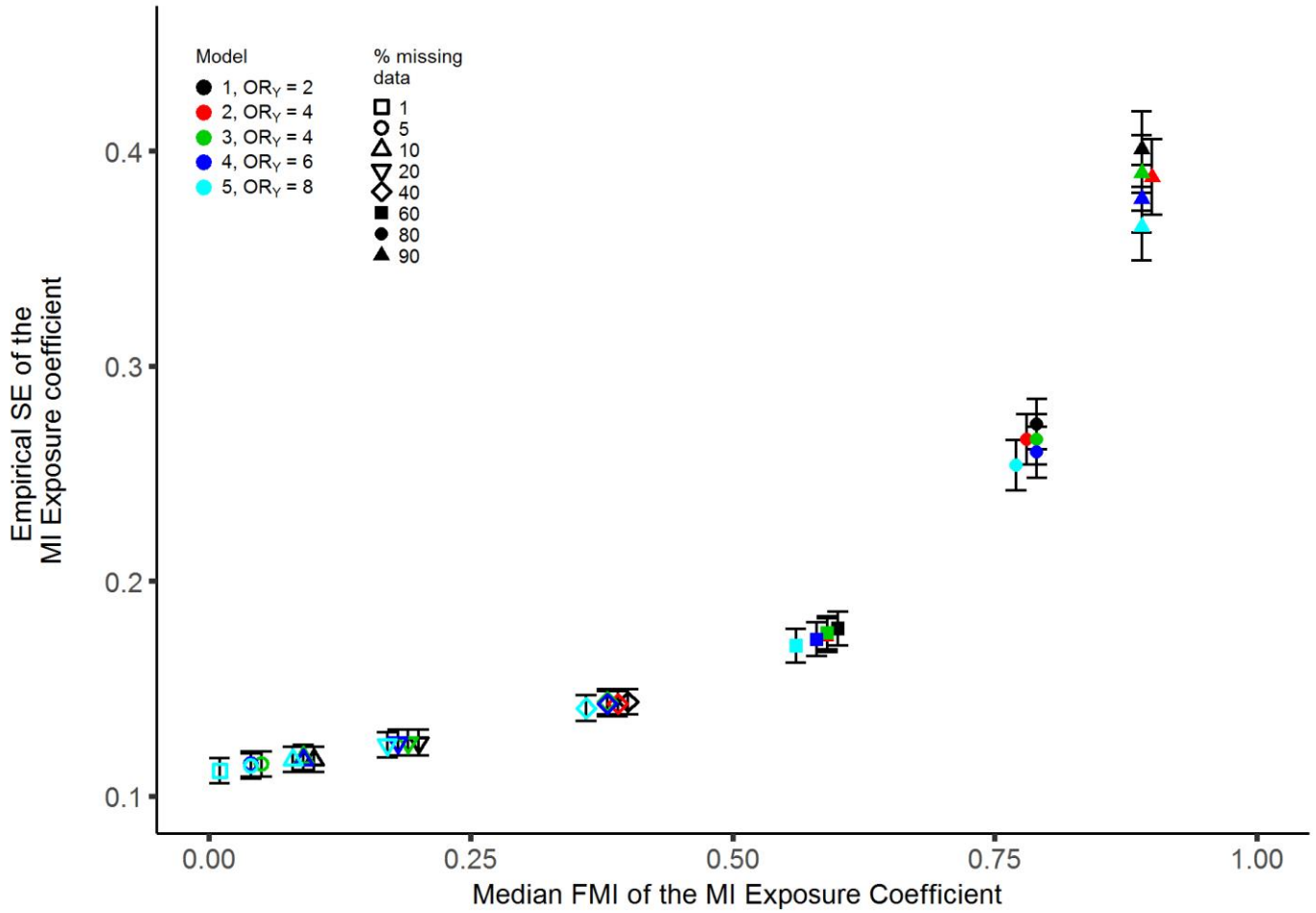

Figure B.10-1: Empirical SE of the MI exposure coefficient plotted against FMI for simulated MCAR binary outcome data. Error bars are 95% confidence intervals based on Monte Carlo standard errors across simulations. FMI = fraction of missing information; SE = standard error.

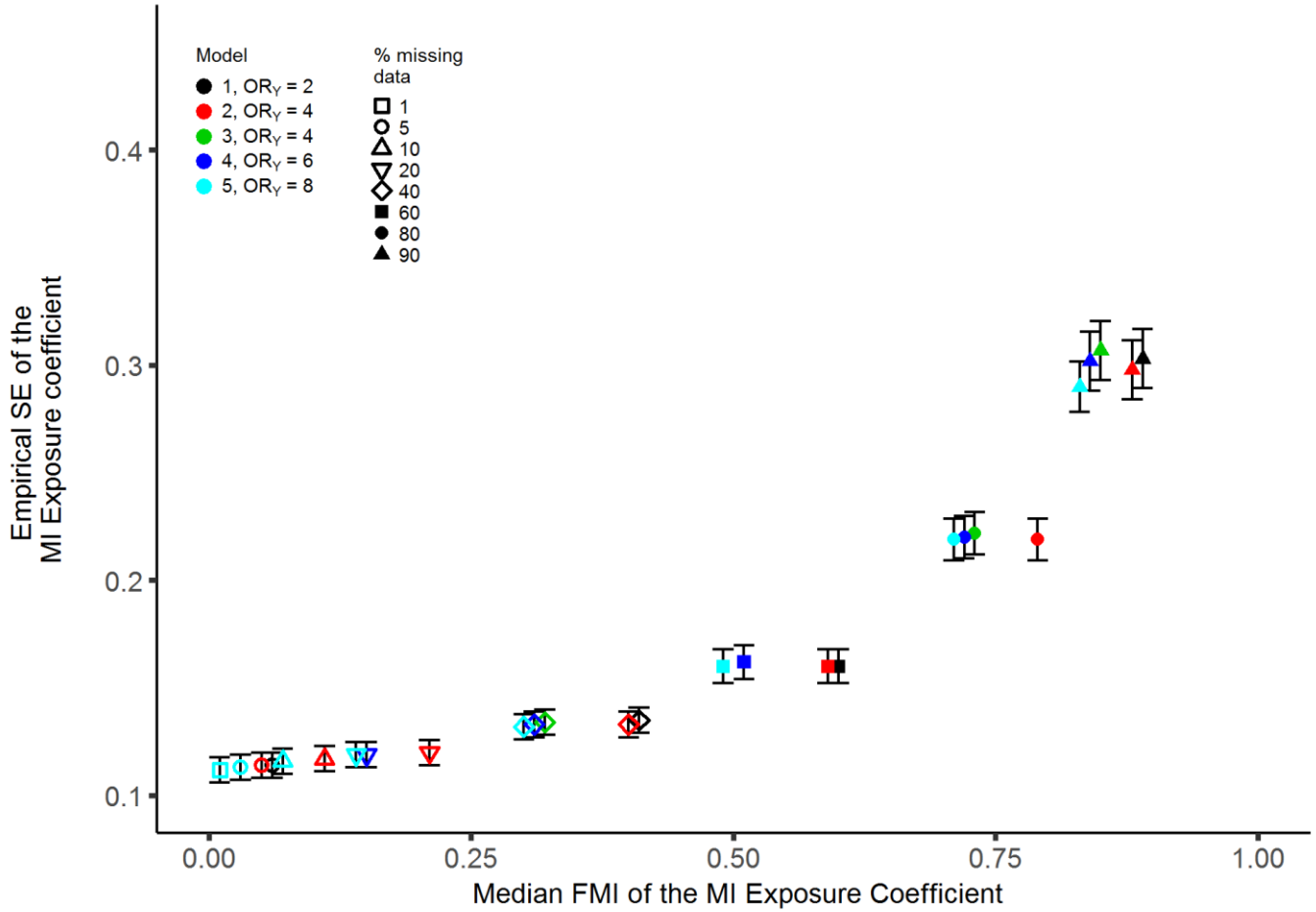

Figure B.10-2: Empirical SE of the MI exposure coefficient plotted against FMI for simulated MAR binary outcome data. Error bars are 95% confidence intervals based on Monte Carlo standard errors across simulations. FMI = fraction of missing information; SE = standard error.

Appendix C – Supplementary material to Chapter 5

C.1 – Methods

C.1.1 – Exclusion criteria

Table C.1.1-1: ICD-10 diagnoses used for exclusion.

ICD-10 Code	Description
Q89.8	Williams syndrome
Q87.1	Prader-Willi syndrome
Q87.2	Rubinstein-Taybi syndrome
Q44.7	Alagille syndrome
D82.1	DiGeorge syndrome
Q85.0	Neurofibromatosis (non-malignant)
Q85.1	Tuberous sclerosis
Q90-Q99	Chromosomal abnormalities, not elsewhere specified
E70-E72	Metabolic disorders

C.1.2 – Psychiatric history diagnosis codes

Table C.1.2-1: ICD-8 and ICD-10 diagnosis codes used to define parental psychiatric history.

Disorder	ICD-8 codes	ICD-10 codes
Anxiety disorders	300.x9 (excluding 300.49), 305.x9, 305.68, 307.99	F40-F48
Depressive disorders	296.09 ,296.29, 298.09, 300.49	F32-F39
Affective and non-affective psychoses	295.x9, 296.39, 296.89, 297.x9, 298.29-298.99, 299.04, 299.05, 299.09, 301.83, 296.19, 298.19	F20-F29, F30-F31
Substance use disorders	291.x9, 294.39, 303.x9, 303.20, 303.28, 303.90, 304.x9	F10-F16, F18-F19

C.2 – Results

C.2.1 – Cohort descriptives

Table C.2.1-1: Distribution of changes in smoking from one pregnancy to the next.

	Parity, N(%)					
	0/1	1/2	2/3	3/4	4/5	5/6
Smoked in neither	215,006 (81.84)	71,405 (81.69)	16,330 (76.71)	4,309 (75.84)	1,539 (77.96)	662 (80.63)
Smoked in both	26,027 (9.91)	10,016 (11.46)	3,401 (15.98)	1,014 (17.85)	330 (16.72)	124 (15.10)
Started smoking	6,710 (2.55)	2,274 (2.60)	595 (2.80)	147 (2.59)	46 (2.33)	15 (1.83)
Stopped smoking	14,982 (5.70)	3,712 (4.25)	962 (4.52)	212 (3.73)	59 (2.99)	20 (2.44)

Table C.2.1-2: Distribution of changes in smoking from one pregnancy to the next separated by the exposure and outcome status of the latter pregnancy.

	Maternal smoking in pregnancy, N(%)		Intellectual Disability, N(%)	
	No	Yes	No	Yes
Smoked in neither	315,170 (93.93)	-	313,663 (81.49)	1507 (68.01)
Smoked in both	-	41,611 (80.66)	41,118 (10.68)	493 (22.25)
Started smoking	-	9,975 (19.34)	9,892 (2.57)	83 (3.75)
Stopped smoking	20,359 (6.07)	-	20,226 (5.25)	133 (6.00)

Table C.2.1-3: Proportions of ID in each level of the family level exposure variable.

	Siblings in a family exposed to smoking in pregnancy, N(%)			
	None	Some	All	Total
Counts for individuals				
-No ID	829,246 (99.39)	66,288 (99.24)	163,404 (98.54)	1,058,938 (99.25)
-ID	5,118 (0.61)	510 (0.76)	2,423 (1.46)	8,051 (0.75)
- Total	834,364 (100.0)	66,798 (100.0)	165,827 (100.0)	1,066,990 (100.0)
Counts for families				
-No ID	500,808 (99.01)	28,255 (98.29)	121,407 (98.08)	650,470 (98.81)
-At least 1 case of ID	4,999 (0.99)	493 (1.71)	2,373 (1.92)	7,865 (1.19)
-Total	505,807 (100.0)	28,748 (100.0)	123,780 (100.0)	658,335 (100.0)

Table C.2.1-4: Distribution of family smoking variable among families with children born before the cohort start date versus families with children all born after the cohort start date.

Family level exposure value	Families with excluded older siblings, N(%)	Families with first born included in cohort, N(%)
0 (smoked in no pregnancies)	145,837 (70.41)	359,970 (79.78)
0.01-0.33	84 (0.04)	459 (0.10)
0.34-0.66	3,924 (1.89)	21,278 (4.72)
0.67-0.99	449 (0.22)	2,554 (0.57)
1 (smoked in all pregnancies)	56,827 (27.44)	66,953 (14.84)

Table C.2.1-5: Distribution of smoking among only children and first-born children.

	N(%)	
	Only child	First born child
Non-smoker	134,458 (76.81)	242,190 (84.47)
Smoker	40,585 (23.19)	44,526 (15.53)

C.2.2 – Missing data assessment

Table C.2.2-1: Descriptive statistics for those included in the main cohort versus those excluded for having missing covariates.

Characteristic	Excluded for missing covariates, N(%)	Included in main cohort, N(%)	O.R.(95% CI) [a]	p-value
Intellectual disability				
-Yes	571 (1.09)	8,051 (0.75)	1.46 (1.34-1.59)	<.001
-No	51,586 (98.91)	1,058,938 (99.25)	Ref	-
Maternal smoking				
-Yes	1,798 (13.26)	198,377 (18.59)	0.67 (0.64-0.70)	<.001
-No	11,759 (86.74)	868,612 (81.41)	Ref	-
Maternal age				
-<20	1,044 (2.01)	14,682 (1.38)	1.49 (1.40-1.59)	<.001
-20-24	7,416 (14.25)	125,515 (11.76)	1.24 (1.21-1.28)	<.001
-25-29	17,455 (33.53)	366,516 (34.35)	Ref	-
-30-34	17,229 (33.10)	380,506 (35.66)	0.95 (0.93-0.97)	<.001
-35+	8,913 (17.12)	179,770 (16.85)	1.04 (1.01-1.07)	.003
Paternal age				
-<20	254 (0.49)	4,238 (0.40)	1.24 (1.09-1.41)	.001
-20-24	3,383 (6.51)	61,559 (5.77)	1.13 (1.09-1.18)	<.001
-25-29	12,811 (24.64)	264,433 (24.78)	Ref	-
-30-34	18,145 (34.90)	392,876 (36.82)	0.95 (0.93-0.98)	<.001
-35+	17,397 (33.46)	343,882 (32.23)	1.04 (1.02-1.07)	<.001
Maximum parental education				
-Primary	5,594 (13.17)	107,273 (10.05)	Ref	-
-General/Vocational	19,380 (45.64)	462,159 (43.31)	0.80 (0.78-0.83)	<.001
-Higher	17,486 (41.18)	497,557 (46.63)	0.67 (0.65-0.70)	<.001
Parental income decile				
-1	9,483 (18.19)	102,424 (9.60)	2.20 (2.12-2.28)	<.001
-2	6,164 (11.82)	105,749 (9.91)	1.38 (1.33-1.44)	<.001
-3	5,361 (10.28)	106,556 (9.99)	1.19 (1.15-1.24)	<.001
-4	5,020 (9.63)	106,894 (10.02)	1.11 (1.07-1.16)	<.001
-5	4,526 (8.68)	107,385 (10.06)	Ref	-
-6	4,595 (8.81)	107,326 (10.06)	1.02 (0.97-1.06)	.463
-7	4,523 (8.67)	107,392 (10.06)	1.00 (0.96-1.04)	.973

Characteristic	Excluded for missing covariates, N(%)	Included in main cohort, N(%)	O.R.(95% CI) [a]	p-value
-8	4,307 (8.26)	107,607 (10.09)	0.95 (0.91-0.99)	.017
-9	3,970 (7.61)	107,946 (10.12)	0.87 (0.84-0.91)	<.001
-10	4,196 (8.05)	107,710 (10.09)	0.92 (0.89-0.96)	<.001
Maternal country of origin				
-Denmark	35,852 (71.82)	928,190 (86.99)	Ref	-
-Africa	1,839 (3.68)	17,130 (1.61)	2.78 (2.65-2.92)	<.001
-Americas	533 (1.07)	6,013 (0.56)	2.29 (2.10-2.51)	<.001
-Europe	5,341 (10.70)	46,295 (4.34)	2.99 (2.90-3.08)	<.001
-Middle East	2,335 (4.68)	25,403 (2.38)	2.38 (2.28-2.49)	<.001
-Oceania	2,284 (4.58)	29,118 (2.73)	2.03 (1.94-2.12)	<.001
-Scandinavia	1,738 (3.48)	14,840 (1.39)	3.03 (2.88-3.19)	<.001
Paternal country of origin				
-Denmark	35,622 (72.28)	931,110 (87.27)	Ref	-
-Africa	1,829 (3.71)	18,980 (1.78)	2.52 (2.40-2.65)	<.001
-Americas	443 (0.90)	5,626 (0.53)	2.06 (1.87-2.27)	<.001
-Europe	5,393 (10.94)	50,035 (4.69)	2.82 (2.73-2.90)	<.001
-Middle East	2,521 (5.12)	30,348 (2.84)	2.17 (2.08-2.26)	<.001
-Oceania	1,903 (3.86)	20,046 (1.88)	2.48 (2.36-2.60)	<.001
-Scandinavia	1,572 (3.19)	10,844 (1.02)	3.79 (3.59-4.00)	<.001
Maternal psychiatric history				
-Affective disorder				
-Yes	755 (1.45)	18,343 (1.72)	0.84 (0.78-0.90)	<.001
-No	51,402 (98.55)	1,048,646 (98.28)	Ref	-
-Anxiety disorder				
-Yes	1,659 (3.18)	39,859 (3.74)	0.85 (0.81-0.89)	<.001
-No	50,498 (96.82)	1,027,130 (96.26)	Ref	-
-Psychotic disorder				
-Yes	263 (0.50)	5,072 (0.48)	1.06 (0.94-1.20)	.350
-No	51,894 (99.50)	1,061,917 (99.52)	Ref	-
-Substance use disorder				
-Yes	738 (1.41)	18,020 (1.69)	0.84 (0.78-0.90)	<.001

Characteristic	Excluded for missing covariates, N(%)	Included in main cohort, N(%)	O.R.(95% CI) [a]	p-value
-No	51,419 (98.59)	1,048,969 (98.31)	Ref	-
Paternal psychiatric history				
-Affective disorder				
-Yes	319 (0.61)	6,708 (0.63)	0.97 (0.87-1.09)	.630
-No	51,838 (99.39)	1,060,281 (99.37)	Ref	-
-Anxiety disorder				
-Yes	842 (1.61)	1,7403 (1.63)	0.99 (0.92-1.06)	.769
-No	51,315 (98.39)	1,049,586 (98.37)	Ref	-
-Psychotic disorder				
-Yes	283 (0.54)	4,977 (0.47)	1.16 (1.03-1.31)	.013
-No	51,874 (99.46)	1,062,012 (99.53)	Ref	-
-Substance use disorder				
-Yes	1,048 (2.01)	25,539 (2.39)	0.84 (0.79-0.89)	<.001
-No	51,109 (97.99)	1,041,450 (97.61)	Ref	-
Child sex				
-Female	25,306 (48.52)	519,856 (48.72)	Ref	-
-Male	26,851 (51.48)	547,133 (51.28)	1.01 (0.99-1.03)	.365
Parity				
-0	22,836 (43.79)	461,759 (43.28)	Ref	-
-1	18,321 (35.13)	400,499 (37.54)	0.93 (0.91-0.94)	<.001
-2	7,423 (14.23)	152,496 (14.29)	0.98 (0.96-1.01)	.247
-3+	3,569 (6.84)	52,235 (4.9)	1.38 (1.33-1.43)	<.001
Cohort year				
-1995-1997	13,575 (26.03)	180,394 (16.91)	Ref	-
-1998-2000	11,862 (22.74)	182,132 (17.07)	0.87 (0.84-0.89)	<.001
-2001-2003	6,704 (12.85)	177,831 (16.67)	0.50 (0.49-0.52)	<.001
-2004-2006	6,070 (11.64)	180,674 (16.93)	0.45 (0.43-0.46)	<.001
-2007-2009	7,310 (14.02)	178,532 (16.73)	0.54 (0.53-0.56)	<.001
-2010-2012	6,636 (12.72)	167,426 (15.69)	0.53 (0.51-0.54)	<.001

[a] Odds Ratio for exclusion from the cohort due to missing covariates.

Table C.2.2-2: Number of observations with missing data in each variable.

Variable	Number of missing values	% of those excluded [a, b]	% of total sample [a, c]

Maternal smoking in pregnancy	38,600	74.01	3.45
Highest parental education	9,697	18.59	0.87
Paternal country of origin	2,874	5.51	0.26
Maternal country of origin	2,235	4.29	0.20
Paternal age	167	0.32	0.01
Maternal age	100	0.19	0.01
Income decile	12	0.02	<0.01
Parity	8	0.02	<0.01

[a] Groups not mutually exclusive so percentages may not add to 100%.
[b] Percentage denominator equals 93,190
[c] Percentage denominator equals 1,119,146

Table C.2.2-3: Distribution of missing smoking data by each cohort year group.

Year group	Excluded for missing smoking data, N(%) [a]	Included in main cohort, N(%) [b]
1995-1997	10,770 (27.90)	183,199 (16.95)
1998-2000	9,652 (25.01)	184,342 (17.06)
2001-2003	4,865 (12.60)	179,670 (16.63)
2004-2006	4,478 (11.60)	182,266 (16.87)
2007-2009	5,041 (13.06)	180,801 (16.73)
2010-2012	3,794 (9.83)	170,268 (15.76)

[a] Percentage denominator equals 38,600
[b] Percentage denominator equals 1,066,989

Table C.2.2-4: Distribution of missing smoking data by parity.

Parity	Excluded for missing smoking data, N(%) [a]	Included in main cohort, N(%) [b]
0	16,771 (43.45)	467,824 (43.30)
1	14,066 (36.44)	404,754 (37.46)
2	5,496 (14.24)	154,423 (14.29)
3+	2,265 (5.87)	53,539 (4.95)

[a] Percentage denominator equals 38,600
[b] Percentage denominator equals 1,066,989

C.2.3 – Secondary analyses

Table C.2.3-1: Multinomial logistic GEE analyses of the association between maternal smoking in pregnancy and each ID severity category

	Unadjusted analyses	
Coefficient	O.R.	95% CI
- F70 – Mild	2.17	2.04 - 2.31
- F71 – Moderate	1.53	1.35 - 1.73
- F72 – Severe	1.20	0.96 - 1.50
- F73 – Profound	1.37	0.97 - 1.94
- F78/F79 – Other/unspecified	1.77	1.59 - 1.97

All odds ratios are relative to the group with no ID

Table C.2.3-2: Multinomial logistic GEE analyses of the association between maternal smoking in pregnancy and each category of ID and ASD comorbidity

		ID only		ASD only		ID + ASD	
Model	Coefficient	O.R.	95% CI	O.R.	95% CI	O.R.	95% CI
Unadjusted	- Population averaged	2.21	2.09 - 2.34	1.37	1.32 - 1.42	1.40	1.29 - 1.53
Adjusted for confounders [a]	- Population averaged	1.42	1.33 - 1.51	1.24	1.20 - 1.29	1.21	1.10 - 1.32
Adjusted for family smoking variable	- Within-family	0.88	0.74 - 1.05	1.16	1.05 - 1.29	0.98	0.75 - 1.28
	- Between-family	2.76	2.29 - 3.32	1.21	1.09 - 1.35	1.50	1.13 - 1.99
Adjusted for confounders [a] and family smoking variable	- Within-family	0.91	0.76 - 1.10	1.07	0.96 - 1.19	0.96	0.73 - 1.28
	- Between-family	1.63	1.33 - 1.98	1.19	1.06 - 1.34	1.29	0.95 - 1.74

All odds ratios are relative to the group with no ID or ASD

[a] Adjusted for child parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

Table C.2.3-3: Multinomial logistic GEE analyses of the association between maternal smoking in pregnancy and each category of ID and ADHD comorbidity

		ID only		ADHD only		ID + ADHD	
Model	Coefficient	O.R.	95% CI	O.R.	95% CI	O.R.	95% CI
Unadjusted	- Population averaged	1.76	1.66 - 1.86	2.27	2.20 - 2.34	2.43	2.22 - 2.66
Adjusted for confounders [a]	- Population averaged	1.30	1.22 - 1.38	1.63	1.58 - 1.69	1.54	1.40 - 1.70
Adjusted for family smoking variable	- Within-family	0.92	0.77 - 1.09	0.99	0.90 - 1.09	0.88	0.67 - 1.16
	- Between-family	2.07	1.72 - 2.48	2.56	2.31 - 2.83	3.12	2.33 - 4.17
Adjusted for confounders [a] and family smoking variable	- Within-family	0.94	0.78 - 1.13	0.95	0.86 - 1.05	0.90	0.67 - 1.21
	- Between-family	1.44	1.19 - 1.75	1.86	1.67 - 2.07	1.84	1.35 - 2.50

All odds ratios are relative to the group with no ID or ADHD

[a] Adjusted for child parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

Table C.2.3-4: Logistic GEE analyses of the association between maternal smoking during pregnancy and offspring ID, including an interaction between maternal smoking and offspring sex

Model	Coefficient	O.R.	95% CI
Unadjusted	- Smoking in pregnancy	1.93	1.78, 2.10
	- Male sex	1.93	1.82, 2.04
	- Interaction term	0.98	0.89, 1.08
Adjusted for confounders [a]	- Smoking in pregnancy	1.37	1.26, 1.49
	- Male sex	1.94	1.83, 2.05
	- Interaction term	0.98	0.89, 1.08

[a] Adjusted for child parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

Table C.2.3-5: Logistic GEE analyses of the association between maternal smoking cessation in pregnancy and offspring ID

Model	Coefficient [a]	O.R.	95% CI
Unadjusted	- Stopped smoking in 1st trimester	1.25	1.04, 1.50
	- Continued smoking after 1st trimester	2.01	1.91, 2.12
Adjusted for confounders [b]	- Stopped smoking in 1st trimester	1.09	0.91, 1.32
	- Continued smoking after 1st trimester	1.39	1.31, 1.47

[a] Reference group is non-smoking mothers

[b] Adjusted for child sex, parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

Table C.2.3-6: Logistic GEE analyses of the association between dosage of maternal smoking during pregnancy and offspring ID

Model	Coefficient	O.R. [a]	95% CI
Unadjusted	- Population averaged	1.05	1.05, 1.06
Adjusted for confounders [b]	- Population averaged	1.03	1.02, 1.04
Adjusted for family dosage variable	- Within-family	1.00	0.98, 1.03
	- Between-family	1.06	1.03, 1.09
Adjusted for confounders [b] and family dosage variable	- Within-family	1.01	0.98, 1.03
	- Between-family	1.03	1.00, 1.06

[a] OR for ID per 1 additional cigarette smoked in pregnancy

[b] Adjusted for child sex, parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

C.2.4 – Sensitivity analyses

Table C.2.4-1: Logistic GEE analyses of the association between maternal smoking during pregnancy and offspring ID using a stricter outcome definition

Model	Coefficient	O.R.	95% CI
Unadjusted	- Population averaged	1.89	1.77, 2.01
Adjusted for confounders [a]	- Population averaged	1.35	1.26, 1.44
Adjusted for family smoking variable	- Within-family	0.90	0.73, 1.10
	- Between-family	2.26	1.82, 2.80
Adjusted for confounders [a] and family smoking variable	- Within-family	0.91	0.74, 1.13
	- Between-family	1.54	1.23, 1.93

[a] Adjusted for child sex, parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

Table C.2.4-2: Logistic GEE analyses of the association between maternal smoking during pregnancy and offspring ID using an exposure variable centred on the family-level smoking variable

Model	Coefficient	O.R.	95% CI
Adjusted for family smoking variable with centred individual smoke exposure variable	- Within-family	0.91	0.78, 1.06
	- Between-family	2.05	1.95, 2.15
Adjusted for confounders [a] and family smoking variable with centred individual smoke exposure variable	- Within-family	0.93	0.79, 1.09
	- Between-family	1.40	1.33, 1.48

[a] Adjusted for child sex, parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

Table C.2.4-3 Logistic GEE analyses of the association between maternal smoking during pregnancy and offspring ID repeated in each year group category

		O.R. (95% CI)					
Model	Coefficient	1995-1997	1998-2000	2001-2003	2004-2006	2007-2009	2010-2012
Unadjusted	- Population averaged	1.82 (1.67, 1.98)	1.73 (1.58, 1.89)	2.12 (1.90, 2.36)	2.17 (1.90, 2.49)	2.04 (1.72, 2.41)	1.93 (1.54, 2.44)
Adjusted for confounders[a]	- Population averaged	1.28 (1.17, 1.41)	1.21 (1.10, 1.33)	1.50 (1.33, 1.69)	1.49 (1.28, 1.72)	1.44 (1.19, 1.73)	1.50 (1.17, 1.94)
Adjusted for family smoking variable	- Within-family	0.98 (0.69, 1.40)	0.74 (0.56, 0.97)	0.95 (0.71, 1.28)	0.99 (0.68, 1.43)	0.89 (0.56, 1.42)	1.72 (0.70, 4.22)
	- Between-family	1.93 (1.35, 2.78)	2.55 (1.91, 3.40)	2.46 (1.78, 3.39)	2.44 (1.63, 3.64)	2.51 (1.53, 4.11)	1.13 (0.45, 2.87)
Adjusted for family smoking variable and confounders[a]	- Within-family	1.03 (0.71, 1.48)	0.78 (0.58, 1.05)	1.01 (0.73, 1.39)	0.97 (0.65, 1.44)	0.80 (0.49, 1.31)	1.46 (0.57, 3.73)
	- Between-family	1.27 (0.87, 1.85)	1.62 (1.19, 2.21)	1.57 (1.11, 2.23)	1.63 (1.06, 2.49)	1.93 (1.14, 3.29)	1.03 (0.39, 2.74)

[a] Adjusted for child sex and parity, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

Table C.2.4-4: Cox proportional hazards analyses of the association between maternal smoking during pregnancy and time to offspring first diagnosis of ID

Model	Coefficient	H.R.	95% CI
Unadjusted	- Population averaged	1.87	1.78, 1.96
Adjusted for confounders [a]	- Population averaged	1.23	1.17, 1.30
Adjusted for family smoking variable	- Within-family	0.91	0.79, 1.05
	- Between-family	2.20	1.89, 2.57
Adjusted for confounders [a] and family smoking variable	- Within-family	0.89	0.76, 1.04
	- Between-family	1.44	1.21, 1.70

[a] Adjusted for child sex, parity and year of birth, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.

Table C.2.4-5: Analyses of the association between maternal smoking during pregnancy and offspring ID repeated in restricted cohorts

		O.R. (95% CI)		
Model	Coefficient	Single-child cohort [a]	Multiple-child cohort [b]	Multiple-child cohort with no missing older siblings [b]
Unadjusted	- Population averaged	1.54 (1.38, 1.71)	1.94 (1.84, 2.04)	1.87 (1.76, 1.99)
Adjusted for confounders [b]	- Population averaged	1.13 (1.01, 1.27)	1.37 (1.29, 1.45)	1.33 (1.24, 1.42)
Adjusted for family smoking variable	- Within-family	-	0.91 (0.78, 1.06)	0.92 (0.77, 1.10)
	- Between-family	-	2.33 (1.97, 2.75)	2.23 (1.85, 2.69)
Adjusted for confounders [c] and family smoking variable	- Within-family	-	0.96 (0.81, 1.13)	0.92 (0.76, 1.11)
	- Between-family	-	1.49 (1.25, 1.79)	1.52 (1.24, 1.86)

[a] Logistic regression model.
[b] Logistic GEE model with exchangeable covariance structure.
[c] Adjusted for child sex, mother and father's age, education and income in the year of the child's birth, the psychiatric history of mother and father prior to the child's birth and mother and father's country of origin.
[d] Adjusted as for [c] but with additional adjustment for child parity.

Appendix D – Supplementary material to Chapter 6

D.1 – Methods

D.1.1 – Exclusion criteria

Table D.1.1-1 – ICD-10 diagnosis codes for genetic and chromosomal abnormalities associated with intellectual disability that were used in exclusion criteria.

ICD-10 Code	Description
Q89.8	Williams syndrome
Q87.1	Prader-Willi syndrome
Q87.2	Rubinstein-Taybi syndrome
Q44.7	Alagille syndrome
D82.1	DiGeorge syndrome
Q85.0	Neurofibromatosis (non-malignant)
Q85.1	Tuberous sclerosis
Q90-Q99	Chromosomal abnormalities, not elsewhere specified
E70-E72	Metabolic disorders

D.1.2 – Exposure imputation procedure

Table D.1.2-1 – Imputation procedure for missing smoking values

Original value			Imputed value		
Smoking 3 months prior to pregnancy	Smoking at first antenatal contact	Smoking at 30-32 weeks pregnancy	Smoking 3 months prior to pregnancy	Smoking at first antenatal contact	Smoking at 30-32 weeks pregnancy
Non-smoker	Missing	Missing	X	Non-smoker	Non-smoker
Non-smoker	Missing	Non-smoker	X	Non-smoker	X
Any value	Non-smoker	Missing	X	X	Non-smoker
1-9 cig or 10+ cig	Missing	1-9 cig or 10+ cig	X	Value at 30-32 weeks pregnancy	X
Missing	1-9 cig or 10+ cig	Any value	Value at first antenatal visit	X	X

X – value not imputed

Table D.1.2-2 - Imputation procedure for missing snus use values

Original value			Imputed value		
Snus use at 3 months prior to pregnancy	Snus use at first antenatal contact	Snus use at 30-32 weeks pregnancy	Snus use 3 months prior to pregnancy	Snus use at first antenatal contact	Snus use at 30-32 weeks pregnancy
Non-user	Missing	Missing	X	Non-user	Non-user
Non-user	Missing	Any value	X	Non-user	X
Any value	Non-user	Missing	X	X	Non-user
Snus-user	Missing	Snus-user	X	Snus-user	X
Missing	Snus-user	Any value	Snus-user	X	X
Missing	Missing	Snus-user	Snus-user	X	X

X – value not imputed

D.1.3 – Confounder derivations

Table D1.3-1: ICD-9 and ICD-10 codes used to define parental psychiatric history.

Disorder	ICD-9 codes	ICD-10 codes
Anxiety disorders	300.0, 300.2, 300.3, 308, 309	F40-F43
Depressive disorders	296.1, 298.0, 300.4, 311	F32-F39
Psychotic disorders (affective and non-affective)	295, 296.0, 296.2, 296.3, 296.4, 297, 298.2, 298.3, 298.4, 298.8, 298.9	F20-F29, F30-F31
Substance use disorders	291.0-291.9 303.0-303.9 304.0-304.9 305.0-305.9 (*minus 305.1*)	F10-F16, F18-F19

D.2 – Results

D.2.1 – Cohort descriptives

Table D.2.1-1: Descriptives separated by categories of family level smoking

		N(%)					
Variable	Level	Never smoked in pregnancy	Sometimes smoked in pregnancy	Always smoked in pregnancy	Never used snus in pregnancy	Sometimes used snus in pregnancy	Always used snus in pregnancy
Total		959785	37999	76304	1052233	13327	8528
Intellectual disability	No	954023 (99.40)	37675 (99.15)	75398 (98.81)	1045430 (99.35)	13219 (99.19)	8447 (99.05)
	Yes	5762 (0.60)	324 (0.85)	906 (1.19)	6803 (0.65)	108 (0.81)	81 (0.95)
Small for gestational age	No	938591 (97.79)	36969 (97.29)	72638 (95.20)	1026839 (97.59)	13070 (98.07)	8289 (97.20)
	Yes	18264 (1.90)	909 (2.39)	3401 (4.46)	22134 (2.10)	231 (1.73)	209 (2.45)
Sex	Female	466545 (48.61)	18486 (48.65)	36897 (48.36)	511285 (48.59)	6481 (48.63)	4162 (48.80)
	Male	493240 (51.39)	19513 (51.35)	39407 (51.64)	540948 (51.41)	6846 (51.37)	4366 (51.20)
Parity	1	432971 (45.11)	12623 (33.22)	31949 (41.87)	468521 (44.53)	4779 (35.86)	4243 (49.75)
	2	354399 (36.92)	15280 (40.21)	23462 (30.75)	385200 (36.61)	5419 (40.66)	2522 (29.57)
	3 or more	172415 (17.96)	10096 (26.57)	20893 (27.38)	198512 (18.87)	3129 (23.48)	1763 (20.67)
Highest parental education	High School	34024 (3.54)	4126 (10.86)	11793 (15.46)	48939 (4.65)	608 (4.56)	396 (4.64)
	Gymnasium	367788 (38.32)	24354 (64.09)	52220 (68.44)	432316 (41.09)	7080 (53.13)	4966 (58.23)
	University	557973 (58.14)	9519 (25.05)	12291 (16.11)	570978 (54.26)	5639 (42.31)	3166 (37.12)
Adjusted family income	1	99291 (10.35)	6600 (17.37)	15462 (20.26)	119113 (11.32)	1391 (10.44)	849 (9.96)
	2	187435 (19.53)	12185 (32.07)	27522 (36.07)	221285 (21.03)	3556 (26.68)	2301 (26.98)
	3	215754 (22.48)	9637 (25.36)	17706 (23.20)	237259 (22.55)	3509 (26.33)	2329 (27.31)
	4	229068 (23.87)	6453 (16.98)	10818 (14.18)	241616 (22.96)	2902 (21.78)	1821 (21.35)
	5	228237 (23.78)	3124 (8.22)	4796 (6.29)	232960 (22.14)	1969 (14.77)	1228 (14.40)

Variable	Level	N(%)					
		Never smoked in pregnancy	Sometimes smoked in pregnancy	Always smoked in pregnancy	Never used snus in pregnancy	Sometimes used snus in pregnancy	Always used snus in pregnancy
Maternal anxiety diagnosis	No	934219 (97.34)	35801 (94.22)	70041 (91.79)	1019456 (96.89)	12625 (94.73)	7980 (93.57)
	Yes	25566 (2.66)	2198 (5.78)	6263 (8.21)	32777 (3.11)	702 (5.27)	548 (6.43)
Maternal depression diagnosis	No	942002 (98.15)	36523 (96.12)	72215 (94.64)	1029797 (97.87)	12848 (96.41)	8095 (94.92)
	Yes	17783 (1.85)	1476 (3.88)	4089 (5.36)	22436 (2.13)	479 (3.59)	433 (5.08)
Maternal psychosis diagnosis	No	956910 (99.70)	37795 (99.46)	75518 (98.97)	1048530 (99.65)	13239 (99.34)	8454 (99.13)
	Yes	2875 (0.30)	204 (0.54)	786 (1.03)	3703 (0.35)	88 (0.66)	74 (0.87)
Maternal addiction diagnosis	No	949629 (98.94)	36400 (95.79)	71084 (93.16)	1036110 (98.47)	12887 (96.70)	8116 (95.17)
	Yes	10156 (1.06)	1599 (4.21)	5220 (6.84)	16123 (1.53)	440 (3.30)	412 (4.83)
Any maternal psychiatric diagnosis	No	915371 (95.37)	33944 (89.33)	64705 (84.80)	994412 (94.50)	12102 (90.81)	7506 (88.02)
	Yes	44414 (4.63)	4055 (10.67)	11599 (15.20)	57821 (5.50)	1225 (9.19)	1022 (11.98)
Any paternal psychiatric diagnosis	No	927360 (96.62)	35224 (92.70)	68321 (89.54)	1010377 (96.02)	12573 (94.34)	7955 (93.28)
	Yes	32425 (3.38)	2775 (7.30)	7983 (10.46)	41856 (3.98)	754 (5.66)	573 (6.72)
Any maternal neurodevelopmental diagnosis	No	958597 (99.88)	37837 (99.57)	75623 (99.11)	1050316 (99.82)	13277 (99.62)	8464 (99.25)
	Yes	1188 (0.12)	162 (0.43)	681 (0.89)	1917 (0.18)	50 (0.38)	64 (0.75)
Any paternal neurodevelopmental diagnosis	No	958374 (99.85)	37832 (99.56)	75622 (99.11)	1050050 (99.79)	13287 (99.70)	8491 (99.57)
	Yes	1411 (0.15)	167 (0.44)	682 (0.89)	2183 (0.21)	40 (0.30)	37 (0.43)
Maternal country of origin	Africa	29014 (3.02)	572 (1.51)	668 (0.88)	30046 (2.86)	163 (1.22)	45 (0.53)
	Americas	10380 (1.08)	311 (0.82)	516 (0.68)	11108 (1.06)	67 (0.50)	32 (0.38)
	Asia	30643 (3.19)	647 (1.70)	1162 (1.52)	32133 (3.05)	189 (1.42)	130 (1.52)
	Europe	46087 (4.80)	2765 (7.28)	6315 (8.28)	54845 (5.21)	207 (1.55)	115 (1.35)

		N(%)					
Variable	Level	Never smoked in pregnancy	Sometimes smoked in pregnancy	Always smoked in pregnancy	Never used snus in pregnancy	Sometimes used snus in pregnancy	Always used snus in pregnancy
	Middle East	53770 (5.60)	2510 (6.61)	3393 (4.45)	59316 (5.64)	279 (2.09)	78 (0.91)
	Oceania	417 (0.04)	10 (0.03)	21 (0.03)	443 (0.04)	0 (0.00)	5 (0.06)
	Scandinavia	15002 (1.56)	682 (1.79)	1907 (2.50)	17284 (1.64)	187 (1.40)	120 (1.41)
	Swedish	774472 (80.69)	30502 (80.27)	62322 (81.68)	847058 (80.50)	12235 (91.81)	8003 (93.84)
Birth year	1999-2001	206094 (21.47)	7412 (19.51)	17054 (22.35)	226659 (21.54)	2275 (17.07)	1626 (19.07)
	2002-2004	234118 (24.39)	11248 (29.60)	21848 (28.63)	261263 (24.83)	3619 (27.16)	2332 (27.35)
	2005-2007	242212 (25.24)	10416 (27.41)	18140 (23.77)	264888 (25.17)	4029 (30.23)	1851 (21.70)
	2008-2010	277361 (28.90)	8923 (23.48)	19262 (25.24)	299423 (28.46)	3404 (25.54)	2719 (31.88)
Any maternal smoking in pregnancy	No	959785 (100.00)	20164 (53.06)	0 (0.00)	960336 (91.27)	11917 (89.42)	7696 (90.24)
	Yes	0 (0.00)	17835 (46.94)	76304 (100.00)	91897 (8.73)	1410 (10.58)	832 (9.76)
Any maternal snus use in pregnancy	No	947238 (98.69)	36795 (96.83)	75365 (98.77)	1052233 (100.00)	7165 (53.76)	0 (0.00)
	Yes	12547 (1.31)	1204 (3.17)	939 (1.23)	0 (0.00)	6162 (46.24)	8528 (100.00)

Table D.2.1-2: Descriptives separated by timing of smoking exposure

		N(%)					
Variable	Level	Non-smoker before pregnancy	Smoker before pregnancy	Non-smoker at first prenatal visit	Smoker at first prenatal visit	Non-smoker at 32 weeks pregnancy	Smoker at 32 weeks pregnancy
Total		838763	190333	982536	85952	999168	59003
Intellectual disability	No	833835 (99.41)	188685 (99.13)	976574 (99.39)	84956 (98.84)	993072 (99.39)	58374 (98.93)
	Yes	4928 (0.59)	1648 (0.87)	5962 (0.61)	996 (1.16)	6096 (0.61)	629 (1.07)
Small for gestational age	No	820424 (97.81)	183904 (96.62)	960797 (97.79)	81939 (95.33)	976992 (97.78)	56256 (95.34)
	Yes	15945 (1.90)	5811 (3.05)	18765 (1.91)	3717 (4.32)	19130 (1.91)	2597 (4.40)
Sex	Female	407647 (48.60)	92382 (48.54)	477699 (48.62)	41481 (48.26)	485596 (48.60)	28658 (48.57)
	Male	431116 (51.40)	97951 (51.46)	504837 (51.38)	44471 (51.74)	513572 (51.40)	30345 (51.43)
Parity	1	358221 (42.71)	100193 (52.64)	438512 (44.63)	36406 (42.36)	447758 (44.81)	23352 (39.58)
	2	322316 (38.43)	53835 (28.28)	364783 (37.13)	26385 (30.70)	369539 (36.98)	18842 (31.93)
	3 or more	158226 (18.86)	36305 (19.07)	179241 (18.24)	23161 (26.95)	181871 (18.20)	16809 (28.49)
Highest parental education	High School	28235 (3.37)	19611 (10.30)	36827 (3.75)	12830 (14.93)	38101 (3.81)	9312 (15.78)
	Gymnasium	303311 (36.16)	120070 (63.08)	383353 (39.02)	58695 (68.29)	393293 (39.36)	40359 (68.40)
	University	507217 (60.47)	50652 (26.61)	562356 (57.24)	14427 (16.78)	567774 (56.82)	9332 (15.82)
Adjusted family income	1	85455 (10.19)	30183 (15.86)	103670 (10.55)	16936 (19.70)	105757 (10.58)	12097 (20.50)
	2	162951 (19.43)	54845 (28.82)	195638 (19.91)	30396 (35.36)	199880 (20.00)	21734 (36.84)
	3	187459 (22.35)	45527 (23.92)	221863 (22.58)	20060 (23.34)	225983 (22.62)	13677 (23.18)
	4	199104 (23.74)	36978 (19.43)	232428 (23.66)	12733 (14.81)	235930 (23.61)	8113 (13.75)
	5	203794 (24.30)	22800 (11.98)	228937 (23.30)	5827 (6.78)	231618 (23.18)	3382 (5.73)
Maternal anxiety diagnosis	No	817032 (97.41)	178512 (93.79)	955405 (97.24)	79228 (92.18)	971212 (97.20)	54002 (91.52)
	Yes	21731 (2.59)	11821 (6.21)	27131 (2.76)	6724 (7.82)	27956 (2.80)	5001 (8.48)
Maternal depression diagnosis	No	823744 (98.21)	182294 (95.78)	963749 (98.09)	81507 (94.83)	979790 (98.06)	55721 (94.44)
	Yes	15019 (1.79)	8039 (4.22)	18787 (1.91)	4445 (5.17)	19378 (1.94)	3282 (5.56)
	No	836333 (99.71)	188998 (99.30)	979516 (99.69)	85128 (99.04)	996034 (99.69)	58432 (99.03)

		N(%)					
Variable	Level	Non-smoker before pregnancy	Smoker before pregnancy	Non-smoker at first prenatal visit	Smoker at first prenatal visit	Non-smoker at 32 weeks pregnancy	Smoker at 32 weeks pregnancy
Maternal psychosis diagnosis	Yes	2430 (0.29)	1335 (0.70)	3020 (0.31)	824 (0.96)	3134 (0.31)	571 (0.97)
Maternal addiction diagnosis	No	831165 (99.09)	181236 (95.22)	971255 (98.85)	80365 (93.50)	987208 (98.80)	55048 (93.30)
	Yes	7598 (0.91)	9097 (4.78)	11281 (1.15)	5587 (6.50)	11960 (1.20)	3955 (6.70)
Any maternal psychiatric diagnosis	No	801630 (95.57)	168338 (88.44)	935312 (95.19)	73416 (85.42)	950336 (95.11)	49935 (84.63)
	Yes	37133 (4.43)	21995 (11.56)	47224 (4.81)	12536 (14.58)	48832 (4.89)	9068 (15.37)
Any paternal psychiatric diagnosis	No	811324 (96.73)	175432 (92.17)	948196 (96.50)	77339 (89.98)	963744 (96.45)	52746 (89.40)
	Yes	27439 (3.27)	14901 (7.83)	34340 (3.50)	8613 (10.02)	35424 (3.55)	6257 (10.60)
Any maternal neurodevelopmental diagnosis	No	837836 (99.89)	189249 (99.43)	981220 (99.87)	85253 (99.19)	997769 (99.86)	58475 (99.11)
	Yes	927 (0.11)	1084 (0.57)	1316 (0.13)	699 (0.81)	1399 (0.14)	528 (0.89)
Any paternal neurodevelopmental diagnosis	No	837662 (99.87)	189191 (99.40)	980982 (99.84)	85254 (99.19)	997542 (99.84)	58467 (99.09)
	Yes	1101 (0.13)	1142 (0.60)	1554 (0.16)	698 (0.81)	1626 (0.16)	536 (0.91)
Maternal country of origin	Africa	27036 (3.22)	1972 (1.04)	29325 (2.98)	777 (0.90)	29563 (2.96)	504 (0.85)
	Americas	8345 (0.99)	2412 (1.27)	10546 (1.07)	584 (0.68)	10729 (1.07)	359 (0.61)
	Asia	27854 (3.32)	3568 (1.87)	31034 (3.16)	1283 (1.49)	31364 (3.14)	855 (1.45)
	Europe	38783 (4.62)	14536 (7.64)	48081 (4.89)	6810 (7.92)	49041 (4.91)	4940 (8.37)
	Middle East	49000 (5.84)	8355 (4.39)	55563 (5.66)	3799 (4.42)	56021 (5.61)	2776 (4.70)
	Oceania	368 (0.04)	61 (0.03)	420 (0.04)	24 (0.03)	429 (0.04)	15 (0.03)
	Scandinavia	12953 (1.54)	3763 (1.98)	15413 (1.57)	2076 (2.42)	15732 (1.57)	1347 (2.28)
	Swedish	674424 (80.41)	155666 (81.79)	792154 (80.62)	70599 (82.14)	806289 (80.70)	48207 (81.70)
Birth year	1999-2001	151182 (18.02)	39577 (20.79)	209698 (21.34)	19511 (22.70)	213522 (21.37)	11018 (18.67)
	2002-2004	211205 (25.18)	53511 (28.11)	239870 (24.41)	25400 (29.55)	244159 (24.44)	17628 (29.88)
	2005-2007	220992 (26.35)	48280 (25.37)	248702 (25.31)	20622 (23.99)	253077 (25.33)	15097 (25.59)
	2008-2010	255384 (30.45)	48965 (25.73)	284266 (28.93)	20419 (23.76)	288410 (28.87)	15260 (25.86)

		N(%)					
Variable	Level	Non-smoker before pregnancy	Smoker before pregnancy	Non-smoker at first prenatal visit	Smoker at first prenatal visit	Non-smoker at 32 weeks pregnancy	Smoker at 32 weeks pregnancy
Any maternal smoking in pregnancy	No	835467 (99.61)	100033 (52.56)	974797 (99.21)	0 (0.00)	979949 (98.08)	0 (0.00)
	Yes	3296 (0.39)	90300 (47.44)	7739 (0.79)	85952 (100.00)	19219 (1.92)	59003 (100.00)
Any maternal snus use in pregnancy	No	826823 (98.58)	187740 (98.64)	969089 (98.63)	84883 (98.76)	985508 (98.63)	58190 (98.62)
	Yes	11940 (1.42)	2593 (1.36)	13447 (1.37)	1069 (1.24)	13660 (1.37)	813 (1.38)

Table D.2.1-3: Descriptives separated by timing of snus exposure

		N(%)					
Variable	Level	Non-snus before pregnancy	Snus before pregnancy	Non-snus at first prenatal visit	Snus at first prenatal visit	Non-snus at 32 weeks	Snus at 32 weeks
Total		997689	27741	1055180	12858	1065864	4954
Intellectual disability	No	991345 (99.36)	27567 (99.37)	1048343 (99.35)	12752 (99.18)	1058957 (99.35)	4895 (98.81)
	Yes	6344 (0.64)	174 (0.63)	6837 (0.65)	106 (0.82)	6907 (0.65)	59 (1.19)
Small for gestational age	No	973637 (97.59)	27122 (97.77)	1029772 (97.59)	12538 (97.51)	1040200 (97.59)	4839 (97.68)
	Yes	21140 (2.12)	536 (1.93)	22179 (2.10)	282 (2.19)	22379 (2.10)	99 (2.00)
Sex	Female	484661 (48.58)	13528 (48.77)	512730 (48.59)	6231 (48.46)	517993 (48.60)	2426 (48.97)
	Male	513028 (51.42)	14213 (51.23)	542450 (51.41)	6627 (51.54)	547871 (51.40)	2528 (51.03)
Parity	1	441382 (44.24)	14875 (53.62)	469187 (44.47)	5503 (42.80)	474211 (44.49)	1945 (39.26)
	2	366859 (36.77)	8373 (30.18)	386625 (36.64)	4422 (34.39)	390180 (36.61)	1835 (37.04)
	3 or more	189448 (18.99)	4493 (16.20)	199368 (18.89)	2933 (22.81)	201473 (18.90)	1174 (23.70)
Highest parental education	High School	46680 (4.68)	813 (2.93)	49034 (4.65)	556 (4.32)	49546 (4.65)	247 (4.99)
	Gymnasium	407050 (40.80)	13798 (49.74)	434395 (41.17)	7322 (56.95)	439622 (41.25)	2926 (59.06)
	University	543959 (54.52)	13130 (47.33)	571751 (54.19)	4980 (38.73)	576696 (54.11)	1781 (35.95)
Adjusted family income	1	112936 (11.32)	2023 (7.29)	119285 (11.30)	1223 (9.51)	120493 (11.30)	534 (10.78)
	2	210612 (21.11)	6393 (23.05)	222430 (21.08)	3499 (27.21)	224861 (21.10)	1406 (28.38)
	3	224803 (22.53)	7368 (26.56)	238291 (22.58)	3524 (27.41)	240835 (22.60)	1388 (28.02)
	4	228347 (22.89)	6931 (24.98)	242282 (22.96)	2789 (21.69)	244661 (22.95)	991 (20.00)
	5	220991 (22.15)	5026 (18.12)	232892 (22.07)	1823 (14.18)	235014 (22.05)	635 (12.82)
Maternal anxiety diagnosis	No	965752 (96.80)	26196 (94.43)	1022137 (96.87)	12056 (93.76)	1032377 (96.86)	4619 (93.24)
	Yes	31937 (3.20)	1545 (5.57)	33043 (3.13)	802 (6.24)	33487 (3.14)	335 (6.76)
	No	975872 (97.81)	26539 (95.67)	1032546 (97.85)	12262 (95.36)	1042920 (97.85)	4696 (94.79)

Variable	Level	Non-snus before pregnancy	Snus before pregnancy	Non-snus at first prenatal visit	Snus at first prenatal visit	Non-snus at 32 weeks	Snus at 32 weeks
		N(%)					
Maternal depression diagnosis	Yes	21817 (2.19)	1202 (4.33)	22634 (2.15)	596 (4.64)	22944 (2.15)	258 (5.21)
Maternal psychosis diagnosis	No	994132 (99.64)	27549 (99.31)	1051436 (99.65)	12761 (99.25)	1062079 (99.64)	4897 (98.85)
	Yes	3557 (0.36)	192 (0.69)	3744 (0.35)	97 (0.75)	3785 (0.36)	57 (1.15)
Maternal addiction diagnosis	No	982050 (98.43)	26760 (96.46)	1038874 (98.45)	12313 (95.76)	1049283 (98.44)	4706 (94.99)
	Yes	15639 (1.57)	981 (3.54)	16306 (1.55)	545 (4.24)	16581 (1.56)	248 (5.01)
Any maternal psychiatric diagnosis	No	941549 (94.37)	24916 (89.82)	996884 (94.48)	11421 (88.82)	1006783 (94.46)	4338 (87.57)
	Yes	56140 (5.63)	2825 (10.18)	58296 (5.52)	1437 (11.18)	59081 (5.54)	616 (12.43)
Any paternal psychiatric diagnosis	No	957031 (95.92)	26178 (94.37)	1013058 (96.01)	12057 (93.77)	1023251 (96.00)	4599 (92.83)
	Yes	40658 (4.08)	1563 (5.63)	42122 (3.99)	801 (6.23)	42613 (4.00)	355 (7.17)
Any maternal neurodevelopmental diagnosis	No	995799 (99.81)	27627 (99.59)	1053245 (99.82)	12781 (99.40)	1063891 (99.81)	4915 (99.21)
	Yes	1890 (0.19)	114 (0.41)	1935 (0.18)	77 (0.60)	1973 (0.19)	39 (0.79)
Any paternal neurodevelopmental diagnosis	No	995543 (99.78)	27647 (99.66)	1052970 (99.79)	12814 (99.66)	1063636 (99.79)	4934 (99.60)
	Yes	2146 (0.22)	94 (0.34)	2210 (0.21)	44 (0.34)	2228 (0.21)	20 (0.40)
Maternal country of origin	Africa	28876 (2.89)	137 (0.49)	30009 (2.84)	91 (0.71)	30206 (2.83)	29 (0.59)
	Americas	10590 (1.06)	95 (0.34)	11081 (1.05)	50 (0.39)	11175 (1.05)	19 (0.38)
	Asia	31059 (3.11)	341 (1.23)	32138 (3.05)	178 (1.38)	32345 (3.03)	64 (1.29)
	Europe	52800 (5.29)	296 (1.07)	54716 (5.19)	152 (1.18)	55054 (5.17)	71 (1.43)
	Middle East	57001 (5.71)	223 (0.80)	59202 (5.61)	144 (1.12)	59579 (5.59)	51 (1.03)
	Oceania	415 (0.04)	10 (0.04)	439 (0.04)	5 (0.04)	446 (0.04)	1 (0.02)
	Scandinavia	16258 (1.63)	364 (1.31)	17299 (1.64)	177 (1.38)	17466 (1.64)	78 (1.57)
	Swedish	800690 (80.25)	26275 (94.72)	850296 (80.58)	12061 (93.80)	859593 (80.65)	4641 (93.68)
Birth year	1999-2001	183455 (18.39)	3773 (13.60)	226637 (21.48)	2097 (16.31)	228992 (21.48)	786 (15.87)
	2002-2004	257752 (25.83)	6900 (24.87)	261691 (24.80)	3623 (28.18)	264932 (24.86)	1419 (28.64)

Variable	Level	Non-snus before pregnancy	Snus before pregnancy	Non-snus at first prenatal visit	Snus at first prenatal visit	Non-snus at 32 weeks	Snus at 32 weeks
				N(%)			
	2005-2007	262196 (26.28)	7082 (25.53)	266003 (25.21)	3306 (25.71)	268987 (25.24)	1135 (22.91)
	2008-2010	294286 (29.50)	9986 (36.00)	300849 (28.51)	3832 (29.80)	302953 (28.42)	1614 (32.58)
Any maternal smoking in pregnancy	No	906723 (90.88)	26331 (94.92)	963349 (91.30)	11760 (91.46)	972735 (91.26)	4467 (90.17)
	Yes	90966 (9.12)	1410 (5.08)	91831 (8.70)	1098 (8.54)	93129 (8.74)	487 (9.83)
Any maternal snus use in pregnancy	No	994642 (99.69)	16188 (58.35)	1053486 (99.84)	0 (0.00)	1059398 (99.39)	0 (0.00)
	Yes	3047 (0.31)	11553 (41.65)	1694 (0.16)	12858 (100.00)	6466 (0.61)	4954 (100.00)

Table D.2.1-4: Descriptives of cohort for sensitivity analysis (ii) separated by change in smoking behaviour across first two pregnancies

			N(%)	
Variable	Level	Total	Stopped	Started
Total		663875	8913	4995
Intellectual disability	No	659146 (99.29)	8843 (99.21)	4946 (99.02)
	Yes	4729 (0.71)	70 (0.79)	49 (0.98)
Small for gestational age	No	644096 (97.02)	8528 (95.68)	4839 (96.88)
	Yes	17473 (2.63)	347 (3.89)	134 (2.68)
Sex	Female	322324 (48.55)	4290 (48.13)	2443 (48.91)
	Male	341551 (51.45)	4623 (51.87)	2552 (51.09)
Parity	1	461255 (69.48)	7443 (83.51)	3818 (76.44)
	2	117966 (17.77)	938 (10.52)	750 (15.02)
	3 or more	84654 (12.75)	532 (5.97)	427 (8.55)
Highest parental	High School	32379 (4.88)	857 (9.62)	733 (14.67)
education	Gymnasium	280306 (42.22)	5810 (65.19)	3115 (62.36)
	University	351190 (52.90)	2246 (25.20)	1147 (22.96)
Adjusted family income	1	78696 (11.85)	1411 (15.83)	1230 (24.62)
	2	125083 (18.84)	2064 (23.16)	1225 (24.52)
	3	129914 (19.57)	2016 (22.62)	932 (18.66)
	4	159749 (24.06)	2158 (24.21)	1031 (20.64)
	5	170433 (25.67)	1264 (14.18)	577 (11.55)
Maternal anxiety	No	644529 (97.09)	8569 (96.14)	4841 (96.92)
diagnosis	Yes	19346 (2.91)	344 (3.86)	154 (3.08)
Maternal depression	No	650406 (97.97)	8686 (97.45)	4889 (97.88)
diagnosis	Yes	13469 (2.03)	227 (2.55)	106 (2.12)
Maternal psychosis	No	661469 (99.64)	8883 (99.66)	4976 (99.62)
diagnosis	Yes	2406 (0.36)	30 (0.34)	19 (0.38)
Maternal addiction	No	653174 (98.39)	8547 (95.89)	4847 (97.04)
diagnosis	Yes	10701 (1.61)	366 (4.11)	148 (2.96)
Any maternal psychiatric	No	628798 (94.72)	8175 (91.72)	4657 (93.23)
diagnosis	Yes	35077 (5.28)	738 (8.28)	338 (6.77)
Any paternal psychiatric	No	638056 (96.11)	8350 (93.68)	4725 (94.59)
diagnosis	Yes	25819 (3.89)	563 (6.32)	270 (5.41)
Any maternal	No	662551 (99.80)	8896 (99.81)	4975 (99.60)
neurodevelopmental	Yes	1324 (0.20)	17 (0.19)	20 (0.40)
diagnosis				
Any paternal	No	662436 (99.78)	8885 (99.69)	4978 (99.66)
neurodevelopmental	Yes	1439 (0.22)	28 (0.31)	17 (0.34)
diagnosis				
Maternal country of	Africa	17491 (2.63)	84 (0.94)	85 (1.70)
origin	Americas	7448 (1.12)	50 (0.56)	50 (1.00)
	Asia	22311 (3.36)	137 (1.54)	105 (2.10)

Variable	Level	Total	N(%) Stopped	Started
	Europe	37280 (5.62)	589 (6.61)	445 (8.91)
	Middle East	36707 (5.53)	343 (3.85)	467 (9.35)
	Oceania	282 (0.04)	0 (0.00)	2 (0.04)
	Scandinavia	11224 (1.69)	158 (1.77)	88 (1.76)
	Swedish	531132 (80.00)	7552 (84.73)	3753 (75.14)
Birth year	1999-2001	206662 (31.13)	3037 (34.07)	2354 (47.13)
	2002-2004	163022 (24.56)	3687 (41.37)	1419 (28.41)
	2005-2007	139817 (21.06)	1805 (20.25)	1014 (20.30)
	2008-2010	154374 (23.25)	384 (4.31)	208 .16)

D.2.2 – Missing data

Table D.2.2-1: Descriptives for missing covariate data

Variable	Level	Included, N(%)	Excluded, N(%)	O.R. for exclusion (95% CI)
Total		1132473 (100.00)	824 (100.00)	
Intellectual disability	No	1124940 (99.33)	811 (98.42)	Ref
	Yes	7533 (0.67)	13 (1.58)	2.39 (1.38-4.14)
Small for gestational age	No	1104238 (97.51)	785 (95.27)	Ref
	Yes	24090 (2.13)	37 (4.49)	2.16 (1.55-3.01)
Any maternal smoking in pregnancy	No	980066 (86.54)	701 (85.07)	Ref
	Yes	103099 (9.10)	80 (9.71)	1.08 (0.86-1.37)
	Missing	49308 (4.35)	43 (5.22)	1.22 (0.90-1.66)
Any maternal snus use in pregnancy	No	1060047 (93.60)	768 (93.20)	Ref
	Yes	14972 (1.32)	9 (1.09)	0.83 (0.43-1.60)
	Missing	57454 (5.07)	47 (5.70)	1.13 (0.84-1.52)
Sex	Female	549959 (48.56)	402 (48.79)	Ref
	Male	582514 (51.44)	422 (51.21)	0.99 (0.86-1.14)
Parity	1	503786 (44.49)	393 (47.69)	Ref
	2	413596 (36.52)	183 (22.21)	0.57 (0.48-0.68)
	3 or more	215091 (18.99)	248 (30.10)	1.48 (1.26-1.73)
Highest parental education	High School	53744 (4.75)	19 (2.31)	Ref
	Gymnasium	467625 (41.29)	40 (4.85)	0.24 (0.14-0.42)
	University	611104 (53.96)	72 (8.74)	0.33 (0.20-0.55)
Adjusted family income	1	128677 (11.36)	430 (52.18)	19.75 (14.44-27.03)
	2	238887 (21.09)	161 (19.54)	3.98 (2.85-5.58)
	3	254172 (22.44)	43 (5.22)	Ref
	4	257905 (22.77)	20 (2.43)	0.46 (0.27-0.78)
	5	252832 (22.33)	61 (7.40)	1.43 (0.97-2.11)
Maternal anxiety diagnosis	No	1096769 (96.85)	804 (97.57)	Ref
	Yes	35704 (3.15)	20 (2.43)	0.76 (0.49-1.19)
Maternal depression diagnosis	No	1108057 (97.84)	804 (97.57)	Ref
	Yes	24416 (2.16)	20 (2.43)	1.13 (0.72-1.76)
Maternal psychosis diagnosis	No	1128347 (99.64)	822 (99.76)	Ref
	Yes	4126 (0.36)	2 (0.24)	0.67 (0.17-2.67)
Maternal addiction diagnosis	No	1114428 (98.41)	821 (99.64)	Ref
	Yes	18045 (1.59)	3 (0.36)	0.23 (0.07-0.70)
Any maternal psychiatric diagnosis	No	1069221 (94.41)	788 (95.63)	Ref
	Yes	63252 (5.59)	36 (4.37)	0.77 (0.55-1.08)
Any paternal psychiatric diagnosis	No	1087032 (95.99)	806 (97.82)	Ref
	Yes	45441 (4.01)	18 (2.18)	0.53 (0.33-0.85)
Any maternal neurodevelopmental diagnosis	No	1130342 (99.81)	816 (99.03)	Ref
	Yes	2131 (0.19)	8 (0.97)	5.20 (2.59-10.45)
Any paternal neurodevelopmental diagnosis	No	1130111 (99.79)	822 (99.76)	Ref
	Yes	2362 (0.21)	2 (0.24)	1.16 (0.29-4.67)
Maternal country of origin	Africa	31815 (2.81)	153 (18.57)	84.64 (61.78-115.96)
	Americas	11914 (1.05)	16 (1.94)	23.64 (13.49-41.41)
	Asia	33954 (3.00)	71 (8.62)	36.80 (25.73-52.64)

Variable	Level	Included, N(%)	Excluded, N(%)	O.R. for exclusion (95% CI)
	Europe	57821 (5.11)	137 (16.63)	41.70 (30.30-57.39)
	Middle East	62334 (5.50)	93 (11.29)	26.26 (18.70-36.87)
	Oceania	491 (0.04)	4 (0.49)	143.38 (51.66-397.96)
	Scandinavia	18925 (1.67)	59 (7.16)	54.87 (37.78-79.68)
	Swedish	915219 (80.82)	52 (6.31)	Ref
Birth year	1999-2001	250416 (22.11)	85 (10.32)	Ref
	2002-2004	276150 (24.38)	128 (15.53)	1.37 (1.04-1.80)
	2005-2007	293076 (25.88)	214 (25.97)	2.15 (1.67-2.77)
	2008-2010	312831 (27.62)	397 (48.18)	3.74 (2.96-4.73)

Table D.2.2-2: Descriptives for missing exposure data

Variable	Level	Included, N(%)	Excluded, N(%)	O.R. for exclusion (95% CI)
Total		1074088 (100.00)	58385 (100.00)	
Intellectual disability	No	1067096 (99.35)	57844 (99.07)	Ref
	Yes	6992 (0.65)	541 (0.93)	1.43 (1.31-1.56)
Small for gestational age	No	1048198 (97.59)	56040 (95.98)	Ref
	Yes	22574 (2.10)	1516 (2.60)	1.26 (1.19-1.32)
Any maternal smoking in pregnancy	No	979949 (91.24)	117 (0.20)	Ref
	Yes	94139 (8.76)	8960 (15.35)	797.18 (664.20-956.79)
	Missing	0 (0.00)	49308 (84.45)	
Any maternal snus use in pregnancy	No	1059398 (98.63)	649 (1.11)	Ref
	Yes	14690 (1.37)	282 (0.48)	31.34 (27.22-36.07)
	Missing	0 (0.00)	57454 (98.41)	
Sex	Female	521928 (48.59)	28031 (48.01)	Ref
	Male	552160 (51.41)	30354 (51.99)	1.02 (1.01-1.04)
Parity	1	477543 (44.46)	26243 (44.95)	Ref
	2	393141 (36.60)	20455 (35.03)	0.95 (0.93-0.96)
	3 or more	203404 (18.94)	11687 (20.02)	1.05 (1.02-1.07)
Highest parental education	High School	49943 (4.65)	3801 (6.51)	Ref
	Gymnasium	444362 (41.37)	23263 (39.84)	0.69 (0.66-0.71)
	University	579783 (53.98)	31321 (53.65)	0.71 (0.69-0.74)
Adjusted family income	1	121353 (11.30)	7324 (12.54)	1.32 (1.29-1.37)
	2	227142 (21.15)	11745 (20.12)	1.13 (1.11-1.17)
	3	243097 (22.63)	11075 (18.97)	Ref
	4	246339 (22.93)	11566 (19.81)	1.03 (1.00-1.06)
	5	236157 (21.99)	16675 (28.56)	1.55 (1.51-1.59)
Maternal anxiety diagnosis	No	1040061 (96.83)	56708 (97.13)	Ref
	Yes	34027 (3.17)	1677 (2.87)	0.90 (0.86-0.95)
Maternal depression diagnosis	No	1050740 (97.83)	57317 (98.17)	Ref
	Yes	23348 (2.17)	1068 (1.83)	0.84 (0.79-0.89)
Maternal psychosis diagnosis	No	1070223 (99.64)	58124 (99.55)	Ref
	Yes	3865 (0.36)	261 (0.45)	1.24 (1.10-1.41)
Maternal addiction diagnosis	No	1057113 (98.42)	57315 (98.17)	Ref
	Yes	16975 (1.58)	1070 (1.83)	1.16 (1.09-1.24)
Any maternal psychiatric diagnosis	No	1014020 (94.41)	55201 (94.55)	Ref
	Yes	60068 (5.59)	3184 (5.45)	0.97 (0.94-1.01)
Any paternal psychiatric diagnosis	No	1030905 (95.98)	56127 (96.13)	Ref
	Yes	43183 (4.02)	2258 (3.87)	0.96 (0.92-1.00)
Any maternal neurodevelopmental diagnosis	No	1072057 (99.81)	58285 (99.83)	Ref
	Yes	2031 (0.19)	100 (0.17)	0.91 (0.74-1.11)
Any paternal neurodevelopmental diagnosis	No	1071828 (99.79)	58283 (99.83)	Ref
	Yes	2260 (0.21)	102 (0.17)	0.83 (0.68-1.01)
Maternal country of origin	Africa	30254 (2.82)	1561 (2.67)	0.93 (0.89-0.98)
	Americas	11207 (1.04)	707 (1.21)	1.14 (1.06-1.23)
	Asia	32452 (3.02)	1502 (2.57)	0.84 (0.79-0.88)
	Europe	55167 (5.14)	2654 (4.55)	0.87 (0.84-0.91)
	Middle East	59673 (5.56)	2661 (4.56)	0.81 (0.78-0.84)

Variable	Level	Included, N(%)	Excluded, N(%)	O.R. for exclusion (95% CI)
	Oceania	448 (0.04)	43 (0.07)	1.74 (1.27-2.38)
	Scandinavia	17591 (1.64)	1334 (2.28)	1.37 (1.30-1.45)
	Swedish	867296 (80.75)	47923 (82.08)	Ref
Birth year	1999-2001	230560 (21.47)	19856 (34.01)	Ref
	2002-2004	267214 (24.88)	8936 (15.31)	0.39 (0.38-0.40)
	2005-2007	270768 (25.21)	22308 (38.21)	0.96 (0.94-0.98)
	2008-2010	305546 (28.45)	7285 (12.48)	0.28 (0.27-0.28)

D.3 – Reanalysis of snus use in pregnancy-SGA association without exclusions that may lead to bias

Table D.3-1: Repeat of primary analysis of snus use in pregnancy at any time and offspring risk of being born SGA using data that did not exclude those with metabolic, genetic and chromosomal abnormalities.

		Snus use in pregnancy	
Model	Coefficient	O.R. [a]	95% CI
1 - Conventional unadjusted	Population averaged	1.03	(0.92-1.15)
2 - Conventional adjusted [b]	Population averaged	1.06	(0.95-1.18)
3 - Within-between unadjusted	Within-family	1.02	(0.82-1.28)
	Between-family	1.01	(0.78-1.31)
4 - Within-between adjusted [b]	Within-family	1.08	(0.85-1.37)
	Between-family	0.97	(0.74-1.27)

[a] Estimates produced using a total sample size of 1,083,663 individuals from 711,737 families including 23,432 cases of SGA.

[b] Model adjusted for year of birth, sex, parity, highest parental education, income, parental psychiatric history, maternal country of origin and maternal age at birth.

Table D.3-2: Repeat of secondary analysis of the effect of timing of snus use in pregnancy and offspring risk of being born SGA using data that did not exclude those with metabolic, genetic and chromosomal abnormalities.

		Snus use in pregnancy	
Model	Coefficient	O.R.	95% CI
1 - Conventional unadjusted [a] (population-averaged estimates)	Non-user	1.00	
	User before pregnancy only	0.82	(0.73-0.92)
	Quit during pregnancy	0.92	(0.76-1.11)
	Used late into pregnancy	0.99	(0.79-1.24)
2 - Conventional adjusted [a, b] (population-averaged estimates)	Non-user	1.00	
	User before pregnancy only	0.79	(0.70-0.89)
	Quit during pregnancy	0.93	(0.76-1.12)
	Used late into pregnancy	1.05	(0.84-1.33)
3 - Unadjusted conditional logistic [c] (within-family estimates)	Non-user	1.00	
	User before pregnancy only	1.11	(0.83-1.49)
	Quit during pregnancy	1.06	(0.69-1.63)
	Used late into pregnancy	1.14	(0.59-2.20)
4 - Adjusted conditional logistic [b, c] (within-family estimates)	Non-user	1.00	
	User before pregnancy only	0.87	(0.63-1.20)
	Quit during pregnancy	1.20	(0.75-1.93)
	Used late into pregnancy	1.61	(0.78-3.25)

[a] Estimates produced using a total sample size of 1,077,716 individuals from 709,097 families including 23,282 cases of SGA.

[b] Model adjusted for year of birth, sex, parity, highest parental education, income, parental psychiatric history, maternal country of origin and maternal age at birth.

[c] Estimates produced using a total sample size of 24,123 individuals including 10,923 cases of SGA.

Appendix E – Supplementary material to Chapter 7

E.1 – Methods

E1.1 – Multiple-sourced indicator variable for ID

Table E.1.1-1: Search terms and number of hits for free text fields.

Search term	Number of Hits in ALSPAC
Intellectual Disability	0
Developmental Disabilities	0
Intellectual disab	0
developmental disab	0
learning disab	15
mental retard	<5
mental handicap	0
handicap	5
intellectual	5
retard	7
learning disability	6
learning disabled	<5
learning difficulties	137
learning difficulty	41
difficulty learning	<5
mental disability	0
mentally disabled	<5
mentally retarded	0
mental retardation	<5
low IQ	<5
development delay	20
developmental delay	29

Table E.1.1-2: Read codes used to indicate intellectual disability in GP records

Read code	Description
13Z4E	Learning difficulties
6664	Mental handicap problem
69DB	Learning disability health examination
8Ce6	Preferred place of care - learning disability unit
8H4f	Referral to learning disabilities psychiatrist
8HHP	Referral to learning disability team
8Hg2	Discharge from learning disability team
918e	On learning disability register
9HB	Learning disabilities administration status
9HB0	Learning disabilities health action plan declined
9HB1	Learning disabilities health action plan offered
9HB2	Learning disabilities health action plan reviewed
9HB3	Learning disabilities health assessment
9HB4	Learning disabilities health action plan completed
9HB5	Learning disabilities annual health assessment
9HB6	Learning disabilities annual health assessment declined
9HB7	Did not attend learning disabilities annual health check
9N0y	Seen in learning disabilities clinic
9hL	Exception reporting: learning disability quality indicators
9mA	Learning disability annual health check invitation
9mA0	Learning disability annual health check verbal invitation
9mA1	Learning disability annual health check telephone invitation
9mA2	Learning disability annual health check letter invitation
9mA20	Learning disability annual health check invtation 1st letter
9mA21	Learning disability annual health check invtation 2nd letter
9mA22	Learning disability annual health check invtation 3rd letter
E3	Mental retardation
E30	Mild mental retardation, IQ in range 50-70
E31	Other specified mental retardation
E310	Moderate mental retardation, IQ in range 35-49
E311	Severe mental retardation, IQ in range 20-34
E312	Profound mental retardation with IQ less than 20
E31z	Other specified mental retardation NOS
E3y	Other specified mental retardation
E3z	Mental retardation NOS
Eu7	[X]Mental retardation
Eu70	[X]Mild mental retardation
Eu700	[X]Mld mental retard with statement no or min impairm behav
Eu701	[X]Mld mental retard sig impairment behav req attent/treatmt
Eu70y	[X]Mild mental retardation, other impairments of behaviour
Eu70z	[X]Mild mental retardation without mention impairment behav
Eu71	[X]Moderate mental retardation
Eu710	[X]Mod mental retard with statement no or min impairm behav
Eu711	[X]Mod mental retard sig impairment behav req attent/treatmt
Eu71y	[X]Mod retard oth behav impair
Eu71z	[X]Mod mental retardation without mention impairment behav
Eu72	[X]Severe mental retardation
Eu720	[X]Sev mental retard with statement no or min impairm behav
Eu721	[X]Sev mental retard sig impairment behav req attent/treatmt
Eu72y	[X]Severe mental retardation, other impairments of behaviour
Eu72z	[X]Sev mental retardation without mention impairment behav
Eu73	[X]Profound mental retardation
Eu731	[X]Profound ment retard sig impairmnt behav req attent/treat
Eu73y	[X]Profound mental retardation, other impairments of behavr
Eu73z	[X]Prfnd mental retardation without mention impairment behav
Eu7y	[X]Other mental retardation
Eu7y0	[X]Oth mental retard with statement no or min impairm behav
Eu7y1	[X]Oth mental retard sig impairment behav req attent/treatmt
Eu7yy	[X]Other mental retardation, other impairments of behaviour
Eu7yz	[X]Other mental retardation without mention impairment behav
Eu7z	[X]Unspecified mental retardation
Eu7z0	[X]Unsp mental retard with statement no or min impairm behav
Eu7z1	[X]Unsp mentl retard sig impairment behav req attent/treatmt
Eu7zy	[X]Unspecified mental retardatn, other impairments of behav
Eu7zz	[X]Unsp mental retardation without mention impairment behav

Read code	Description
Eu814	[X]Moderate learning disability
Eu815	[X]Severe learning disability
Eu816	[X]Mild learning disability
Eu817	[X]Profound learning disability
Eu81z	[X]Learning disorder NOS
Eu841	[X]Mental retardation with autistic features
Eu844	[X]Overactive disorder assoc mental retard/stereotype movts
R034y	[D]Global retardation
Z7CBE	Intellectual functioning disability
Z7CD2	Learning difficulties
ZL1B5	Under care of psychiatrist for mental handicap
ZL5B5	Referral to psychiatrist for mental handicap
ZL9D5	Seen by psychiatrist for mental handicap
ZLD2f	Discharge by psychiatrist for mental handicap
ZLE94	Discharge from mental handicap psychiatry service
ZS34	Learning disability
ZV623	[V]Educational handicap
1286	FH: Mental retardation
94Z9	Preferred place of death: learning disability unit
9hL0	Excepted from learning disability quality indicators: informed dissent
9hL1	Excepted from learning disability quality indicators: patient unsuitable
Eu730	[X]Profound mental retardation with the statement of no, or minimal, impairment of behaviour
PKyG	Mental retardation, congenital heart disease, blepharophimosis, blepharoptosis and hypoplastic teeth
R034E	[D]Developmental delay
Eu81z-1	[x]learning disability nos
E30-1	educationally subnormal
6664	Mental handicap problem

Table E.1.1-3: International classification of disease diagnosis codes for intellectual disability obtained from hospital episode statistics data.

Diagnosis classification system	ICD code	Description
ICD-9	317	Mild intellectual disabilities
	318.0	Moderate intellectual disabilities
	318.1	Severe intellectual disabilities
	318.2	Profound intellectual disabilities
	319	Unspecified intellectual disabilities
ICD-10	F70	Mild intellectual disabilities
	F71	Moderate intellectual disabilities
	F72	Severe intellectual disabilities
	F73	Profound intellectual disabilities
	F78	Other intellectual disabilities
	F79	Unspecified intellectual disabilities

E1.2 – Exclusion criteria

Table E.1.2-1: Codes (read and ICD) used to identify genetic, chromosomal and metabolic abnormalities.

Source (code type)	Code	Description
GP records (read code)	1JB0	Suspected Downs syndrome
	677C4	Carrier of fragile X gene mutation
	C301	Phenylketonuria
	F1y0	Fragile X associated tremor ataxia syndrome
	PJ0	Down's syndrome - trisomy 21
	PJ02	Partial trisomy 21 in Down's syndrome
	PJ0z	Down's syndrome NOS
	PJ2	Edward's syndrome - trisomy 18
	PJ22	Partial trisomy 18 in Edward's syndrome
	PJ2z	Edward's syndrome NOS
	PJ31	Cri-du-chat syndrome
	PJ534	Individual with autosomal fragile site
	PJyy2	Fragile X chromosome
	PJyy4	Fragile X syndrome
	PK5	Tuberous sclerosis
	PK61	Sturge-Weber syndrome
	PKy0	Prader-Willi syndrome
	PKy4	William syndrome
	PKy60	Cornelia de Lange syndrome
	PKy80	Noonan's syndrome
	PKy93	Prader - Willi syndrome
	PKyz5	Angelman syndrome
	PKyz7	Angelman syndrome
	ZC2C6	Dietary advice for phenylketonuria
	PJ00	Trisomy 21, meiotic nondisjunction
	PJ01	Trisomy 21, mosaicism
	PJ20	Trisomy 18, meiotic nondisjunction
	PJ21	Trisomy 18, mosaicism
	PJ0-2	trisomy 21
	PJ50w	Whole chromosome trisomy, meitotic nondisjunction
	PJ11.	Trisomy 13, mosaicism
	PJ507	Other trisomy C syndromes
	PJ1z.	Patau's syndrome NOS
	PKy03	Weaver syndrome = Sotos syndrome
	PJ506	Trisomy 12
	PJ636	Turner's phenotype, ring chromosome karyotype
	PJ334	Jacobsen syndrome
	PJ12.	Trisomy 13, translocation
	PJ333	Smith-Magenis syndrome
	PJ504	Trisomy 10
	PJ523	Triploidy
	PKy95	Biemond's syndrome
	PJ524	Polyploidy
	PKyz0	Ullrich - Feichtiger syndrome, chimaera
	PJ50x	Whole chromosome trisomy, mitotic nondisjunction
	PJ50z	Whole chromosome trisomy syndrome NOS
	PJ10.	Trisomy 13, meiotic nondisjunction
	PJ332	Deletion of short arm of chromosome 18
	PJ30.	Deletion of long arm of chromosome 21
	PJ501	Trisomy 7
	PJ71.	Klinefelter's syndrome, male with more than two X chromosomes
	PJ37.	Whole chromosome monosomy, mosaicism
	PJ505	Trisomy 11
	PJ1..	Patau's syndrome - trisomy 13
	PJ508	Trisomy 22
	PJ502	Trisomy 8
	PJ5y.	Pseudotrisomy 18
	PJ32.	Deletion of short arm of chromosome 4
	PJ50.	Whole chromosome trisomy syndromes
	PJ330	Deletion of long arm of chromosome 13
	PJ3y0	Shprintzen syndrome

Source (code type)	Code	Description
	PJ510	Major partial trisomy
	PJ331	Deletion of long arm of chromosome 18
	PJ500	Trisomy 6
	PJ36.	Whole chromosome monosomy, meiotic nondisjunction
	PKy92	Menke's syndrome
	PJ503	Trisomy 9
	PJ73.	Klinefelter's syndrome, XXYY
	PJ50y	Other specified whole chromosome trisomy syndrome
	Eu842	[X]Rett's syndrome
	PJ9..	Mowat-Wilson syndrome
	PKyz.	Cockayne's syndrome
	PKy94	Zellweger's syndrome
HES data (ICD-9 codes)	270.0	Disturbances of amino-acid transport
	270.1	Phenylketonuria [PKU]
	270.2	Other disturbances of aromatic amino-acid metabolism
	270.3	Disturbances of branched-chain amino-acid metabolism
	270.4	Disturbances of sulphur-bearing amino-acid metabolism
	270.5	Disturbances of histidine metabolism
	270.6	Disorders of urea cycle metabolism
	270.7	Other disturbances of straight-chain amino-acid metabolism
	270.8	Other specified disorders of amino-acid metabolism
	270.9	Unspecified disorder of amino-acid metabolism
	271.8	Other specified disorders of carbohydrate transport and metabolism
	272.8	Other disorders of lipoid metabolism
	277.81	Primary carnitine deficiency
	277.82	Carnitine deficiency due to inborn errors of metabolism
	277.83	Iatrogenic carnitine deficiency
	277.84	Other secondary carnitine deficiency
	277.85	Disorders of fatty acid oxidation
	277.86	Peroxisomal disorders
	277.89	Other specified disorders of metabolism
	279.11	Digeorge's syndrome
	330.8	Other specified cerebral degenerations in childhood
	751.60	Unspecified anomaly of gallbladder, bile ducts, and liver
	751.69	Other anomalies of gallbladder, bile ducts, and liver
	758.0	Down's syndrome
	758.1	Patau's syndrome
	758.2	Edwards' syndrome
	758.31	Cri-du-chat syndrome
	758.32	Velo-cardio-facial syndrome
	758.33	Other microdeletions
	758.39	Other autosomal deletions
	758.4	Balanced autosomal translocation in normal individual
	758.5	Other conditions due to autosomal anomalies
	758.6	Gonadal dysgenesis
	758.7	Klinefelter's syndrome
	758.81	Other conditions due to sex chromosome anomalies
	758.9	Conditions due to anomaly of unspecified chromosome
	759.5	Tuberous sclerosis
	759.81	Prader-Willi syndrome
	759.83	Fragile X syndrome
	759.89	Other specified congenital anomalies
HES data (ICD-10 codes)	F84.2	Rett's syndrome
	Q89.8	Williams syndrome
	Q87.1	Prader-Willi syndrome
	Q87.2	Rubinstein-Taybi syndrome
	Q44.7	Alagille syndrome
	D82.1	DiGeorge syndrome
	Q85.0	Neurofibromatosis (non-malignant)
	Q85.1	Tuberous sclerosis
	Q90	Down syndrome
	Q91	Trisomy 18 and Trisomy 13
	Q92	Other trisomies and partial trisomies of the autosomes, not elsewhere classified

Source (code type)	Code	Description
	Q93	Monosomies and deletions from the autosomes, not elsewhere classified
	Q95	Balanced rearrangements and structural markers, not elsewhere classified
	Q96	Turner's syndrome
	Q97	Other sex chromosome abnormalities, female phenotype, not elsewhere classified
	Q98	Other sex chromosome abnormalities, male phenotype, not elsewhere classified
	Q99	Other chromosomal abnormalities, not elsewhere specified
	E70	Disorders of aromatic amino-acid metabolism
	E71	Disorders of branched-chain amino-acid metabolism and fatty-acid metabolism
	E72	Other disorders of amino-acid metabolism

Personal Reflections

At the start of this PhD I considered causal inference to be this magical set of methods that revealed the true value of an unknown estimate. All other researchers must be kicking themselves for not having put the time in to try to figure out how to use these methods. Slowly, over 4 years I came to the realisation that these methods are simply imperfect ways of reducing the impacts of the different forms of bias that can hide the truth behind associations. Becoming aware of this was humbling and helped me become a bit less arrogant about the quality of the work I was producing and the comparison to methods used by others in the field. This also opened up a whole new avenue of research that I did not think I would enjoy when I started.

If the first page of a google search is to be believed, on February 5th 1675, Isaac Newton wrote a letter to Robert Hooke where he provided the now famous (and perhaps overused) metaphor for the progression of knowledge: *"If I have seen further it is by standing on the shoulders of Giants"*. As an epidemiologist and statistician, some of the most interesting work has been assessing the methods we are using to assess causality and trying to identify under what scenarios do they show what we think they are showing. Performing simulation studies has not felt like standing on the shoulder of giants, though the work of the researchers who developed these methods is genuinely quite amazing. Instead it has felt more like standing at the top of a large human pyramid, throwing large stones to see if the foundations on which you stand on are really all that sturdy, fully aware that you may have to jump onto another nearby pyramid when you, or a peer next to you, lands a well-placed throw. This is perhaps too generous an analogy given that none of the work I've undertaken so far is likely to be paradigm shifting. It does, however, speak to the concept of testing ideas to find where they fall apart and the sometimes hostile nature with which scientists do this (academic twitter has truly been eye opening with this regard). Going forward I would like to be one of the researchers that challenges others' ideas with tact and grace though it has been fun developing the accuracy of my throw and deciding which nearby pyramid would be best to jump to.

In my continuing development as a researcher I need to work on selecting the most relevant information rather than providing everything in the hope that it provides a complete picture. Too much information can sometimes be just as confusing as restricted information (Bradbury's dystopia in Fahrenheit 451 is not necessarily better than Orwell's dystopia in 1984). The length of the appendices (nearly 70 pages) in this thesis is a testament to the fact that this is something I still need to work on.

This thesis was not written in a linear fashion: Chapter 2 was written after Chapters 5 and 6. As a result, some of the decisions made in empirical chapters were based on an incomplete understanding of the methodological considerations, but could not be changed due to later data restrictions. In the perfect world I would have fully understood all of the methods before I implemented them. The reality of implementing anything is that understanding and mastery comes with time. You walk before you can run. You cycle the bike before realising you can go faster by strapping your feet to the pedals. I've been told many times that the purpose of the PhD is not to write the perfect thesis, but to display the development of understanding. I've learnt a lot over 4 years and have been very lucky to have had the opportunity to spend such a large portion of time developing understanding of an area that I find interesting. I look forward to a career where I can continually learn and never feel comfortable in the knowledge I already have.

It has been an interesting experience completing this PhD during the course of the COVID-19 global pandemic. What has been nice is that (most) people now know what an epidemiologist is. I hope that fewer people ask me if it means that I work with skin (maybe a reasonable mistake). I thought that training in this area would mean that I would have a solid understanding of all the data and studies and questions that people have produced but I instead feel none the wiser than the average person. This may be showing the difference in working in non-communicable disease epidemiology versus working in infectious disease epidemiology. A number of experts in fields other than epidemiology produced work that was criticised heavily for questionable decisions due to not fully understanding the field and therefore drawing conclusions from their data that were potentially harmful. Many were criticised for wandering out of their field (see for example https://www.theguardian.com/world/2020/may/31/covid-19-expert-karl-friston-germany-may-have-more-immunological-dark-matter and the response from GDS https://twitter.com/mendel_random/status/1267075832897507328). I don't disagree with this criticism. Thinking that you can better another field by stumbling in and having a go dismisses the time experts within that field have spent grappling with the minutiae that needs to be understood in order to provide usable data and interpretation. However, one of the best bits of being an epidemiologist is that it is a crossing point for so many different fields (mathematics, statistics, biology, medicine, chemistry, psychology, economics, geography, history and likely several more). As a result there is ample opportunity to work with many people who have training in a different field to you, meaning that each new collaboration provides an opportunity to develop understanding of an area of science you had previously not thought much about. In essence my experience of the pandemic has reinforced the idea that it's good to explore areas you know nothing about, but to be aware of your limitations and collaborate with people that can teach you more.