# Causal Inference Approaches to the Study of Prenatal Exposures

Elizabeth Diemer

# Causal Inference Approaches to the Study of Prenatal Exposures

Causal inferentie methoden in onderzoek naar prenatale risicofactoren

## Thesis

to obtain the degree of Doctor from the

Erasmus University Rotterdam

by the command of the

Rector Magnificus

Prof. Dr. F.A. van der Duijn Schouten

and in accordance with the decision of the Doctorate Board

The public defense shall be held on

Tuesday, June 8th, 2021 at 3:30pm

by

Elizabeth Wicker Diemer

born in San Jose, California, United States of America

**Erasmus University Rotterdam**

**Doctoral Committee**

**Manuscripts that form the basis of this thesis**

**Diemer EW**, Tuhkahnen J, Sammallahti S, Heinonen K, Neumann A, Robinson SL, Suderman M, Jin J, Page C, Fore R, Rifas-Shiman SL, Oken E, Perron P, Bouchard L, Hivert MF, Räikköne K, Lahti J, Yeung EH, Guan W, Mumford SL, Magnus MC, Håberg S, Nystad W, Parr CL, London S, Felix JF, Tiemeier H. (in progress). Epigenome-wide meta-analysis of prenatal Vitamin D insufficiency and cord blood DNA methylation. (Chapter 2)

**Diemer EW**, Labrecque JA, Neumann A, Tiemeier H, Swanson SA (2021). Mendelian randomisation approaches to the study of prenatal exposures: A systematic review. *Paediatric and Perinatal Epidemiology.* 35(1):130-142. (Chapter 3)

**Diemer EW**, Labrecque JA, Tiemeier H, Swanson SA (2020). Application of the instrumental inequalities to a Mendelian randomization study with multiple proposed instruments. *Epidemiology.* 31(1):65-74. (Chapter 4)

Guo K, **Diemer EW**, Labrecque JA, Swanson SA (in progress). Falsification of the instrumental variable conditions in Mendelian randomization studies in the UK Biobank. (Chapter 5)

**Diemer EW**, Havdahl A, Andreassen OA, Munafo MR, Njolstad PR, Tiemeier H, Zuccolo L, Swanson SA (revise and resubmit). Bounding the average causal effect in Mendelian randomization studies with multiple proposed instruments: An application to prenatal alcohol exposure and attention deficit hyperactivity disorder. *American Journal of Epidemiology.* (Chapter 6)

**Diemer EW**, Zuccolo L, Swanson SA (in progress). Partial identification of the average causal effect using bounds computed in multiple study populations: the challenge of combining Mendelian randomization studies. (Chapter 7)

# Contents

# Chapter 1

# Introduction

The prenatal period is a potentially critical time for the development of psychiatric and behavioral disorders in the offspring. Substantial neural formation and organization occur during this period, with the majority of cortical neurons forming prior to 20 weeks, and cortical neuron migration peaking between gestational weeks 12 and 20 (Monk et al., 2019). Thus, the fetal central nervous system is highly plastic, and variations in centralnervous system organization that occur during this period will impact later neural network patterns and behavioral development (Anderson and Thomason, 2013). Beyond the central nervous system, fetal development may impact psychiatric health through a number of mechanisms. DNA methylation, the binding of a methyl group to a DNA position, primarily at sites where a cytosine is next to a guanine, can alter the binding of transcription factors, impacting gene expression, and through that expression, any number of offspring characteristics (Felix et al., 2018). DNA methylation changes substantially during fetal development, meaning environmental impacts during the prenatal period could result in durable, long-lasting alterations to DNA methylation (Marsit, 2015). In addition, large doses of teratogenic substances like alcohol and tobacco have been associated with a number of systemic changes in offspring, including HPA axis dysregulation, increased oxidative stress, and cell apoptosis (Ornoy et al., 2018). Fetal tissue development also relies on a sufficient supply of maternal micronutrients, and micronutrient deficiencies during pregnancy can negatively impact offspring health in a number of ways. Examples of this include the relationship between folate deficiency and neural tube defects, and offspring rickets resulting from severe maternal vitamin D deficiency (Gernand et al., 2016).

Consumption of large volumes of teratogenic substances like alcohol have clear and well-established negative effects on offspring health (Streissguth et al., 1980). However, evaluation of the effect of low to moderate exposure to these substances during pregnancy on offspring psychiatric health is challenging. First, because of the potential for harm, randomized control trials of moderate use of such substances are often unethical to conduct in pregnant women. Second, observational studies of the topic are also difficult. Many common methods for estimating causal effects in observational data rely on an assumption that individuals at different levels of the exposure are conditionally exchangeable with regards to counterfactual outcome, meaning that, within levels of the other variables in the model, had individuals who received one level of the exposure received another level of the exposure, they would have the same outcome as those individuals who actually received the different level of the exposure (Hernan and Robins, 2018). To meet this assumption, many commonly used methods, including outcome regression, the g-formula, inverse probability of treatment weighting, and g-estimation, rely on identification, measurement, and adjustment for confounders (common causes of the exposure and outcome) (Hernan and Robins, 2018). Observational studies of the use of substances during pregnancy are troublesome because of the large number of potential confounders that are difficult to accurately measure. Such confounders include shared genetic factors, socioeconomic status, maternal health behaviors, overall diet, social support, and engagement with healthcare providers. Previous studies have found that alcohol consumption during pregnancy was associated with older maternal age, cigarette smoking, use of illicit drugs, pregnancy unwantedness, domestic violence, single parenthood, primiparity, and pre-pregnancy drinking (Giglia and Binns, 2007; O'Keeffe et al., 2015; Walker et al., 2011). A retrospective study of Dutch mothers found that higher education, older maternal age, smoking, and primiparity were all associated with consuming any alcohol during pregnancy (Lanting et al., 2015). Many of these factors are also associated with differences in risk of offspring psychiatric disorders. Small qualitative studies have found that women who chose to abstain from alcohol reported doing so due to "a sense that alcohol was generically harmful", and guilt over breaking perceived social norms against drinking during pregnancy (Jones and Telenta, 2012). Women who did consume alcohol during pregnancy generally reported they felt alcohol was important to their social functioning, and perceived low and moderate consumption of alcohol during pregnancy to be low risk (Meurk et al., 2014). This suggests that women who abstain completely from alcohol may be especially conscientious, and may be more likely to engage in other behaviors to reduce harm to the fetus than mothers who consume moderate amounts of alcohol. However, especially conscientious moth-

14

ers may also experience more stress during pregnancy than less conscientious mothers, which might also impact offspring psychiatric health (O'Donnell et al., 2014; Van den Bergh et al., 2017). Both this additional stress and engagement in other health behaviors could potentially impact offspring behavioral health outcomes, resulting in confounding.

Estimates of the average causal effect of differences in micronutrient status are similarly vulnerable to bias from unmeasured confounding. Vitamin D status is primarily determined by endogenous production in the skin, with a relatively small contribution from oral intake of foods such as fatty fish, egg yolks, mushrooms, and yeast (Macdonald et al., 2011). Because endogenous vitamin D production differs according to sun exposure and skin tone, the effects of vitamin D status on offspring psychiatric health will generally be confounded by race/ethnicity (Clemens et al., 1982; Kessler et al., 2006; Merikangas et al., 2010), as well as safe access to unpolluted outdoor spaces (Macdonald et al., 2011; McCormick, 2017). Moreover, pregnant women's use of vitamin D supplements may itself be related to their general tendency towards health-seeking behaviors, discussions with their healthcare providers, and their ability to afford supplements (Barnes et al., 2019). Altogether, this suggests that causal inference methods that rely on confounder identification and adjustment may produce biased estimates in the setting of pregnancy exposures and offspring outcomes.

It is understandable, then, that some in the research community have argued for broader use of alternative causal inference methods (Gage et al., 2016), as well as increased research into physiologic mechanisms by which prenatal exposures could impact later life outcomes (Sujan et al., 2019). In particular, some researchers have recommended using Mendelian randomization (MR), a method that has been growing in popularity in recent years, to study the effects of prenatal exposures on the offspring. Under certain assumptions, MR, an application of instrumental variable (IV) methods proposing single nucleotide polymorphisms (SNPs) as instruments, allows for unbiased estimation of causal effects even in the presence of unmeasured exposure-outcome confounding. Specifically, an MR study proposing a single SNP as an instrument requires that (Hernán and Robins, 2006):

1. The SNP $Z$ is associated with the exposure $A$.

2. $Z$ does not affect the outcome $Y$ except through $A$.

3. Individuals at different levels of $Z$ are comparable (i.e. exchangeable) with regards to counterfactual outcome $Y^a$.

(Throughout this dissertation, unless otherwise noted, we use $Z$ to denote SNPs or proposed instruments, $X$ or $A$ to denote exposures, and $Y$ to denote outcomes). These assumptions alone are only sufficient for sharp causal null hypothesis testing and bounding (Hernán and Robins, 2006). In order to obtain point estimates for the average causal effect, investigators must also make one of a set of possible homogeneity assumptions. Essentially, these are assumptions about the extent to which the causal effect of interest varies across the study population. Formally, investigators must assume one of the following holds (Hernan and Robins, 2018; Tchetgen et al., 2017; Wang and Tchetgen, 2018):

4a. The effect of $A$ on $Y$ is identical (constant) for all individuals in the population.

4b. No additive effect modification of the $A$-$Y$ relationship by $Z$ in either the treated or untreated.

4c. No multiplicative effect modification of the $A$-$Y$ relationship by $Z$ in either the treated or untreated.

4d. No additive effect modification of the $A$-$Y$ relationship by the confounders $U$.

4e. The $Z$-$A$ association on the additive scale is constant across levels of the confounders $U$.

Settings in which these conditions might be violated are discussed in Chapter 3 and Chapter 8. Some alternative estimators have been proposed, though these require similarly strong homogeneity conditions (Bowden et al., 2015; Bowden et al., 2016; Hartwig et al., 2017; Tchetgen et al., 2017).

Prenatal exposures present an especially compelling case for the use of MR. Because an individual's genes are with them for life, the MR conditions can be violated if the relationship between a genetic variant proposed as an instrument and the exposure change over time (which will necessarily occur if an individual's exposure level changes over time, and exposure levels at the second time point are affected by the genetic variants directly) (Labrecque and Swanson, 2019). However, when maternal genetic variants are proposed as instruments for an offspring's exposure to exposures during pregnancy, said offspring is only directly exposed to maternal genetic variants and levels of exposure in utero.

An investigator can then more reasonably argue that a SNP-exposure relationship remains constant over the 9-month period of pregnancy, and that MR studies of prenatal exposures are less likely to be biased in this way (though children are of course indirectly exposed to parental exposures through behaviors and the passive effects of behaviors like smoking).

However, it is important to note that MR conditions 2,3, and all versions of 4a-e are unverifiable, and, like any causal inference approach, should be carefully considered and weighed within the context of a specific research question and study population. Unlike most MR studies, where the proposed genetic instrument, exposure, and outcome are measured within the same person, pregnancy MR combines data on proposed genetic instruments and exposures in mothers with outcome data in offspring. While this separation has distinct advantages, it also presents a unique causal structure that may complicate the interpretation of certain estimates and could result in unique biases. To this point, no study has investigated what types of violations of the MR conditions are discussed in pregnancy MR studies, and what methods are used to mitigate bias resulting from these violations.

Importantly, MR studies frequently propose large numbers of SNPs as joint instruments. When multiple genetic variants are proposed as instruments, the MR conditions must hold for all SNPs proposed as instruments both individually and jointly. While this means that MR studies impose increasingly large numbers of assumptions as they propose larger numbers of SNPs as instruments, the availability of multiple proposed instruments could also provide opportunities for novel applications of existing IV methods. Historically, although the additional homogeneity conditions are often implausible, bounding approaches have been unpopular. Under the primary IV conditions alone, one can estimate bounds, meaning upper and lower limits on the average causal effect (Balke and Pearl, 1997; Manski, 1990; Robins, 1989). Importantly, bounds are distinct from confidence intervals (and in fact have their own confidence intervals). In contrast to a confidence interval, a bound will not collapse to point in an infinite sample. The unpopularity of this approach may be because, in the all-binary setting, IV bounds are often wide. However, in the context of MR studies with multiple proposed instruments, it may be possible to narrow bounds enough to identify directions of effect without additional point-estimating assumptions. Similarly, the instrumental inequalities, a falsification method implied by the IV model, are rarely used in applied studies because the method is only able to detect extreme biases when used with a single binary proposed instrument (Glymour et al., 2012). In MR studies with multiple proposed instruments, the instrumental inequalities may prove to be

more informative. Moreover, by applying both the instrumental inequalities and bounding approaches across different combinations of a set of SNPs proposed as instruments, we may be able to identify subsets of SNPs that are less likely to provide biased estimates, and to evaluate how strongly our conclusions depend upon a particular assumption in the model.

## 1.1 Aims

Given this, the ultimate aim of this dissertation was to explore how to improve the analysis of observational data to study the effect of maternal nutritional and substance use exposures on offspring psychiatric outcomes. To do so, we investigated potential physiologic mechanisms by which prenatal exposures might impact psychiatric health in children, and explored the use of MR to study effects of pregnancy exposures on offspring outcomes. In Chapter 2, I discuss a study of the associations between maternal mid-pregnancy vitamin D sufficiency and offspring DNA methylation in cord blood. In Chapter 3, I then review the use of MR to study pregnancy exposures and offspring outcomes in the existing literature, with particular attention to the reporting of methodologic limitations. Next, in Chapter 4, I explore the use of the instrumental inequalities in MR studies with multiple proposed instruments through simulations as well as an application to the study of prenatal vitamin D sufficiency and offspring psychiatric symptoms in real data. I also provide software for the implementation and visualization of the instrumental inequalities. In Chapter 5, to evaluate how the instrumental inequalities performed in other study settings, I then apply the instrumental inequalities to MR studies of the effects of several commonly studied exposures on coronary heart disease in a large sample on adults in the United Kingdom. In Chapter 6, to investigate the use of bounding approaches in MR studies with multiple proposed instruments, I then calculate bounds on the average causal effect of maternal alcohol consumption during pregnancy on offspring attention deficit –hyperactivity disorder in two European cohorts. I also provide software for the implementation and visualization of the IV bounds. Finally, in Chapter 7, I describe how information about bounds on a causal effect of interest generated in different study populations can be combined, using an application to the study of maternal pregnancy alcohol consumption and offspring attention-deficit hyperactivity disorder. In Chapter 8, I discuss the broader implications of the findings of this dissertation and directions for future research, including those involving the mathematical connections between the instrumental inequalities and the bounds.

## 1.2   Setting

With the exception of Chapters 3 and 5, the studies presented in this dissertation were all conducted within one or more prospective cohort studies of European or North American mothers and children. The study of maternal pregnancy vitamin D sufficiency and offspring DNA methylation shown in Chapter 2 included analyses in 7 cohorts from the Pregnancy and Childhood Epigenetics Consortium, based in the Netherlands, United Kingdom, Norway, Finland, Canada, and the United States (Felix et al., 2018). The study on the use of the instrumental inequalities in Chapter 4 was embedded within Generation R, a prospective cohort from fetal life onward, based in Rotterdam, the Netherlands (Jaddoe et al., 2006). The studies on bounding approaches presented in Chapters 6 and 7 are based on results in the Avon Longitudinal Study of Parents and Children (ALSPAC) and the Norwegian Mother, Father, and Child Cohort Study (MoBa). ALSPAC is a longitudinal prospective cohort study that recruited pregnant women in former Avon county, United Kingdom in 1991 and 1992, and continues to follow the offspring of those pregnancies today (Boyd et al., 2013). MoBa is a large population-based cohort study conducted by the Norwegian Institute of Public Health, which recruited pregnant women across Norway between 1999 and 2008, and has continued to collect data on both parents and offspring of those pregnancies (Magnus et al., 2006). The study presented in Chapter 5, which focuses on applying the instrumental inequalities to non-pregnancy MR context, used data from the UKBiobank, a prospective cohort study of approximately 500,000 individuals in the United Kingdom, aged between 40 and 69 at initial recruitment (Bycroft et al., 2018).

# References

Anderson, A. L., & Thomason, M. E. (2013). Functional plasticity before the cradle: A review of neural functional imaging in the human fetus. *Neuroscience and Biobehavioral Reviews*, *37*(9), 2220–2232.

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176.

Barnes, L., Barclay, L., McCaffery, K., & Aslani, P. (2019). Complementary medicine products: Information sources, perceived benefits and maternal health literacy. *Women and Birth*, *32*(6), 493–520.

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International journal of epidemiology*, *44*(2), 512–525.

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

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., & Davey Smith, G. (2013). Cohort profile: The 'children of the 90s'—the index offspring of the avon longitudinal study of parents and children. *International journal of epidemiology*, *42*(1), 111–127.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliot, L., Sharp, K., Mother, A., Vukcevic, D., Delanueau, O., O'Connell, J., & Cortes, A. (2018). The uk biobank resource with deep phenotypic and genomic data. *Nature*, *562*(7726), 203–209.

Clemens, T. L., Henderson, S. L., Adams, J. S., & Holick, M. F. (1982). Increased skin pigment reduces the capacity of skin to synthesise vitamin d3. *The Lancet*, *319*(8263), 74–76.

Felix, J. F., Joubert, B. R., Baccarelli, A. A., Sharp, G. C., Almqvist, C., Annesi-Maesano, I., Arshad, H., Baïz, N., Bakermans-Kranenburg, M. J., & Bakulski, K. M. (2018). Cohort profile: Pregnancy and childhood epigenetics (pace) consortium. *International journal of epidemiology*, *47*(1), 22–23u.

Gage, S. H., Munafò, M. R., & Davey Smith, G. (2016). Causal inference in developmental origins of health and disease (dohad) research. *Annual review of psychology*, *67*, 567–585.

Gernand, A. D., Schulze, K. J., Stewart, C. P., West, K. P., & Christian, P. (2016). Micronutrient deficiencies in pregnancy worldwide: Health effects and prevention. *Nature Reviews Endocrinology*, *12*(5), 274–289.

Giglia, R. C., & Binns, C. W. (2007). Patterns of alcohol intake of pregnant and lactating women in perth, australia. *Drug and alcohol review*, *26*(5), 493–500.

Glymour, M. M., Tchetgen Tchetgen, E. J., & Robins, J. M. (2012). Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American journal of epidemiology*, *175*(4), 332–339.

Hartwig, F. P., Davey Smith, G., & Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology*, *46*(6), 1985–1998.

Hernan, M. A., & Robins, J. M. (2018). *Causal inference.* Chapman; Hall/CRC.

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 360–372.

Jaddoe, V. W. V., Mackenbach, J. P., Moll, H. A., Steegers, E. A. P., Tiemeier, H., Verhulst, F. C., Witteman, J. C. M., & Hofman, A. (2006). The generation r study: Design and cohort profile. *European journal of epidemiology*, *21*(6), 475.

Jones, S. C., & Telenta, J. (2012). What influences australian women to not drink alcohol during pregnancy? *Australian journal of primary health*, *18*(1), 68–73.

Kessler, R. C., Adler, L., Barkley, R., Biederman, J., Conners, C. K., Demler, O., Faraone, S. V., Greenhill, L. L., Howes, M. J., & Secnik, K. (2006). The prevalence and correlates of adult adhd in the united states: Results from the national comorbidity survey replication. *American Journal of psychiatry*, *163*(4), 716–723.

Labrecque, J. A., & Swanson, S. A. (2019). Interpretation and potential biases of mendelian randomization estimates with time-varying exposures. *American journal of epidemiology*, *188*(1), 231–238.

Lanting, C. I., van Dommelen, P., van der Pal-de, K. M., Gravenhorst, J. B., & van Wouwe, J. P. (2015). Prevalence and pattern of alcohol consumption during pregnancy in the netherlands. *BMC Public Health*, *15*(1), 1–5.

Macdonald, H. M., Mavroeidi, A., Fraser, W. D., Darling, A. L., Black, A. J., Aucott, L., O'Neill, F., Hart, K., Berry, J. L., & Lanham-New, S. A. (2011). Sunlight and dietary contributions to the seasonal vitamin d status of cohorts of healthy postmenopausal women living at northerly

latitudes: A major cause for concern? *Osteoporosis International*, *22*(9), 2461–2472.

Magnus, P., Irgens, L. M., Haug, K., Nystad, W., Skjærven, R., & Stoltenberg, C. (2006). Cohort profile: The norwegian mother and child cohort study (moba). *International journal of epidemiology*, *35*(5), 1146–1150.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, *80*(2), 319–323.

Marsit, C. J. (2015). Influence of environmental exposure on human epigenetic regulation. *Journal of Experimental Biology*, *218*(1), 71–79.

McCormick, R. (2017). Does access to green space impact the mental well-being of children: A systematic review. *Journal of Pediatric Nursing*, *37*, 3–7.

Merikangas, K. R., He, J.-p., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., Benjet, C., Georgiades, K., & Swendsen, J. (2010). Lifetime prevalence of mental disorders in us adolescents: Results from the national comorbidity survey replication–adolescent supplement (ncs-a). *Journal of the American Academy of Child and Adolescent Psychiatry*, *49*(10), 980–989.

Meurk, C. S., Broom, A., Adams, J. S., Hall, W., & Lucke, J. (2014). Factors influencing women's decisions to drink alcohol during pregnancy: Findings of a qualitative study with implications for health communication. *BMC pregnancy and childbirth*, *14*(1), 1–9.

Monk, C., Lugo-Candelas, C., & Trumpff, C. (2019). Prenatal developmental origins of future psychopathology: Mechanisms and pathways. *Annual review of clinical psychology*, *15*, 317–344.

O'Donnell, K. J., Glover, V., Barker, E. D., & O'Connor, T. G. (2014). The persisting effect of maternal mood in pregnancy on childhood psychopathology. *Development and psychopathology*, *26*(2), 393–403.

O'Keeffe, L. M., Kearney, P. M., McCarthy, F. P., Khashan, A. S., Greene, R. A., North, R. A., Poston, L., McCowan, L. M. E., Baker, P. N., & Dekker, G. A. (2015). Prevalence and predictors of alcohol use during pregnancy: Findings from international multicentre cohort studies. *BMJ open*, *5*(7).

Ornoy, A., Koren, G., & Yanai, J. (2018). Is post exposure prevention of teratogenic damage possible: Studies on diabetes, valproic acid, alcohol and anti folates in pregnancy: Animal studies with reflection to human. *Reproductive Toxicology*, *80*, 92–104.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

Streissguth, A., Landesman-Dwyer, S., Martin, J., & Smith, D. (1980). Teratogenic effects of alcohol in humans and laboratory animals. *Science*, *209*(4454), 353–361.

Sujan, A. C., Öberg, A. S., Quinn, P. D., & D'Onofrio, B. M. (2019). Annual research review: Maternal antidepressant use during pregnancy and offspring neurodevelopmental problems–a critical review and recommendations for future research. *Journal of Child Psychology and Psychiatry*, *60*(4), 356–376.

Tchetgen, E. J. T., Sun, B., & Walter, S. (2017). The genius approach to robust mendelian randomization inference. *arXiv preprint arXiv:1709.07779*.

Van den Bergh, B. R. H., van den Heuvel, M. I., Lahti, M., Braeken, M., de Rooij, S. R., Entringer, S., Hoyer, D., Roseboom, T., Räikkönen, K., & King, S. (2017). Prenatal developmental origins of behavior and mental health: The influence of maternal stress in pregnancy. *Neuroscience and Biobehavioral Reviews*.

Walker, M. J., Al-Sahab, B., Islam, F., & Tamim, H. (2011). The epidemiology of alcohol utilization during pregnancy: An analysis of the canadian maternity experiences survey (mes). *BMC Pregnancy and Childbirth*, *11*(1), 52.

Wang, L., & Tchetgen, E. T. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *80*(3), 531.

# Chapter 2

# Epigenome-wide meta-analysis of prenatal Vitamin D insufficiency and cord blood DNA methylation

Elizabeth W. Diemer, Johanna Tuhkahnen, Sara Sammallahti, Kati Heinonen, Alexander Neumann, Sonia L Robinson, Matthew Suderman, Jianping Jin, Christian Page, Ruby Fore, Sheryl L Rifas-Shiman, Emily Oken, Patrice Perron, Luigi Bouchard, Marie France Hivert, Katri Räikköne, Jari Lahti, Edwina H. Yeung, Weihua Guan, Sunni L. Mumford, Maria C. Magnus, Siri Håberg, Wenche Nystad, Christine L. Parr, Stephanie London, Janine F. Felix, Henning Tiemeier

## 2.1  Abstract

**Background**: Low maternal vitamin D concentrations during pregnancy have been associated with a range of offspring health outcomes. DNA methylation is one mechanism by which maternal vitamin D status during pregnancy could impact offspring health in later life.

**Objective**: We aimed to evaluate whether maternal vitamin D insufficiency during pregnancy was conditionally associated with DNA methylation in offspring cord blood.

**Methods**: Maternal vitamin D insufficiency (plasma 25-hydroxy Vitamin D $\leq$ 75 nmol/L) during pregnancy and offspring cord blood DNA methylation, assessed using Illumina Infinium 450k or Illumina EPIC Beadchip, was collected for 3,738 mother-child pairs in 7 cohorts as part of the Pregnancy and Childhood Epigenetics (PACE) consortium. Associations between maternal vitamin D and offspring DNA methylation, adjusted for fetal sex, maternal smoking, maternal age, maternal pre-pregnancy or early pregnancy BMI, maternal education, gestational age at measurement of 25(OH)D, parity, and cell type composition, were estimated using robust linear regression in each cohort, and a fixed effects meta-analysis was conducted.

**Results**: The prevalence of Vitamin D insufficiency ranged from 44.3% to 78.5% across cohorts. Across 364,678 CpG sites, none were associated with maternal Vitamin D insufficiency at an epigenome-wide significant level after correcting for multiple testing using Bonferroni correction or a less conservative Benjamini-Hochberg False Discovery Rate approach (FDR p > 0.05).

**Conclusions**: In this epigenome-wide association study, we did not find convincing evidence of a conditional association of vitamin D insufficiency on offspring DNA methylation at any measured CpG site.

## 2.2  Introduction

Vitamin D is a fat soluble vitamin and precursor to 1,25-dihydroxyvitamin D $(1,25(OH)_2D)$, which plays a key role in calcium homeostasis and bone health (Holick and Chen, 2008). There are two physiologically active forms of Vitamin D: $D_2$, found primarily in mushrooms and yeast, and $D_3$, which is synthesized in the skin via UV radiation, and is found in a limited number of foods, including fatty fish and egg yolks (Holick, 2006). Both forms are hydroxylated in the liver to form 25-hydroxyvitamin D (25(OH)D), the major circulating form and indicator of vitamin D status, which is then hydroxylated again, primarily in the kidneys, to form $1,25(OH)_2D$ (Holick and Chen, 2008; Jones et al., 1998).

Vitamin D concentrations are associated with a range of health outcomes, though these relationships are often nonlinear. Severely deficiencies can result in the development of rickets or osteomalacia, but extremely high concentrations (typically above 375 nmol/L), can result in vitamin D toxicity and a range of severe symptoms including recurrent vomiting, confusion, polyuria, and dehydration (though toxicity usually results from overconsumption of supplements or comorbid disorders) (Liu et al., 2018; Marcinowska-Suchowierska et al., 2018). Nonlinearities have also been noted in associations between vitamin D concentrations and other health outcomes, including preterm birth and cardiovascular disease (Bodnar et al., 2015; Welles et al., 2014). Individuals are typically considered to be clinically vitamin D deficient at a 25(OH)D concentration $\leq$ 50 nmol/L and vitamin D insufficient at a 25(OH)D concentration $\leq$ 75 nmol/L (Hollis, 2005).

Previous research has found that vitamin D insufficiency is relatively common in pregnancy, impacting between 42.1% and 97% of pregnant study participants, depending on location and race/ethnicity (Bodnar et al., 2007; Ginde et al., 2010; Johnson et al., 2011). Maternal 25(OH)D readily crosses the placental barrier, and is likely hydroxylated in fetal kidneys and the placenta itself to form $1,25(OH)_2D$ (Larqué et al., 2018). Low maternal 25(OH)D concentrations during pregnancy have been associated with a wide range of health outcomes in offspring, including bone health, symptoms of Attention Deficit-Hyperactivity Disorder, symptoms of Autism Spectrum Disorder, asthma, eczema, and autoimmune conditions (Boghossian et al., 2019; Erkkola et al., 2011; Javaid et al., 2006; Magnusson et al., 2016; Morales et al., 2015; Song et al., 2017; Wei et al., 2016). Results of randomized trials of vitamin D supplementation in pregnancy have produced mixed results, but meta-analyses have suggested that Vitamin D supplementation during pregnancy may reduce risk of low birthweight off-

spring (Palacios et al., 2019; Roth et al., 2017). However, the mechanism by which maternal Vitamin D levels impact offspring outcomes is not yet clear.

One possible mechanism is through offspring DNA methylation. $1,25(OH)_2D$ has a known impact on gene expression through direct binding of vitamin D response elements via the vitamin D receptor transcription factor (Pike and Meyer, 2014), but some studies have suggested that vitamin D levels may also impact DNA methylation (Beckett et al., 2016; Fetahu et al., 2014). To this point, research on the relationship between maternal vitamin D and offspring DNA methylation has been limited, and little is known about the potential magnitude of possible effects of vitamin D on methylation. To our knowledge, only one previous epigenome wide association study of maternal pregnancy vitamin D and offspring cord blood methylation has been conducted (Suderman et al., 2016). While the previous study did not identify any associations between maternal vitamin D and offspring methylation after adjusting for multiple tests, the study sample included 1,416 mother child pairs, and was likely underpowered to detect weak or moderate epigenetic effects. The aim of this meta-analysis was therefore to investigate associations between maternal mid-pregnancy Vitamin D insufficiency and DNA methylation in offspring cord blood in a sample expanded from the original 1,416 pairs to a total of 3,738 mother-child pairs.

## 2.3  Methods

**Participating Cohorts**

This study was conducted as part of the Pregnancy and Childhood Epigenetics (PACE) consortium (Felix et al., 2018). A total of 7 cohorts participated in the study: the Avon Longitudinal Study of Parents and Children (ALSPAC), the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial, the Genetics of Glucose regulation in Gestation and Growth (Gen3G) cohort, the Generation R Study, the Norwegian Mother, Father, and Child Study (MoBa1, MoBa2), the Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction (PREDO) study, and Project Viva. To limit confounding by ancestry, samples were restricted to participants of self-reported white European ancestry. We additionally excluded offspring of multiple births, and offspring with known congenital abnormalities. To reduce confounding by familial relatedness, where multiple children of the same mother were included in a cohort, only one randomly selected child was included in the final analytic sample.

Ethical approvals for all study protocols were obtained for all participating cohorts. Details on study methods for each cohort are described in detail in the Appendix.

**Measures**

*Maternal Vitamin D Insufficiency*

With the exception of ALSPAC, maternal vitamin D insufficiency was evaluated using serum or plasma samples taken between gestational weeks 8-25. Within ALSPAC, serum samples could be taken at any point during pregnancy, but all samples were normalized to obtain an estimate of concentration at 28 weeks gestation. Maternal vitamin D sufficiency was defined as maternal serum or plasma total $25(OH)D \leq 75$ nmol/L, as recommended by the Endocrine Society (Hollis, 2005). As noted previously, nonlinearities have been noted in associations between vitamin D concentrations and several outcomes, including preterm birth and adult cardiovascular disease (Bodnar et al., 2015; Welles et al., 2014). Evidence supports an association between maternal pregnancy clinical vitamin D insufficiency at this threshold and various adverse outcomes, including preterm birth, type 1 diabetes mellitus, multiple sclerosis, allergies, and atopic disorders (De-Regil et al., 2016). In order to focus our study on the impact of clinically relevant insufficiency on offspring outcomes, we chose to evaluate the relationship between dichotomous maternal vitamin D insufficiency and offspring methylation, rather than continuous $25(OH)D$.

*Offspring Cord Blood DNA Methylation*

Offspring cord blood DNA methylation was evaluated using Illumina Infinium 450k or EPIC BeadChip in all cohorts. Each cohort normalized methylation beta values using their own preferred published normalization method and conducted their own quality control pipeline for probe and sample filtering, as detailed in the Supplementary Materials. To remove outliers, methylation sets were trimmed using the interquartile range (IQR) strategy, meaning beta values below (25th percentile -3*IQR) and above (75th percentile + 3*IQR) were removed.

*Covariates*

All cohorts ran models adjusted for fetal sex, maternal smoking, maternal age, maternal pre-pregnancy or early pregnancy ($< 15$ weeks gestation) BMI, maternal education, gestational age at measurement of $25(OH)D$, parity, and cell type composition. All covariates were selected based on previous literature in order to reduce confounding bias or to remain consistent with previous analyses (Bakulski et al., 2016; Busche et al., 2015; Jaffe and Irizarry, 2014; Krieger et al., 2018; Suderman et al., 2016}). In analyses in MoBa1, MoBa2,

Generation R, PREDO, and ALSPAC, maternal smoking was divided into 3 categories (no smoking during pregnancy, smoking during first trimester only, smoking throughout pregnancy). However, in EAGeR and Project Viva, insufficient data was available to apply this categorization. Within EAGeR, smoking was dichotomized into smoking during pregnancy vs. no smoking during pregnancy. Within Project Viva, smoking status was grouped into 3 categories (never smoked, former smoker, smoked during pregnancy). Maternal self-reported education was categorized according to each cohort's discretion (see Appendix for cohort-specific details). Cell counts were estimated in each cohort using the Bakulski reference set (Bakulski et al., 2016). To control for ancestry, all samples except MoBa were restricted to mother-child pairs of self-reported white European ancestry. MoBa does not collect data on self-reported ancestry. However, only 5.6% of all MoBa mothers report a first language other than Norwegian, suggesting the sample is primarily of Scandinavian ancestry (Magnus et al., 2006). In addition, where maternal genetic data were available, models were adjusted for the first 4 principal components, or a number determined by the cohort to be sufficient for their sample, from DNA methylation data. In some cases, if maternal genetic data was not available, offspring principal components were used as a proxy for maternal genetic ancestry. Each cohort also adjusted for batch effects using methods appropriate to the cohort, and where necessary, included additional covariates to correct for study design (Appendix).

The primary source of vitamin D for most adults is sun exposure (Holick and Chen, 2008). However, sunlight exposure, and thus vitamin D sufficiency status, varies seasonally (Holick and Chen, 2008). A limited amount of research has suggested that season of birth itself may be associated with DNA methylation (Lockett et al., 2016). As season of 25(OH)D measurement was directly related to season of birth, season of measurement may therefore confound the effect of maternal pregnancy vitamin D sufficiency on offspring methylation. In addition to stimulating vitamin D production in the skin, sunlight exposure appears to degrade folate in the skin (Off et al., 2005; Steindal et al., 2008; Tam et al., 2009). Folate is a source of the one carbon group used to methylate DNA, and maternal folate status has been associated with offspring DNA methylation in both human and animal studies (Crider et al., 2012; Joubert et al., 2016). Maternal sunlight exposure during pregnancy may therefore also impact offspring DNA methylation through folate levels. In order to limit this possible source of confounding, in secondary analyses we additionally adjusted models for season of measurement, which was grouped into 4 categories (February-April, May-July, August-October, and November-January), based

on previous research suggesting vitamin D sufficiency follows a seasonal pattern lagged from astronomical seasons by approximately 8 weeks (Kasahara et al., 2013). All cohorts were located in the Northern Hemisphere, meaning they followed a similar season pattern. Notably, because vitamin D levels in ALSPAC were pre-adjusted for season using a method described previously (Lawlor et al., 2013), ALSPAC results were included only in models adjusted for season of measurement, and not in the base model.

## Statistical Methods

Each cohort performed independent epigenome-wide association studies according to a common pre-specified analysis plan. Associations between maternal vitamin D insufficiency and methylation at each CpG site were evaluated using 3 nested robust linear regression models. In the first model, maternal mid-pregnancy vitamin D sufficiency was modelled as the exposure, and offspring methylation at each CpG site was modelled as the outcome, with adjustment for fetal sex, maternal smoking, maternal age, maternal pre-pregnancy BMI, maternal education, gestational age at measurement of 25(OH)D, parity, and cell type composition. The second model was additionally adjusted for season of measurement.

Prior to meta-analysis, cross-reactive probes flagged by Chen et al or McCartney et al. were removed (Chen et al., 2013; McCartney et al., 2016). We additionally removed all control and polymorphic probes as annotated by meffil (Min et al., 2018), and all probes located on sex chromosomes. As methylation patterns between the Illumina 450K and EPIC arrays are highly correlated (Solomon et al., 2018), analyses conducted on EPIC and 450K arrays were meta-analyzed together. Because only one cohort evaluated DNA methylation using EPIC, probes exclusive to EPIC were removed. In addition, all probes available in less than 3 cohorts or 1000 participants were removed. Flowcharts detailing probe removal are available in the Appendix. QQ plots, PZ plots comparing observed p-values to those calculated from reported beta estimates and standard errors, boxplots of beta distributions, and volcano plots were generated for each analysis and visually inspected to identify possible inflation or bias of test statistics (See Appendix for results). Precision plots were generated across cohorts for each model. Fixed effects inverse-variance weighted meta-analysis of cohort specific results was conducted using Metasoft (Han and Eskin, 2011). Correction for multiple testing was conducted using the Bonferroni method ($p < 1.37 * 10^{-7}$, 364,678 tests). In addition, because the Bonferroni method can be unnecessarily conservative, we also evaluated whether associations met a significance level corresponding to false discovery rate of 0.05,

using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). As a sensitivity analysis, a random effects meta-analysis of cohort specific results was also conducted using Metasoft. In addition to the primary meta-analysis conducted by the first author at Erasmus MC, a shadow meta-analysis was conducted independently by authors at University of Helsinki to minimize human error.

Table 2.1: Description of Included Cohorts

| | ALSPAC | EAGeR | MoBa1 | MoBa2 | Gen3G | Generation R | PREDO | Project Viva |
|---|---|---|---|---|---|---|---|---|
| Country | United Kingdom | United States | Norway | Norway | Canada | Netherlands | Finland | United States |
| N in largest analysis | 499 | 361 | 783 | 177 | 175 | 1154 | 301 | 283 |
| N (%) females | 260 (52.1) | 180 (49.9) | 352 (45.0) | 79 (44.6) | 95 (54.3) | 562 (48.7) | 145 (48.2) | 140 (49.5) |
| N (%) Vitamin D < 75 nmol/L | 306 (61.3) | 162 (44.3) | 389 (49.7) | 91 (51.4) | 128 (73.1) | 706 (61.2) | 198 (65.1) | 222 (78.5) |
| N (%) any smoking during pregnancy | 65 (13.0) | 9 (2.5) | 227 (29.0) | NA | 15 (8.6) | 191 (20.7) | 13 (4.3) | 28 (9.9) |
| N (%) no smoking during pregnancy | 434 (87.0) | 352 (97.5) | 556 (71.0) | NA | 160 (91.4) | 731 (79.3) | 288 (95.7) | 255 (90.1) |
| Mean maternal age , years (sd) [range] | 30.4 (4.4) | 28.3 (4.4) [19, 40] | 29.9 (4.3) [15,41] | 29.4 (4.2) [18,40] | 28.0 (4.1) [19-37] | 31.7 (4.2) [16-46] | 32.6 (5.2) [26] | 32.97 (4.4) [17-44] |
| Mean maternal pre-pregnancy or early pregnancy BMI kg/m2 (sd) [range] | 21.6 (7.0) | 25.1 (5.3) [15.7, 50.9] | 24.0 (4.2) [13.6, 43.2] | 24.1 (4.7) [15.2, 43.0] | 25.7 (5.8) [18.4-54.1] | 23.2 (3.8) [17.3-43.3] | 27.0 (6.4) [37.4] | 24.9 (4.8) [16.7-50.1] |
| N (%) Measurement Feb-Apr | 107 (21.4) | 86 (23.8) | 115 (14.7) | 36 (20.3) | 55 (31.4) | 277 (24.0) | 59 (19.6) | 68 (24.0) |
| N (%) Measurement May-Jul | 139 (27.9) | 100 (27.7) | 279 (35.6) | 90 (50.8) | 41 (23.4) | 358 (31.0) | 66 (21.9) | 69 (24.4) |
| N (%) Measurement Aug-Oct | 149 (29.9) | 83 (23.0) | 268 (34.2) | 33 (18.6) | 49 (28.0) | 254 (22.0) | 95 (31.6) | 64 (22.6) |
| N (%) Measurement Nov-Jan | 104 (20.8) | 92 (25.5) | 121 (15.5) | 18 (10.2) | 30 (17.1) | 265 (23.0) | 81 (26.9) | 82 (29.0) |
| N Parity =0 | 235 | 145 (40.2) | 342 (43.7) | 81 (45.8) | 59 (33.7) | 703 (60.9) | 101 (33.6) | 133 (47.0) |

## 2.4   Results

The prevalence of vitamin D insufficiency varied between 44.3% and 78.5% across cohorts (Table 2.1). In our primary analysis, with dichotomous Vitamin D insufficiency as the exposure, we meta-analysed results from a total of 3,239 mother-child pairs. In secondary analyses additionally adjusting for season of measurement we meta-analyzed results from 3,738 mother-child pairs (ALSPAC participants were included only in models additionally adjusted for season of measurement). Table 2.1 summarizes the characteristics of each cohort. Genomic control lambdas (base model 1.06, season of measurement model 0.93) and QQ plots (See Appendix) suggested only mild genomic inflation in the base model, and did not suggest inflation in the season of measurement model. Maternal mid-pregnancy vitamin D insufficiency was not significantly associated with DNA methylation at any individual CpG site when applying a Bonferroni correction for multiple testing ($p < 1.37 * 10^{-7}$), or when applying a more permissive Benjamini-Hochberg threshold (FDR $< 0.05$) (Figure 2.1). P-values were less than $5 * 10^{-5}$ at only 36 CpG sites in this primary analysis. In secondary analyses, methylation was not significantly associated with maternal mid-pregnancy vitamin D sufficiency after adjustment for season of at any measured CpG site (Figure 2.2). Betas from both regression models tended to be small, with relatively large confidence intervals. Results of the random effects meta-analysis were broadly similar. The meta-analysis results for all probes will be made available in a public respository.
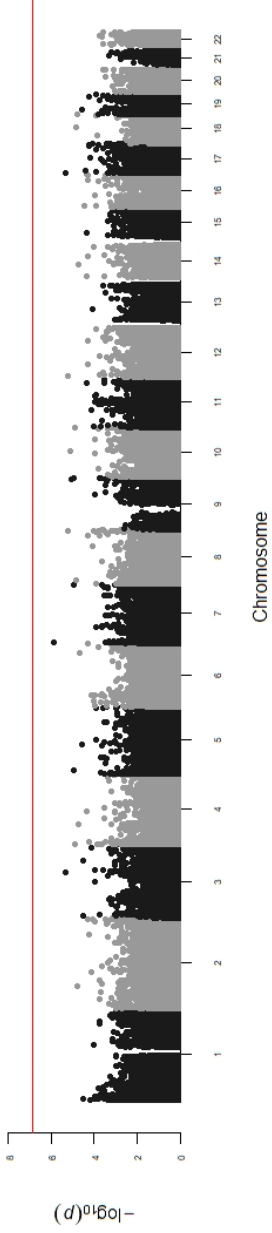
Figure 2.1: Meta-analysis of the association between maternal mid-pregnancy vitamin D insufficiency and offspring DNA methylation in cord blood, adjusted for fetal sex, maternal smoking, maternal age, maternal pre- or early pregnancy BMI, maternal education, gestational age at measurement, parity, cell type composition, principal components of ancestry, and, where appropriate, technical and study design covariates. A total of 364,678 CpG sites were included in this analysis. No sites were epigenome-wide significantly associated with maternal vitamin D insufficiency using either a Bonferroni correction for multiple testing (solid line) or using a Benjamini-Hochberg correction corresponding to a false discovery rate of 0.05.
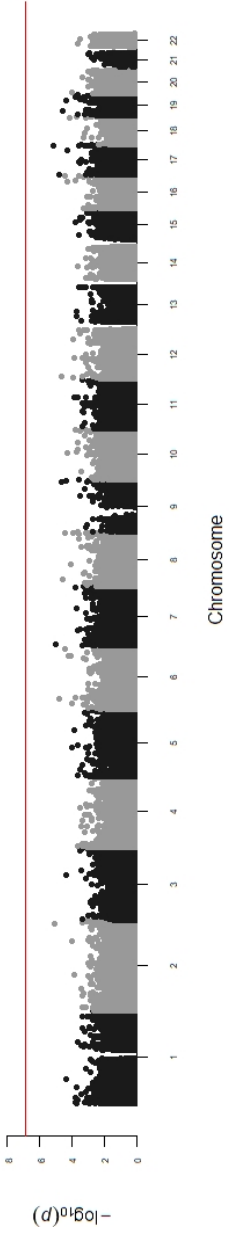
Figure 2.2: Meta-analysis of the association between maternal mid-pregnancy vitamin D insufficiency and offspring DNA methylation in cord blood, adjusted for **season of measurement**, fetal sex, maternal smoking, maternal age, maternal pre- or early pregnancy BMI, maternal education, gestational age at measurement, parity, cell type composition, principal components of ancestry, and, where appropriate, technical and study design covariates. A total of 364,678 CpG sites were included in this analysis. No site were epigenome-wide significantly associated with maternal vitamin D insufficiency using either a Bonferroni correction for multiple testing (solid line) or using a Benjamini-Hochberg correction corresponding to a false discovery rate of 0.05.

## 2.5  Discussion

In our study of European ancestry mothers and children, we did not find evidence of a conditional association between maternal mid-pregnancy vitamin D insufficiency and offspring DNA methylation at any of the measured CpG sites after correction for multiple testing.

This study was, to our knowledge, the largest study of maternal vitamin D during pregnancy and offspring DNA methylation to date. The sample size for our primary analyses (n=3,239 mother-child pairs), was more than double the size of the largest previous analysis of vitamin D during pregnancy. Consistent with that previous study, this study did not find convincing evidence of an association between maternal vitamin D levels and offspring cord blood DNA methylation (Suderman et al., 2016). This lack of an association can, under the assumptions of positivity, consistency, no model misspecification, and conditional exchangeability, be interpreted as evidence that either Vitamin D insufficiency does not have a causal effect on offspring DNA methylation in cord blood at any measured site, or that any possible causal effects of vitamin D insufficiency on offspring cord blood methylation are small (Hernan and Robins, 2018; Hernán, 2018). This analysis was also less vulnerable to reverse causation than many other EWAS designs, because maternal vitamin D sufficiency status during pregnancy occurs prior to offspring DNA methylation in cord blood (measured at birth), and fetal DNA methylation is relatively unlikely to affect maternal vitamin D status during pregnancy.

However, we were only able to estimate associations for 364,678 sites, rather than the approximately 28 million CpG sites contained within the human genome. Although the Illumina 450k array is targeted at regions with potential regulatory functions, it is possible that maternal vitamin D insufficiency impacts offspring DNA methylation at other regions of the epigenome. In particular, previous studies have argued that distal regulatory elements, including enhancers, are severely underrepresented on the 450K (Busche et al., 2015; Pidsley et al., 2016). While previous work has suggested that EWAS designs including more than 1,000 participants are well-powered to detect moderate and small effects (Mansell et al., 2019), it is also likely that our sample size, while large relative to previous studies of maternal vitamin D and offspring DNA methylation, remains too small to detect very weak effects on offspring methylation. Within the cohorts included in this analysis, prevalence of vitamin D insufficiency ranged from 44.3% to 78.5%. This suggests that our analysis was not limited by a low prevalence of the exposure, but rather that

any possible associations of Vitamin D insufficiency with methylation are small, and would require larger sample sizes for detection.

The limitations of our study include possible selection bias. The majority of cohorts in this meta-analysis had a modest participation rate and measured methylation only within a subset of their total sample. If vitamin D insufficiency or offspring methylation were differentially associated with selection into the sample, this could have resulted in bias (Hernan et al., 2004). Previous studies have found inconsistent associations between markers of socioeconomic status and vitamin D insufficiency, though these associations may be partially explained by differences in racial/ethnic background of participants in different socioeconomic groups (Krieger et al., 2018; Malacova et al., 2019; Tønnesen et al., 2016; Voortman et al., 2015). Because some of the studies included in this analysis show evidence of selection on socioeconomic status, a true association between vitamin D insufficiency and socioeconomic status could have resulted in selection bias (Fraser et al., 2013; Jaddoe et al., 2006). It is also possible that our study may have been impacted by residual confounding by supplement use, as women who are wealthier and more health conscious may use vitamin D supplements more often, and may also engage in other behaviors that differentially impact offspring DNA methylation. It is also possible that our results may have been impacted by error in the measurement of vitamin D concentrations, as previous work has found that quality of 25(OH)D measurement varies substantially across cohorts, partly as a result in of differing methods of assessment (Cashman et al., 2015).

Our study was also limited by the measurement of methylation within cord blood. While we did not find any strong associations between maternal vitamin D insufficiency, it is possible that maternal vitamin D insufficiency is more strongly associated with DNA methylation in other offspring tissue types, such as brain tissue, bone, or respiratory tract tissues, though such tissues are obviously much more difficult to obtain. Similarly, our study relies on maternal vitamin D measurements at a single pregnancy time point, generally in mid-pregnancy. However, maternal vitamin D levels may change during pregnancy, and impact methylation more strongly in early or late pregnancy. Importantly, our analysis may have been further limited by the restriction of the sample to participants of white European ancestry to reduce confounding by race/ethnicity. While this restriction was necessary to reduce bias in the analysis, it is possible that the potential effects of vitamin D insufficiency on offspring methylation may be stronger in non-white participants, who are also more likely to experience more severe levels of vitamin D insufficiency (Webb, 2006). Of course, this also limits the generalizability of these results to non-

white women. Our results might be similarly impacted by residual confounding by sunlight exposure. In order to reduce computational burden on participating cohorts, we chose to adjust for sunlight through grouping date of measurement into seasonal categories. However, sunlight exposure varies substantially by latitude, calendar year, and local climate patterns, meaning these categories may be insufficient to completely control for confounding. Our results may also have been impacted by our choice to dichotomize vitamin D at clinical insufficiency levels. While this dichotomization may have limited the power of our analyses to detect epigenetic effects, the relationship between vitamin D and many health outcomes appears to be nonlinear (De-Regil et al., 2016; Holick, 2006; Wang et al., 2018), and our cut-off was selected based on clinical cut-offs relevant to medical decision making.

## 2.6   Conclusions

We did not find strong evidence of an association between maternal mid-pregnancy vitamin D insufficiency and offspring cord blood methylation levels at any measured CpG site among white European ancestry mother-child pairs. Our results, consistent with a previous study of the topic, suggest that large, robust changes in neonatal DNA methylation in response to maternal vitamin D insufficiency are unlikely. However, it is possible that our study was limited by sample size and potential selection bias. Future studies of the relationship between maternal vitamin D insufficiency and offspring DNA methylation could include more racial/ethnically diverse samples, larger sample sizes, measurement of methylation in other offspring cell types, and may consider exploring associations between offspring DNA methylation in cord blood and maternal vitamin D levels on offspring methylation at different periods across gestation.

# Appendix

## Cohort-specific descriptions of data collection

### ALSPAC
*Study population*

ALSPAC is a "transgenerational prospective observational study investigating influences on health and development across the life course" (Boyd et al., 2013; Fraser et al., 2013). Participants comprise a cohort of offspring born to pregnant women recruited in 1991-2 in Bristol, UK. Participants have been followed through a series of ongoing data collection waves involving questionnaires and clinical assessments. Please note that the study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/). While data is available for a total of 14,451 mother-child pairs, DNA methylation was measured in cord blood for approximately 1000 mother-child pairs. For the current study, we restricted analyses to pairs with complete data on mid-pregnancy Vitamin D, offspring cord blood methylation, and all covariates, resulting in a total analytic sample of 499.

*Maternal mid-pregnancy Vitamin D*

25(OH)D concentrations were measured in serum of non-fasting blood samples taken as part of antenatal care. Samples could be taken from any stage of pregnancy. Measurements were made using high-performance liquid chromatography tandem mass spectrometry in one laboratory. Since vitamin D levels may change with time of the year, and pregnant women with length of gestation, measurements were simultaneously adjusted for both factors to obtain an estimate at 28 weeks gestation (Lawlor et al., 2013). Other estimates to 0 and 34 weeks gestation were highly correlated (R ~ 0.6 and 0.9, respectively).

*Offspring cord blood DNA methylation*

DNA methylation was measured for approximately 1000 mother-child pairs in the cord blood and peripheral blood of study children at ages 7 and 15-17 years and in the peripheral blood of mothers approximately 18 years after the birth of the study child. The resulting profiles comprise the Accessible Resource for Integrated Epigenomics Studies (Relton et al., 2015) (ARIES, http://www.ariesepigenomics.org.uk/). All data are available by request from the Avon Longitudinal Study of Parents and Children Executive Committee (http://www.bristol.ac.uk/alspac/researchers/access/) for researchers who

meet the criteria for access to confidential data. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

Genomic DNA was obtained from blood samples and bisulphite converted using the Zymo EZ DNA MethylationTM kit (Zymo, Irvine, CA). DNA methylation was quantified using the Illumina HumanMethylation450 BeadChip according to manufacturer's instructions. During the data generation process a wide range of batch, variables were recorded in a purpose-built laboratory information management system (LIMS) (Boyd et al., 2013).The LIMS also reported quality control (QC) metrics from the standard control probes on the 450k BeadChip for each sample. Samples failing QC were excluded from further analysis and the assay repeated. Sample QC and normalization was completed using the meffil package as previously described (Min et al., 2018). Briefly, probe intensities underwent a functional normalization approach (Fortin et al., 2014) using the first 10 PCs of the Illumina 450K array control probes. This approach includes subset quantile normalization of the data and normal-exponential out-of-band background correction. In addition, twenty surrogate variables (Leek and Storey, 2007) were generated and included in all regression models.

*Covariates*

Maternal smoking status, pre-pregnancy weight, maternal education, maternal age, and parity were obtained by questionnaire during pregnancy. Information on fetal sex was obtained via self report and administrative records. Estimation of six different white blood cell types (CD8+ T and CD4+ T lymphocytes, CD56+ natural killer cells, CD19+ B cells, CD14+ monocytes, and granulocytes) by Houseman method (Jaffe and Irizarry, 2014).


**EAGeR**

*Study Population*

The Effects of Aspirin in Gestation and Reproduction (EAGeR) trial was a block-randomized, double-blind, placebo-controlled trial evaluating the effect of preconception-initiated daily low dose aspirin on live birth. Details of the trial design have been described in detail elsewhere (Schisterman et al., 2013). Briefly, participants who had previously experienced 1-2 prior pregnancy losses, and were currently attempting to conceive were recruited from 2007 to 2011 at 4 United States medical centers. Exclusion criteria for the study included a

known history of infertility treatment, pelvic inflammatory disease, tubal occlusion, endometriosis, anovulation and polycystic ovarian syndrome, or uterine abnormality, resulting in enrollment of 1,228 women. Participants were followed for six menstrual cycles while attempting pregnancy or throughout pregnancy if they conceived. For the current study, we restricted the sample to women who had conceived, and had complete data on midpregnancy vitamin D, offspring cord blood methylation, and all covariates, resulting in an analytic sample of 361 mother-child pairs.

*Maternal midpregnancy Vitamin D*

Serum samples were collected at 8 weeks gestation and cryostored at -80° C prior to analysis. Total 25-hydroxy-vitamin D [25(OH)D] was measured in serum with the 25-hydroxyvitamin D ELISA solid phase sandwich enzyme immunoassay (BioVendor R&D, Ashville, NC, USA). Further details on the measurement of 25(OH)D in the EAGeR trial have been published previously (Mumford et al., 2018).

*Offspring Cord Blood Methylation*

Beginning in 2009, the trial collected 10 ml cord blood from over 90% of deliveries at the Utah trial site. Cord blood was centrifuged and separated into plasma and buffy coat. Samples were subsequently frozen at -80 degrees C. Genome-wide DNA methylation was measured with the Infinium MethylationEPIC Bead Chip. Methylation data were processed using the minfi package in R, which included identification of failed probes and scaling with Illumina control probes to determine methylation values. Quantile normalization was used to normalize beta values between two types of probes. We used principal component analysis (PCA) to detect further outliers and samples mismatched for sex. Samples mismatched for sex were excluded. Beta values were replaced as missing if the detection P-value was $> 0.01$ or bead counts $< 3$. In the analyses the results were set as "NA", when the missing values of the CpG site is greater than 3% to enable the code to run without the need to remove the missing values.

*Covariates*

Cell type mixture was estimated on the full set of normalized methylation data using the Bakulski et al., 2016 Bakulski et al., 2016 reference dataset for cord blood (FlowSorted.CordBlood.450K package). Maternal age, smoking status, and income were measured via self-report. Maternal pre-pregnancy BMI was measured directly, prior to pregnancy.

**Gen3G**

*Study Population*

The Genetics of Glucose regulation in Gestation and Growth (Gen3G) cohort is a prospective observational pre-birth cohort study aimed at investigating glucose regulation determinants in pregnancy and fetal growth, based on the Eastern Townships Region of Quebec, Canada (Guillemette et al., 2016). All women who received prenatal care directly at the Centre Hospitalier Universitaire de Sherbrooke (CHUS), a CHUS affiliated health center, or planned delivery at CHUS between January 2010 and June 2013, were considered eligible. A total of 1034 pregnant were women were recruited at the 1st trimester, of which 10 were excluded due to the presence of a multiple pregnancy. Additional exclusion criteria included known pre-pregnancy diabetes, use of a medication known to influence glucose tolerance, glycated haemoglobin (HBA1c) $>= 6.5\%$ of 1 h glucose $>= 10.3$ mmol/L post 50 g glucose challenge test, miscarriage, medical abortions, or health problems that prohibited participation. A total of 854 participants were followed through delivery. Recruitment and characteristics of the cohort have been described in detail elsewhere (Guillemette et al., 2016). For the current study, we restricted to mother-child pairs with complete data on maternal mid-pregnancy Vitamin D, offspring DNA methylation, and available covariates, resulting in a total analytic sample of 175 pairs.

*Maternal mid-pregnancy Vitamin D*

Non-fasting blood samples were collected at the first study visit (mean 9.6 weeks gestation). Aprotinin was added to blood samples, which were centrifuged at 2500g at 4 degrees C for 10 minutes, and aliquoted for storage at -80 degrees C. 25(OH)D2 and 25(OH)D3 concentrations were assessed using liquid-liquid extraction followed by liquid chromatography-electrospray tandem mass spectrometry (Quattro micro mass spectrometer; Waters, Milford, MA). 25(OH)D concentrations were calculated as the total of the two measurements (Switkowski et al., 2019).

*Offspring DNA methylation*

Cord blood samples were collected via syringe from the umbilical vein after delivery. Bisulfite conversion was performed using the EZ-96 DNA methylation kit (Zymo research Corporation, Irvine, USA). We used the Infinium Human-Methylation450 BeadChip (Illumina Inc., San Diego, USA) to measure the methylation level as a beta value ranging from 0 (no methylation) to 1 (complete methylation). During quality control, we removed samples that were outliers on the MDS plot, samples with $> 5\%$ missingness, probes missing in more than 20% of samples, and duplicates. DASEN normalization was per-

formed using the watermelon package in R. In analyses, ComBat was used to adjust for sample Plate while protecting dichotomized vitamin D in the model statement.

*Covariates*

Fetal sex was collected from medical records. Maternal age and parity were assessed via questionnaire at the beginning of pregnancy. Gestational age was calculated based on reported last menstrual period, and corrected by ultrasound dating when appropriate. Maternal smoking self-reported in the first trimester questionnaire, and grouped into 3 categories (no smoking in pregnancy, stopped smoking in the beginning of pregnancy, and smoked during pregnancy). Maternal early pregnancy BMI was calculated using weight and height measured by research staff according to standard procedures at the first trimester visit. Measures of maternal education were not available in this cohort. The sample included in this analysis is comprised entirely on women of European ancestry. We used the Bakulski -based Houseman method (Bakulski et al., 2016; Houseman et al., 2012) with the estimate Cell Counts function in the Minfi package (Jaffe and Irizarry, 2014) in R (Team, 2020) to estimate relative proportions of six white blood cell subtypes (CD4+ T-lymphocytes, CD8+ T-lymphocytes, natural killer (NK) cells, B-lymphocytes, monocytes and granulocytes).

## Generation R Study

*Study Population*

The GenerationR Study is a prospective birth cohort from fetal life to young adulthood, based in Rotterdam, the Netherlands (Jaddoe et al., 2006; Kooijman et al., 2016). Pregnant women who lived in the Rotterdam area and had a delivery date between April 2002 and January 2006 were recruited by participating midwives and obstetricians. While the study aimed to recruit women during early pregnancy, women were allowed to enroll at any point during pregnancy, or in the first months after birth during routine visits to child health centers. In total 9,778 mothers were enrolled, 8,880 of whom were enrolled during pregnancy. Recruitment and characteristics of the cohort have been described in detail elsewhere (Jaddoe et al., 2006; Kooijman et al., 2016). Participants were only eligible for methylation analysis if they were additionally part of the Generation R Focus cohort, a subset of the study selected characterized by Dutch ethnicity and a high level of completeness of collected data. For the current study, we restricted the sample to mother –child pairs with complete data on maternal mid-pregnancy vitamin D, offspring DNA methy-

lation, and all covariates, resulting in an analytic sample of 1,154 mother-child pairs.

*Maternal mid-pregnancy Vitamin D*

Maternal vitamin D concentrations were measured in serum samples taken in weeks 18.1-24.9 of gestation. Details of the collection procedure have been described in detail elsewhere (Vinkhuyzen et al., 2016). Briefly, 50 $\mu$L milli-Q water and 50 $\mu$L of acetonitrile (ACN) containing 6,19,19-[2H3]-25OHD2 and 6,19,19-[2H3]-25OHD3 at 10 nmol/L each were added to 3 $\mu$L plasma, sonicated, vortexed and centrifuged. The supernatant was filtered using a TiO2/ZrO2 filter plate (Glygen, USA) and evaporated to dryness. Samples were derivatised using 4-phenyl-1,2,4-triazoline-3,5-dione (PTAD) and reconstituted in ACN:H2O (1:3) prior to analysis. Samples were quantified using isotope dilution liquid chromatography-tandem mass spectrometry. The analytical system was comprised of a Shimadzu Nexera UPLC coupled to an AbSciex 5500 QTRAP equipped with an APCI source. Chromatographic separation was achieved using a Kinetex XB-C18 column (50 * 2.1 mm, 1.7 $\mu$m; Phenomenex, USA), and 72% acetonitrile/32% aqueous 0.1% formic acid at a flow rate of 0.5 mL/min. Total Serum 25(OH)D was calculated as the sum of Serum 25(OH)D$_2$ and 25(OH)D$_3$.

*Offspring DNA methylation*

Directly after delivery, obstetricians and midwives collected a maximum of 30 ml cord blood from the umbilical vein. DNA extraction from all children using the Qiagen Flexigene Kit (Qiagen Hilden, Germany)(Miller and Dd, 1988). In a subgroup of 1339 Generation R children of Dutch ancestry, 500 ng DNA per sample underwent bisulfite conversion using the EZ-96 DNA Methylation kit (Shallow) (Zymo Research Corporation, Irvine, USA). Samples were plated onto 96-well plates in no specific order. Samples were processed with the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA), which analyses methylation at 485,577 CpG sites. Preparation and normalization of the HumanMethylation450 BeadChip array data was performed according to the CPACOR workflow1 using the software package R2. In detail, the idat files were read using the minfi package. Probes that had a detection p-value above background (based on sum of methylated and unmethylated intensity values) $\geq$ 1E-16 were set to missing per array. Next, the intensity values were stratified by autosomal and non-autosomal probes and quantile normalized for each of the six probe type categories separately: type II red/green, type I methylated red/green and type I unmethylated red/green. Beta values were calculated as proportion of methylated intensity value on the sum

of methylated+unmethylated+100 intensities. Arrays with observed technical problems such as failed bisulfite conversion, hybridization or extension, as well as arrays with a mismatch between sex of the proband and sex determined by the chr X and Y probe intensities were removed from subsequent analyses. Additionally, only arrays with a call rate > 95% per sample were processed further. Probes on the X and Y chromosomes were excluded from the dataset. The final dataset contained information on 458,563 CpGs.

*Covariates*

Fetal sex, maternal smoking (categorized as no smoking during pregnancy, stopped smoking in early pregnancy, or continued smoking in pregnancy), maternal age at birth, maternal education (categorized as no education, primary education, secondary education (phase 1), secondary education (phase 2), higher education (phase 1), or higher education (phase 2)) , parity, and prepregnancy BMI were assessed via self-report questionnaire during pregnancy. Cell type correction was applied using the reference-based Houseman method3 in the minfi package4 in R2, using the cord blood-specific Bakulski reference (Bakulski et al., 2016). Because the subsample of Generation R data with available methylation data was entirely of European ancestry, principal components were calculated using only Generation R children of European ancestry, and the first 10 principal components were included as covariates in the models.

## MoBa
*Study population*

The Norwegian Mother, Father, and Child Study (MoBa) is a prospective population based birth cohort conducted by the Norwegian Institute of Public Health. Between 1999 and 2008, all pregnant women in Norway were invited to take part in MoBa via postal invitation distributed after their routine ultrasound examination at 17-18 week's gestation, of which 40.6% agreed to participate. Detailed information is available elsewhere (Magnus et al., 2016). The cohort contains linked data on 114,500 children, 95,200 mothers, and 75,200 fathers. MoBa1 and MoBa2 are two subsets of the total MoBa sample on which methylation data were obtained. MoBa1 consists of a case-control sample of 3,000 randomly drawn MoBa children born between July 2002 and July 2003, along with all MoBa children born between July 2002 and July 2004 whose mother reported they had received a diagnosis of asthma by age 3 and were using an inhalation medication for asthma at age 3, and who had remained in the study through age 3 (Håberg et al., 2011). MoBa2 consisted of a second sample of 685 MoBa children with complete data on maternal plasma folate during pregnancy (Joubert et al., 2016). The two samples were analyzed separately.

For the current analysis, we restricted samples to individuals with complete data on maternal vitamin D during pregnancy, offspring methylation, and all covariates, resulting in analytic samples of 783 and 177 for MoBa1 and MoBa2, respectively.

*Maternal Mid-pregnancy Vitamin D*

25OHD concentrations were measured in plasma from blood samples taken at the 17th-18th week of pregnancy (Rønningen et al., 2006). Maternal plasma levels of 25-hydroxyvitamin $D_3$ and 25-hydroxyvitamin $D_2$ were analyzed using a liquid chromatography-tandem mass spectrometry method (LC–MS/MS) at the BEVITAL laboratory. BEVITAL is approved by the Vitamin D External Quality Assurance Scheme. The sum of 25-hydroxyvitamin $D_3$ and -$D_2$, termed 25(OH)D, was used in the analysis. Further details on blood collection, storage, and measurement of 25(OH)D are available elsewhere (Magnus et al., 2013).

*Offspring cord blood DNA methylation*

Details of the assessment of cord blood DNA methylation in both MoBa1 and MoBa2 have been described previously (Håberg et al., 2011; Joubert et al., 2016). Cord blood samples were collected at birth and frozen at -80 degrees C (Rønningen et al., 2006). Bisulfite conversion was performed using the EZ-96 DNA Methylation kit (Zymo Research Corporation, Irvine, CA) and DNA methylation was measured at 485,577 CpGs in cord blood using Illumina's Infinium HumanMethylation450 BeadChip47. Raw intensity (.idat) files were handled in R using the minfi package to calculate the methylation level at each CpG as the beta-value ($\beta$=intensity of the methylated allele (M)/(intensity of the unmethylated allele (U) + intensity of the methylated allele (M) + 100)) and the data were exported for quality control and processing. Probe and sample-specific quality control was performed in the MoBa1 and MoBa2 datasets separately. Control probes (N=65) and probes on X (N=11,230) and Y (N=416) chromosomes were excluded in the datasets. Remaining CpGs missing > 10% of methylation data were also removed (none in MoBa2). Samples indicated by Illumina to have failed or or have an average detection p-value across all probes < 0.05 (N=35 MoBa2) and samples with gender mismatch (N=8 MoBa2) were also removed. As for MoBa1 we accounted for the different probe designs by applying the intra-array normalization strategy Beta Mixture Quantile dilation (BMIQ) (Teschendorff et al., 2013). The Empirical Bayes method via ComBat was applied separately in each dataset for batch correction using the sva package in R (Leek et al., 2012). Finally four samples determined to be ancestry outliers based on the principle components analysis of Illumina HumanCore genotype data were excluded from the analysis.

*Covariates*

Fetal sex, maternal smoking status, maternal age at birth, maternal pre-pregnancy BMI, maternal education, and parity were assessed via maternal questionnaire during pregnancy or from birth registry (Magnus et al., 2006). Maternal age was included as a continuous variable. Maternal smoking status during pregnancy was classified into three groups: non-smoker, stopped smoking in early pregnancy , and smoked throughout pregnancy. Maternal educational level was categorized into four groups based on years of education: less than high school/secondary school, high school/secondary school completion, some college or university, or 4 years of college/university or more. Estimation of six different white blood cell types (CD8+ T and CD4+ T lymphocytes, CD56+ natural killer cells, CD19+ B cells, CD14+ monocytes, and granulocytes) by Houseman method (Jaffe and Irizarry, 2014) was performed using the default implementation of the estimateCellCounts function in the minfi package (Team, 2020).

## PREDO

*Study Population*

The Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction (PREDO) study is prospective, multicenter longitudinal pregnancy cohort. The study recruited women with a singleton, intrauterine pregnancy who visited any of 10 study hospitals in Finland for an ultrasound screening at 12-13 weeks gestation between 2006 and 2010. Recruitment and characteristics of the cohort have been described elsewhere (Girchenko, Lahti, Tuovinen, et al., 2017). The sample is comprised of two subsamples, one of which recruited women with a known risk factor for preeclampsia and intrauterine growth restriction, and one of which recruited women regardless of risk factor status. To be eligible for the high-risk subsample, women must have had one of the following: preeclampsia in previous pregnancy, intrauterine growth restriction in previous pregnancy, gestational diabetes in previous pregnancy, pre-pregnancy obesity, chronic hypertension, Type 1 Diabetes, maternal age at birth < 20 years, maternal age at birth > 40 years, systemic lupus erythematosus, Sjogren's syndrome, previous pregnancy with fetal demise at > 22 weeks gestation or over 500g fetal weight. 110 women without any known risk factors were included to provide a normal pregnancy reference for blood samples. Exclusion criteria for the high-risk subsample included asthma diagnosed by a physician, allergy to ASA, tobacco smoking during pregnancy, previous peptic ulcer, previous placental ablation, inflammatory bowel disease, rheumatoid arthritis, haemophilia or thrombophilia, and multiple pregnancy.

Of 5,332 recruited women, 4,785 were eligible and consented to participate. Of these, 1,083 were part of the high risk subsample, and 3,702 were from the general subsample. 4,777 of these pregnancies resulted in a live birth. For the current study, we restricted the sample to mother-child pairs with complete data on maternal mid-pregnancy Vitamin D, offspring DNA methylation, and all covariates, resulting in a total analytic sample of 301 mother-child dyads.

*Maternal mid-pregnancy Vitamin D*

Maternal 25(OH)D levels were measured from maternal serum samples taken at 14.43 to 22.86 weeks of gestation. 25(OH)D concentrations were measured with a fully automated IDS-iSYS analyzer (Immunodiagnostic Systems Ltd., Bolton, UK). The method was validated against liquid chromatography tandem mass spectrometry (LC-MS/MS) in house, as well as by the manufacturer. The two have good linear agreement, though the method used by PREDO gives 0.72-fold lower results. Intra- and inter-assay CV% were <5% and 7%. The quality and accuracy of the serum 25(OH)D analysis in PREDO is validated on an ongoing basis by participation in the vitamin D External Quality Assessment Scheme (DEQAS, Charing Cross Hospital, London, UK). iSYS shows a 3% positive bias against all laboratory trimmed mean values and a 10% positive bias compared with NIST standards in international comparisons Within the sample, levels ranged between 26.80 and 139.9 nmol/L (mean: 69.58, SD: 19.04).

*Offspring DNA methylation*

Epigenome-wide methylation in cord blood samples was assessed using Illumina 450K microarrays. We randomized all samples on 96-well plates based on gender and maternal risk factors. Quality control was conducted using the R package minfi. Samples were excluded if they were duplicates, outliers in median intensities, or sex discrepant based on X and Y chromosomes. Probes on X or Y chromosomes, probes containing SNPs, cross-hybridizing probes, and CpG sites with low detection p-values in at least 50% of samples were also removed from the analysis. Maternal blood contamination was tested using methylation data at 10 CpGs independently identified as differentially methylated between cord and adult blood and indicative of maternal blood contamination. Samples with DNA methylation values above previously identified thresholds at >4 of the 10 sites were considered contaminated and removed from all future analyses. The final dataset contained data on 428,619 CpG sites. Betas were normalized using the funnorm function, incorporating the first 10 principal components from internal control probes. To check for batch effects, principal components were computed on these normalized betas. Two batches were significantly associated with main principal components and removed iteratively

using the combat package. Additional details on these procedures are available elsewhere (Girchenko, Lahti, Czamara, et al., 2017). Cell type proportions were estimated using the Bakulski reference set (Bakulski et al., 2016).

*Covariates*

Gestational age at measurement of Vitamin D, fetal sex, season of vitamin D measurement, parity, and maternal age at delivery were assessed using data from the Finnish Medical Birth Register and Population Register. Maternal smoking was assessed using Finnish Medical Birth register data, and was categorized into 3 levels; no smoking during pregnancy, quit smoking in the first trimester, continued smoking through pregnancy. Maternal early-pregnancy BMI was based on weight and height measured at the first antenatal clinic visit (mean 8 weeks gestation) derived from the Finnish Medical Birth Register. Maternal education was assessed using self-report questionnaire at 12-13 weeks gestation, and was classified into primary education, secondary education, lower tertiary education, or upper tertiary education, as recommended by Statistics Finland. Ancestry was evaluated using offspring GWAS data, and 2 principal components with eigenvalues > 1 were included in the models. As PREDO is a highly ethnically homogenous sample of Finnish-speaking mothers from Southern Finland, this is likely sufficient to adjust for population stratification.

## Project Viva

*Study Population*

Project Viva is a longitudinal pre-birth cohort established to examine the effects of events during early development on lifetime health outcomes (Oken et al., 2015). Recruitment and characteristics of the cohort have been described in detail elsewhere (Oken et al., 2015). Between April 1999 and November 2002, the study recruited women in early pregnancy from eight obstetric of Atrius Harvard Vanguard Medical Associates, a multispecialty group practice in eastern Massachusetts. Exclusion criteria included multiple gestation, inability to answer questions in English, gestational age >= 22 weeks at recruitment, and plans to move away from the study area before delivery. Of 2670 enrolled participants, 2128 were still enrolled at delivery and had a live birth. For the current study, we restricted to mother-child pairs of white race/ethnicity with complete data on mid-pregnancy Vitamin D, offspring DNA methylation, and non-missing covariates resulting in a total analytic sample of 283 mother-child pairs.

*Maternal mid-pregnancy Vitamin D*

Vitamin D concentrations were assessed in plasma samples collected at 23.8-36.4 weeks gestation. Blood samples were initially refrigerated, then plasma was separated and stored at -80 degrees C. Samples were analyzed in duplicate for 25(OH)D concentration, using an automated chemiluminescence immunoassay (Ersfeld et al., 2004) and a manual radioimmunoassay (Hollis et al., 1993). Values from the two assays were averaged to obtain more stable estimates of 25(OH)D level (Burris et al., 2014).

*Offspring Cord Blood Methylation*

Cord blood samples collected at birth were centrifuged within 24 hours of collection. Genomic DNA was extracted from nucleated cells using commercially available PureGene Kits (Fisher, Catalog Nos. A407-4, A416-4; Qiagen, Catalog Nos.158908, 158912, 158924), and frozen at -80 degrees C. Extracted DNA underwent bisulfite conversion using the Zymo EZ DNA Methylation kit (Zymo Research), and epigenome wide methylation was measured using the Illumina HM450K microarray. Data were preprocessed using the minfi package in R. Failed samples, replicates, non-CpG probes, and probes on X and Y chromosomes were removed. Data were checked for gender mismatch using X and Y chromosomes. CpG sites with low detection p-values were identified and flagged. Raw methylation values were Noob adjusted (background and dye bias adjusted), and methylation values were normalized using a beta-mixture quantile normalization method.

*Covariates*

Analyses in Project Viva were restricted to subjects of self-reported white race. Fetal sex, maternal smoking status, maternal age at enrollment, early pregnancy BMI, maternal education, and parity were assessed by self-administered questionnaires and interviews during pregnancy. Season of vitamin D measurement was categorized into 4 groups (Feb-Apr, May-Jul, Aug-Oct, Nov-Jan).

# Supplementary Results
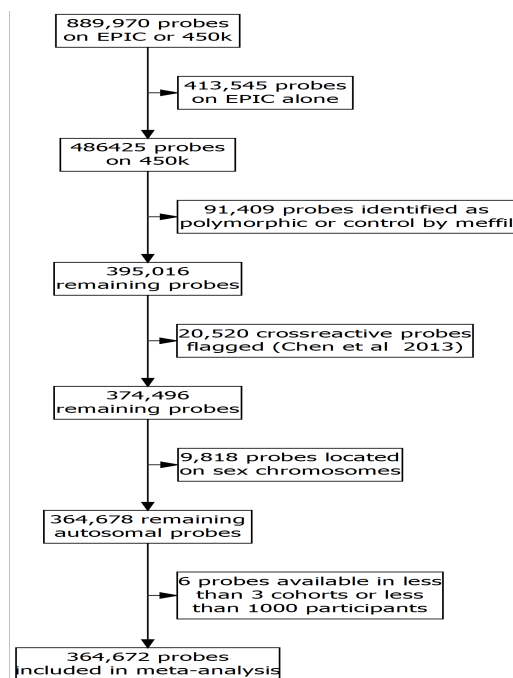


Figure 2.3: Flowchart describing probe inclusion

Table 2.2: Lambdas for each participating cohort (Base Model)

| Cohort | n | $\lambda$ |
|---|---|---|
| EAGeR | 361 | 1.044 |
| MoBa1 | 783 | 1.01 |
| MoBa2 | 177 | 1.465 |
| PREDO | 301 | 0.85 |
| Gen3G | 175 | 1.321 |
| Generation R | 1154 | 1.033 |
| Project Viva | 283 | 0.929 |

Table 2.3: Lambdas for each participating cohort (Season of Measurement Model)

| Cohort | n | $\lambda$ |
|---|---|---|
| ALSPAC | 499 | 0.981 |
| EAGeR | 361 | 1.047 |
| MoBa1 | 783 | 0.861 |
| MoBa2 | 177 | 1.157 |
| PREDO | 301 | 0.817 |
| Gen3G | 175 | 1.126 |
| Generation R | 1154 | 0.841 |
| Project Viva | 283 | 0.958 |
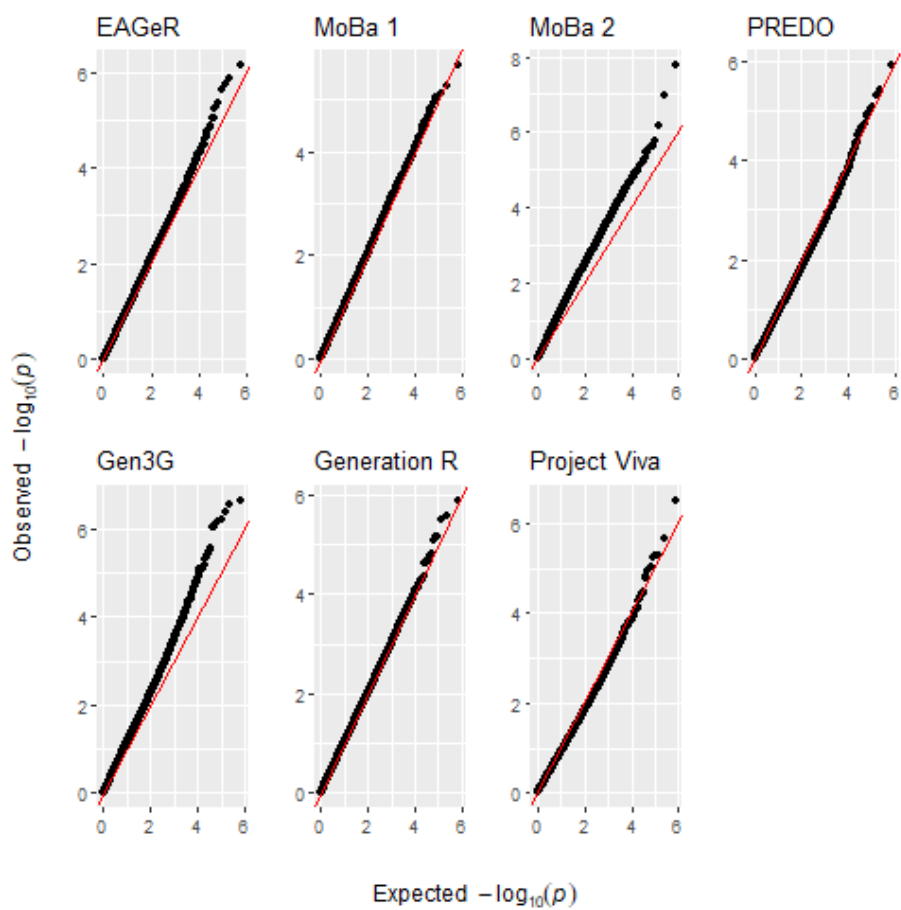
Figure 2.4: Q-Q plots for each participating cohort (Base Model)
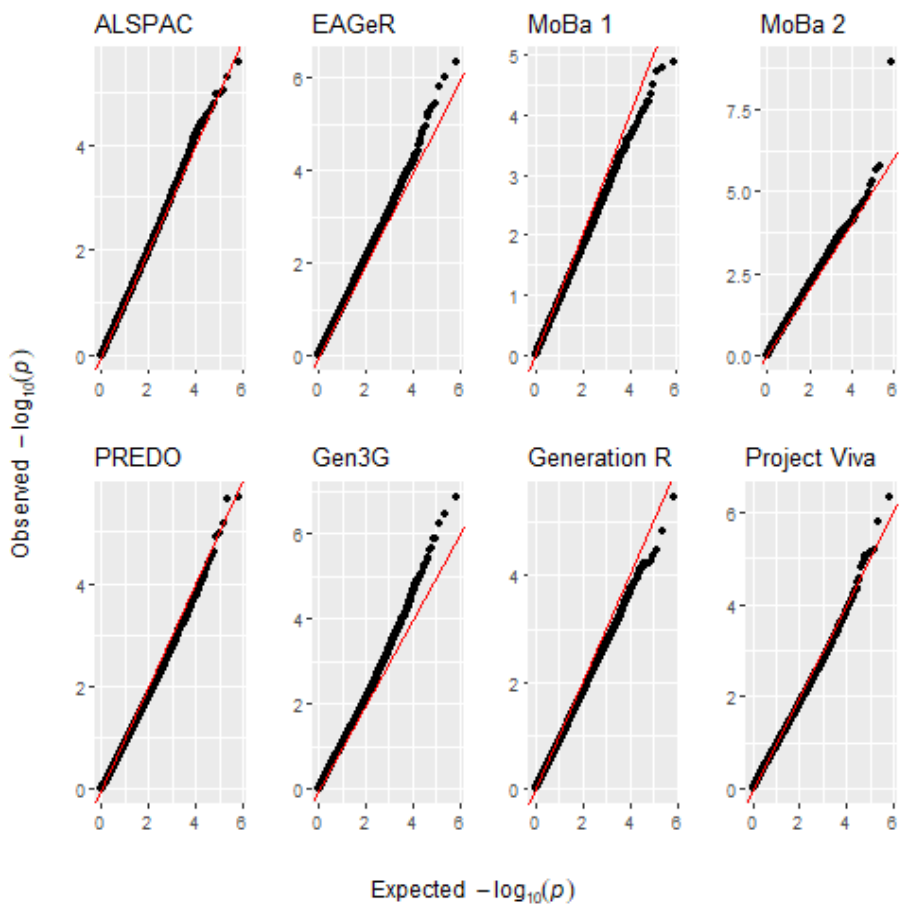
Figure 2.5: Q-Q plots for each participating cohort (Season of Measurement Model)
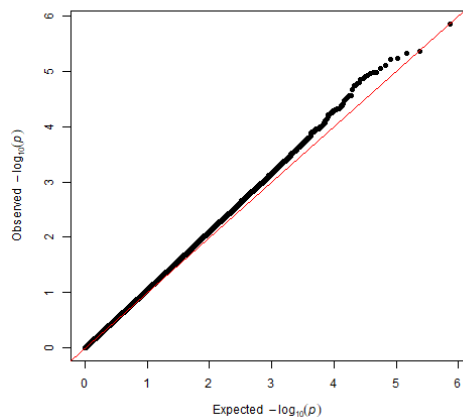
Figure 2.6: Q-Q plot for full meta-analytic results (Base Model)
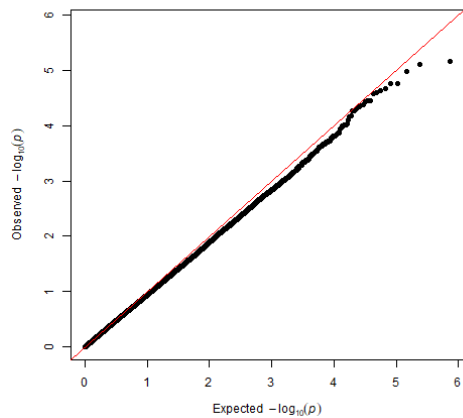


Figure 2.7: Q-Q plots for full meta-analytic results (Season of Measurement Model)

59
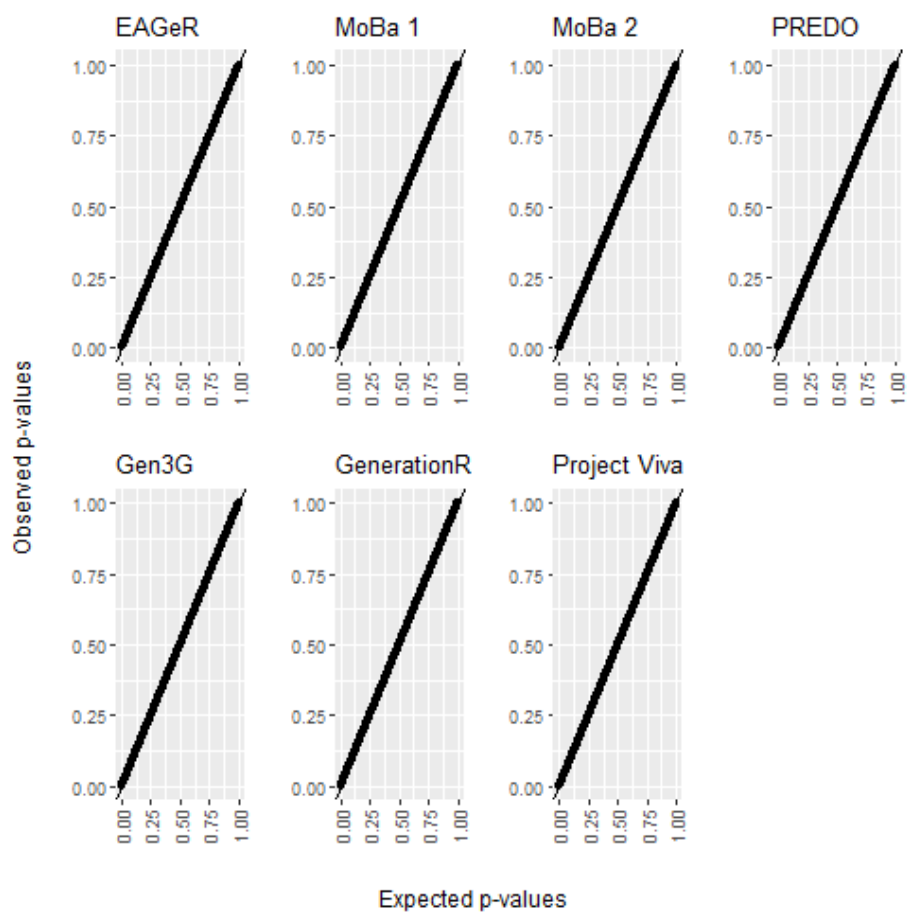
Figure 2.8: P-Z plots for each participating cohort (Base Model)
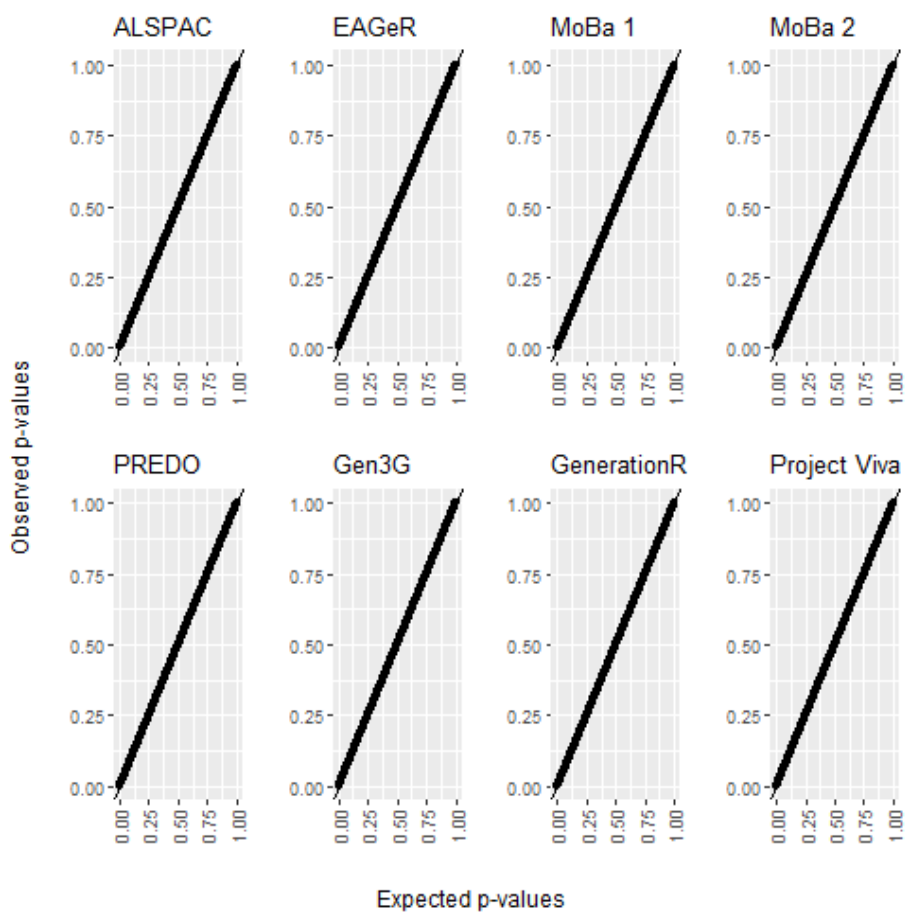
Figure 2.9: P-Z plots for each participating cohort (Season of Measurement Model)
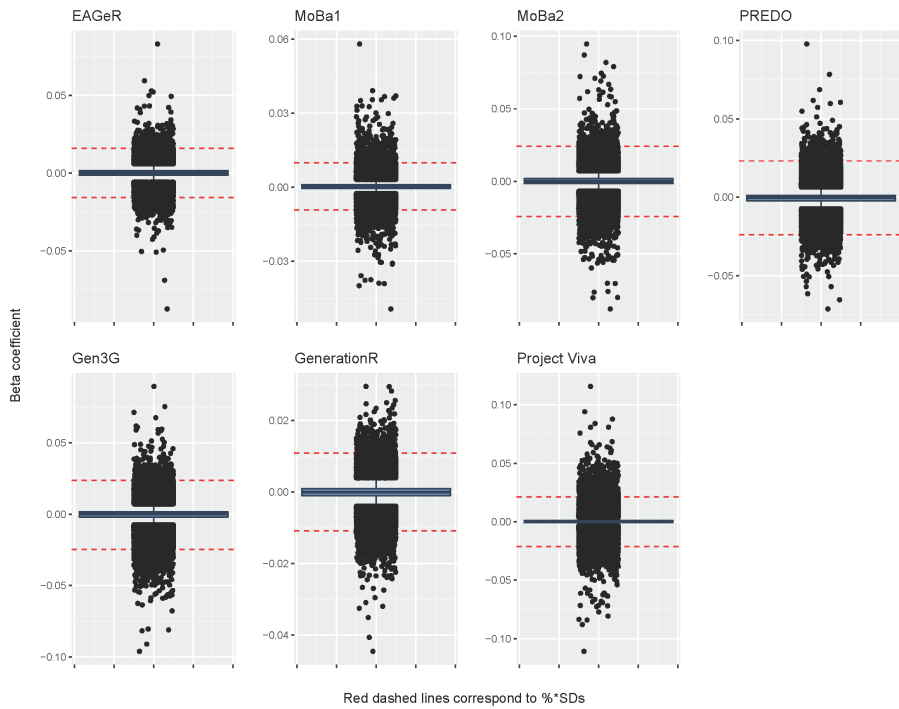
Figure 2.10: Boxplots of Betas for each participating cohort (Base Model)

Red dashed lines correspond to 5*SDs
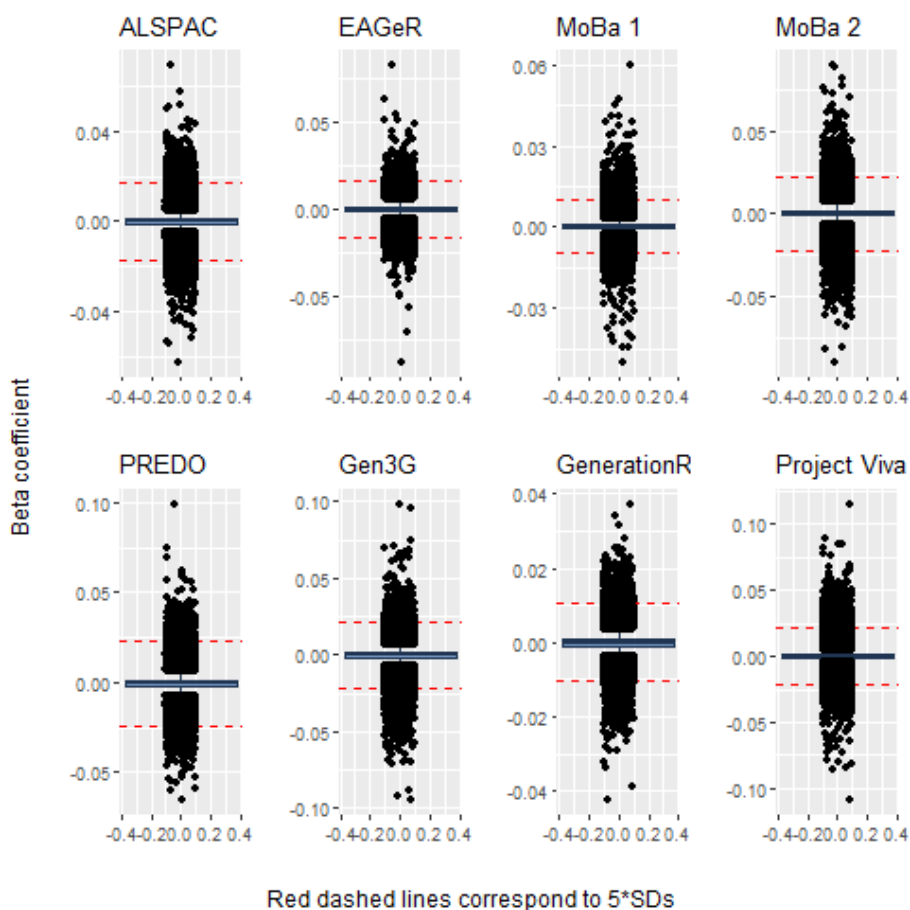
Figure 2.11: Boxplots of Betas for each participating cohort (Season of Measurement Model)
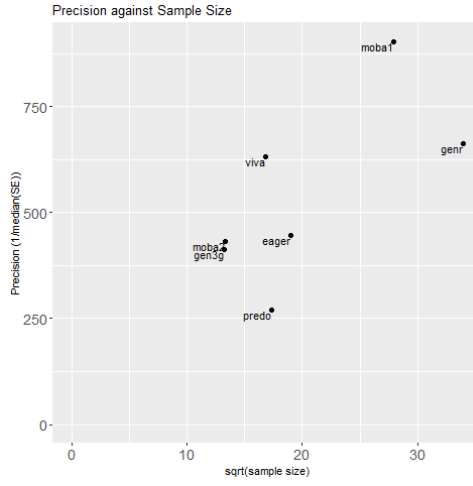
Figure 2.12: Plot of precision relative to sample size (Base Model)
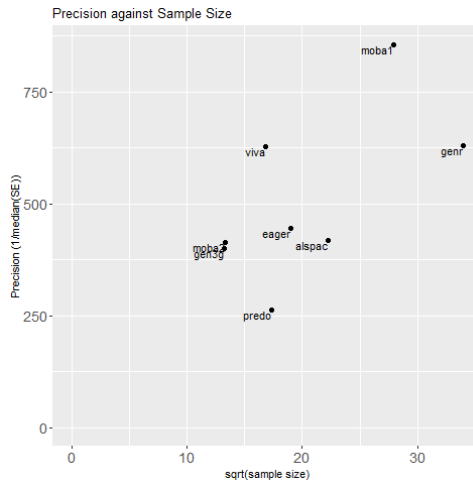


Figure 2.13: Plot of precision relative to sample size (Season of Measurement Model)

64
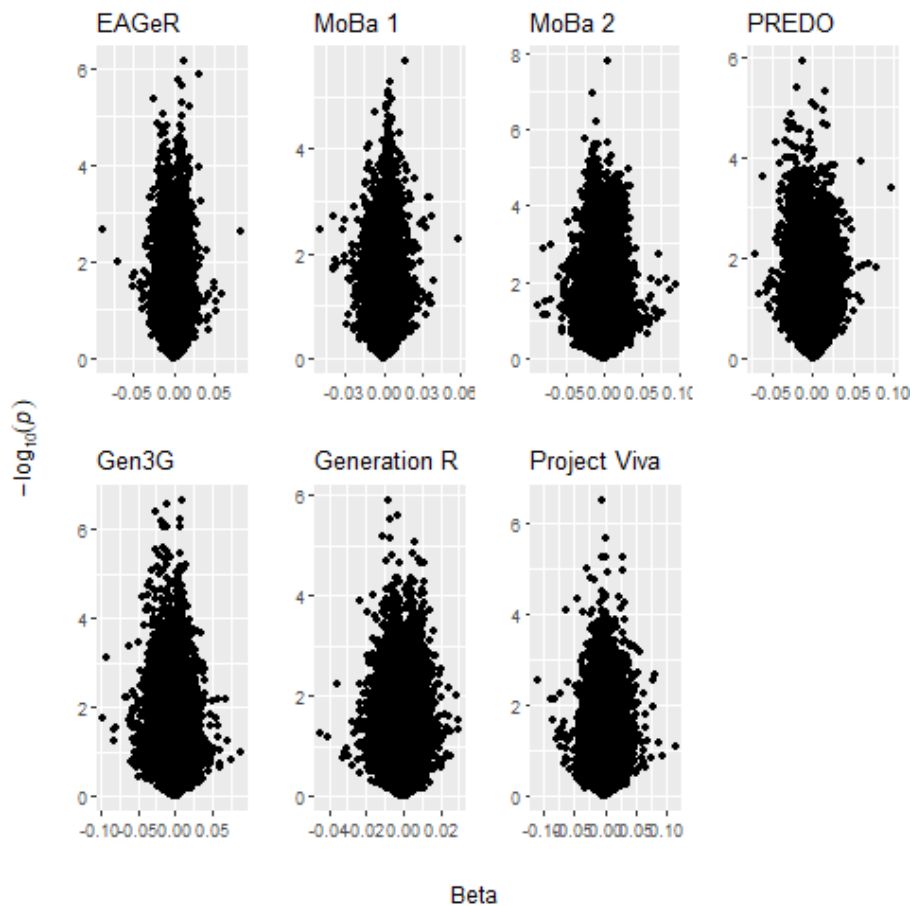
Figure 2.14: Volcano plots for each participating cohort (Base Model)
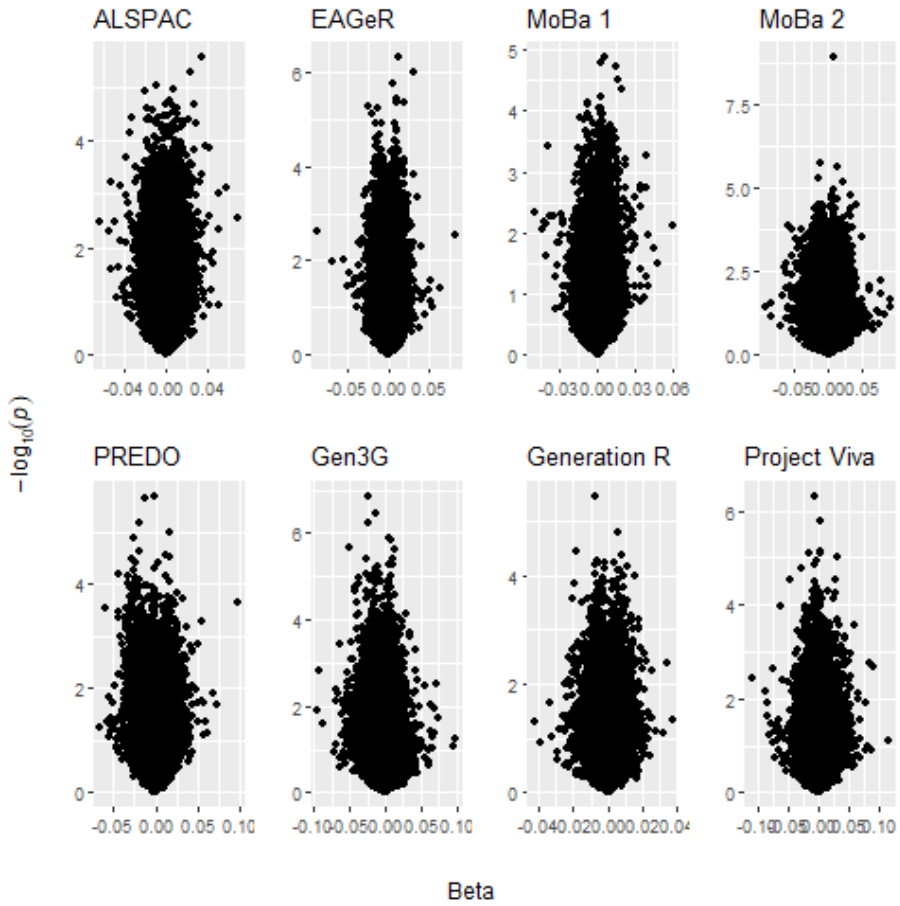
Figure 2.15: Volcano plots for each participating cohort (Season of Birth Model)

# References

Bakulski, K. M., Feinberg, J. I., Andrews, S. V., Yang, J., Brown, S., L. McKenney, S., Witter, F., Walston, J., Feinberg, A. P., & Fallin, M. D. (2016). Dna methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics*, *11*(5), 354–362.

Beckett, E. L., Duesing, K., Martin, C., Jones, P., Furst, J., King, K., Niblett, S., Yates, Z., Veysey, M., & Lucock, M. (2016). Relationship between methylation status of vitamin d-related genes, vitamin d levels, and methyl-donor biochemistry. *Journal of Nutrition and Intermediary Metabolism*, *6*, 8–15.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Bodnar, L. M., Platt, R. W., & Simhan, H. N. (2015). Early-pregnancy vitamin d deficiency and risk of preterm birth subtypes. *Obstetrics and gynecology*, *125*(2), 439.

Bodnar, L. M., Simhan, H. N., Powers, R. W., Frank, M. P., Cooperstein, E., & Roberts, J. M. (2007). High prevalence of vitamin d insufficiency in black and white pregnant women residing in the northern united states and their neonates. *The Journal of nutrition*, *137*(2), 447–452.

Boghossian, N. S., Koo, W., Liu, A., Mumford, S. L., Tsai, M. Y., & Yeung, E. H. (2019). Longitudinal measures of maternal vitamin d and neonatal body composition. *European journal of clinical nutrition*, *73*(3), 424–431.

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., & Davey Smith, G. (2013). Cohort profile: The 'children of the 90s'—the index offspring of the avon longitudinal study of parents and children. *International journal of epidemiology*, *42*(1), 111–127.

Burris, H. H., Rifas-Shiman, S. L., Huh, S. Y., Kleinman, K., Litonjua, A. A., Oken, E., Rich-Edwards, J. W., Camargo Jr, C. A., & Gillman, M. W. (2014). Vitamin d status and hypertensive disorders in pregnancy. *Annals of epidemiology*, *24*(5), 399–403. e1.

Busche, S., Shao, X., Caron, M., Kwan, T., Allum, F., Cheung, W. A., Ge, B., Westfall, S., Simon, M.-M., & Barrett, A. (2015). Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome biology*, *16*(1), 1–18.

Cashman, K. D., Dowling, K. G., Škrabáková, Z., Kiely, M., Lamberg-Allardt, C., Durazo-Arvizu, R. A., Sempos, C. T., Koskinen, S., Lundqvist, A., & Sundvall, J. (2015). Standardizing serum 25-hydroxyvitamin d data from four nordic population samples using the vitamin d standardization program protocols: Shedding new light on vitamin d status in nordic individuals. *Scandinavian journal of clinical and laboratory investigation*, *75*(7), 549–561.

Chen, Y.-a., Lemire, M., Choufani, S., Butcher, D. T., Grafodatskaya, D., Zanke, B. W., Gallinger, S., Hudson, T. J., & Weksberg, R. (2013). Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray. *Epigenetics*, *8*(2), 203–209.

Crider, K. S., Yang, T. P., Berry, R. J., & Bailey, L. B. (2012). Folate and dna methylation: A review of molecular mechanisms and the evidence for folate's role. *Advances in nutrition*, *3*(1), 21–38.

De-Regil, L. M., Palacios, C., Lombardo, L. K., & Peña-Rosas, J. P. (2016). Vitamin d supplementation for women during pregnancy. *Cochrane Database of Systematic Reviews*, (1).

Erkkola, M., Nwaru, B. I., & Viljakainen, H. T. (2011). Maternal vitamin d during pregnancy and its relation to immune-mediated diseases in the offspring. *Vitamins and hormones* (pp. 239–260). Elsevier.

Ersfeld, D. L., Rao, D. S., Body, J.-J., Sackrison Jr, J. L., Miller, A. B., Parikh, N., Eskridge, T. L., Polinske, A., Olson, G. T., & MacFarlane, G. D. (2004). Analytical and clinical validation of the 25 oh vitamin d assay for the liaison® automated analyzer. *Clinical biochemistry*, *37*(10), 867–874.

Felix, J. F., Joubert, B. R., Baccarelli, A. A., Sharp, G. C., Almqvist, C., Annesi-Maesano, I., Arshad, H., Baïz, N., Bakermans-Kranenburg, M. J., & Bakulski, K. M. (2018). Cohort profile: Pregnancy and childhood epigenetics (pace) consortium. *International journal of epidemiology*, *47*(1), 22–23u.

Fetahu, I. S., Höbaus, J., & Kállay, E. (2014). Vitamin d and the epigenome. *Frontiers in physiology*, *5*, 164.

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., Greenwood, C. M. T., & Hansen, K. D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome biology*, *15*(11), 503.

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., & Ness, A. (2013). Cohort profile: The avon longitudinal study of parents and children:

Alspac mothers cohort. *International journal of epidemiology*, *42*(1), 97–110.

Ginde, A. A., Sullivan, A. F., Mansbach, J. M., & Camargo Jr, C. A. (2010). Vitamin d insufficiency in pregnant and nonpregnant women of childbearing age in the united states. *American journal of obstetrics and gynecology*, *202*(5), 436. e1–436. e8.

Girchenko, P., Lahti, J., Czamara, D., Knight, A. K., Jones, M. J., Suarez, A., Hämäläinen, E., Kajantie, E., Laivuori, H., & Villa, P. M. (2017). Associations between maternal risk factors of adverse pregnancy and birth outcomes and the offspring epigenetic clock of gestational age at birth. *Clinical epigenetics*, *9*(1), 49.

Girchenko, P., Lahti, M., Tuovinen, S., Savolainen, K., Lahti, J., Binder, E. B., Reynolds, R. M., Entringer, S., Buss, C., & Wadhwa, P. D. (2017). Cohort profile: Prediction and prevention of preeclampsia and intrauterine growth restriction (predo) study. *International journal of epidemiology*, *46*(5), 1380–1381g.

Guillemette, L., Allard, C., Lacroix, M., Patenaude, J., Battista, M.-C., Doyon, M., Moreau, J., Ménard, J., Bouchard, L., & Ardilouze, J.-L. (2016). Genetics of glucose regulation in gestation and growth (gen3g): A prospective prebirth cohort of mother–child pairs in sherbrooke, canada. *BMJ open*, *6*(2), e010031.

Håberg, S. E., London, S. J., Nafstad, P., Nilsen, R. M., Ueland, P. M., Vollset, S. E., & Nystad, W. (2011). Maternal folate levels in pregnancy and asthma in children at age three years. *The Journal of allergy and clinical immunology*, *127*(1), 262.

Han, B., & Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics*, *88*(5), 586–598.

Hernan, M. A., Hernandez-Diaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 615–625.

Hernan, M. A., & Robins, J. M. (2018). *Causal inference.* Chapman; Hall/CRC.

Hernán, M. A. (2018). The c-word: Scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, *108*(5), 616–619.

Holick, M. F. (2006). Resurrection of vitamin d deficiency and rickets. *The journal of clinical investigation*, *116*(8), 2062–2072.

Holick, M. F., & Chen, T. C. (2008). Vitamin d deficiency: A worldwide problem with health consequences. *The American journal of clinical nutrition*, *87*(4), 1080S–1086S.

Hollis, B. W., Kamerud, J. Q., Selvaag, S. R., Lorenz, J. D., & Napoli, J. L. (1993). Determination of vitamin d status by radioimmunoassay with an 125i-labeled tracer. *Clinical chemistry*, *39*(3), 529–533.

Hollis, B. W. (2005). Circulating 25-hydroxyvitamin d levels indicative of vitamin d sufficiency: Implications for establishing a new effective dietary intake recommendation for vitamin d. *The Journal of nutrition*, *135*(2), 317–322.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., & Kelsey, K. T. (2012). Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, *13*(1), 1–16.

Jaddoe, V. W. V., Mackenbach, J. P., Moll, H. A., Steegers, E. A. P., Tiemeier, H., Verhulst, F. C., Witteman, J. C. M., & Hofman, A. (2006). The generation r study: Design and cohort profile. *European journal of epidemiology*, *21*(6), 475.

Jaffe, A. E., & Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology*, *15*(2), 1–9.

Javaid, M. K., Crozier, S. R., Harvey, N. C., Gale, C. R., Dennison, E. M., Boucher, B. J., Arden, N. K., Godfrey, K. M., Cooper, C., & Princess Anne Hospital Study, G. (2006). Maternal vitamin d status during pregnancy and childhood bone mass at age 9 years: A longitudinal study. *The Lancet*, *367*(9504), 36–43.

Johnson, D. D., Wagner, C. L., Hulsey, T. C., McNeil, R. B., Ebeling, M., & Hollis, B. W. (2011). Vitamin d deficiency and insufficiency is common during pregnancy. *American journal of perinatology*, *28*(01), 007–012.

Jones, G., Strugnell, S. A., & DeLuca, H. F. (1998). Current understanding of the molecular actions of vitamin d. *Physiological reviews*, *78*(4), 1193–1231.

Joubert, B. R., Herman, T., Felix, J. F., Bohlin, J., Ligthart, S., Beckett, E., Tiemeier, H., Van Meurs, J. B., Uitterlinden, A. G., & Hofman, A. (2016). Maternal plasma folate impacts differential dna methylation in an epigenome-wide meta-analysis of newborns. *Nature communications*, *7*(1), 1–8.

Kasahara, A. K., Singh, R. J., & Noymer, A. (2013). Vitamin d (25ohd) serum seasonality in the united states. *PloS one*, *8*(6), e65785.

Kooijman, M. N., Kruithof, C. J., van Duijn, C. M., Duijts, L., Franco, O. H., van Ijzendoorn, M. H., de Jongste, J. C., Klaver, C. C. W., van der Lugt, A., & Mackenbach, J. P. (2016). The generation r study: Design

and cohort update 2017. *European Journal of Epidemiology*, *31*(12), 1243–1264.

Krieger, J.-P., Cabaset, S., Canonica, C., Christoffel, L., Richard, A., Schröder, T., von Wattenwyl, B. L., Rohrmann, S., & Lötscher, K. Q. (2018). Prevalence and determinants of vitamin d deficiency in the third trimester of pregnancy: A multicentre study in switzerland. *British Journal of Nutrition*, *119*(3), 299–309.

Larqué, E., Morales, E., Leis, R., & Blanco-Carnero, J. E. (2018). Maternal and foetal health implications of vitamin d status during pregnancy. *Annals of Nutrition and Metabolism*, *72*(3), 179–192.

Lawlor, D. A., Wills, A. K., Fraser, A., Sayers, A., Fraser, W. D., & Tobias, J. H. (2013). Association of maternal vitamin d status during pregnancy with bone-mineral content in offspring: A prospective cohort study. *The Lancet*, *381*(9884), 2176–2183.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882–883.

Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, *3*(9).

Liu, X., Baylin, A., & Levy, P. D. (2018). Vitamin d deficiency and insufficiency among us adults: Prevalence, predictors and clinical implications. *British Journal of Nutrition*, *119*(8), 928–936.

Lockett, G. A., Soto-Ramírez, N., Ray, M. A., Everson, T. M., Xu, C.-J., Patil, V. K., Terry, W., Kaushal, A., Rezwan, F. I., & Ewart, S. L. (2016). Association of season of birth with dna methylation and allergic disease. *Allergy*, *71*(9), 1314–1324.

Magnus, M. C., Stene, L. C., Håberg, S. E., Nafstad, P., Stigum, H., London, S. J., & Nystad, W. (2013). Prospective study of maternal mid-pregnancy 25-hydroxyvitamin d level and early childhood respiratory disorders. *Paediatric and perinatal epidemiology*, *27*(6), 532–541.

Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Daltveit, A. K., Handal, M., Haugen, M., Høiseth, G., & Knudsen, G. P. (2016). Cohort profile update: The norwegian mother and child cohort study (moba). *International journal of epidemiology*, *45*(2), 382–388.

Magnus, P., Irgens, L. M., Haug, K., Nystad, W., Skjærven, R., & Stoltenberg, C. (2006). Cohort profile: The norwegian mother and child cohort study (moba). *International journal of epidemiology*, *35*(5), 1146–1150.

Magnusson, C., Lundberg, M., Lee, B. K., Rai, D., Karlsson, H., Gardner, R., Kosidou, K., Arver, S., & Dalman, C. (2016). Maternal vitamin d

deficiency and the risk of autism spectrum disorders: Population-based study. *British Journal of Psychiatry Open*, *2*(2), 170–172.

Malacova, E., Cheang, P. R., Dunlop, E., Sherriff, J. L., Lucas, R. M., Daly, R. M., Nowson, C. A., & Black, L. J. (2019). Prevalence and predictors of vitamin d deficiency in a nationally representative sample of adults participating in the 2011–2013 australian health survey. *The British journal of nutrition*, *121*(8), 894–904.

Mansell, G., Gorrie-Stone, T. J., Bao, Y., Kumari, M., Schalkwyk, L. S., Mill, J., & Hannon, E. (2019). Guidance for dna methylation studies: Statistical insights from the illumina epic array. *BMC genomics*, *20*(1), 1–15.

Marcinowska-Suchowierska, E., Kupisz-Urbańska, M., Łukaszkiewicz, J., Płudowski, P., & Jones, G. (2018). Vitamin d toxicity–a clinical perspective. *Frontiers in endocrinology*, *9*, 550.

McCartney, D. L., Walker, R. M., Morris, S. W., McIntosh, A. M., Porteous, D. J., & Evans, K. L. (2016). Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationepic beadchip. *Genomics data*, *9*, 22–24.

Miller, S. A., & Dd, D. (1988). Polesky hf. a simple salting out procedure for extracting dna from human nucleated cells. *Nucleic Acids Res*, *16*(3), 1215.

Min, J. L., Hemani, G., Davey Smith, G., Relton, C., & Suderman, M. (2018). Meffil: Efficient normalization and analysis of very large dna methylation datasets. *Bioinformatics*, *34*(23), 3983–3989.

Morales, E., Julvez, J., Torrent, M., Ballester, F., Rodriguez-Bernal, C. L., Andiarena, A., Vegas, O., Castilla, A. M., Rodriguez-Dehli, C., Tardon, A., et al. (2015). Vitamin d in pregnancy and attention deficit hyperactivity disorder-like symptoms in childhood. *Epidemiology*, *26*(4), 458–465.

Mumford, S. L., Garbose, R. A., Kim, K., Kissell, K., Kuhr, D. L., Omosigho, U. R., Perkins, N. J., Galai, N., Silver, R. M., Sjaarda, L. A., et al. (2018). Association of preconception serum 25-hydroxyvitamin d concentrations with livebirth and pregnancy loss: A prospective cohort study. *The Lancet Diabetes & Endocrinology*, *6*(9), 725–732.

Off, M. K., Steindal, A. E., Porojnicu, A. C., Juzeniene, A., Vorobey, A., Johnsson, A., & Moan, J. (2005). Ultraviolet photodegradation of folic acid. *Journal of Photochemistry and Photobiology B: Biology*, *80*(1), 47–55.

Oken, E., Baccarelli, A. A., Gold, D. R., Kleinman, K. P., Litonjua, A. A., De Meo, D., Rich-Edwards, J. W., Rifas-Shiman, S. L., Sagiv, S., &

Taveras, E. M. (2015). Cohort profile: Project viva. *International journal of epidemiology*, *44*(1), 37–48.

Palacios, C., Kostiuk, L. K., & Peña-Rosas, J. P. (2019). Vitamin d supplementation for women during pregnancy. *Cochrane Database of Systematic Reviews*, (7).

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., & Clark, S. J. (2016). Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. *Genome biology*, *17*(1), 1–17.

Pike, J. W., & Meyer, M. B. (2014). Fundamentals of vitamin d hormone-regulated gene expression. *The Journal of steroid biochemistry and molecular biology*, *144*, 5–11.

Relton, C. L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., Woodward, G., Lyttleton, O., Evans, D. M., & Reik, W. (2015). Data resource profile: Accessible resource for integrated epigenomic studies (aries). *International journal of epidemiology*, *44*(4), 1181–1190.

Rønningen, K. S., Paltiel, L., Meltzer, H. M., Nordhagen, R., Lie, K. K., Hovengen, R., Haugen, M., Nystad, W., Magnus, P., & Hoppin, J. A. (2006). The biobank of the norwegian mother and child cohort study: A resource for the next 100 years. *European journal of epidemiology*, *21*(8), 619–625.

Roth, D. E., Leung, M., Mesfin, E., Qamar, H., Watterworth, J., & Papp, E. (2017). Vitamin d supplementation during pregnancy: State of the evidence from a systematic review of randomised trials. *Bmj*, *359*, j5237.

Schisterman, E. F., Silver, R. M., Perkins, N. J., Mumford, S. L., Whitcomb, B. W., Stanford, J. B., Lesher, L. L., Faraggi, D., Wactawski-Wende, J., & Browne, R. W. (2013). A randomised trial to evaluate the effects of low-dose aspirin in gestation and reproduction: Design and baseline characteristics. *Paediatric and perinatal epidemiology*, *27*(6), 598–609.

Solomon, O., MacIsaac, J., Quach, H., Tindula, G., Kobor, M. S., Huen, K., Meaney, M. J., Eskenazi, B., Barcellos, L. F., & Holland, N. (2018). Comparison of dna methylation measured by illumina 450k and epic beadchips in blood of newborns and 14-year-old children. *Epigenetics*, *13*(6), 655–664.

Song, H., Yang, L., & Jia, C. (2017). Maternal vitamin d status during pregnancy and risk of childhood asthma: A meta-analysis of prospective studies. *Molecular Nutrition & Food Research*, *61*(5), 1600657.

Steindal, A. H., Tam, T. T. T., Lu, X. Y., Juzeniene, A., & Moan, J. (2008). 5-methyltetrahydrofolate is photosensitive in the presence of riboflavin. *Photochemical & Photobiological Sciences*, *7*(7), 814–818.

Suderman, M., Stene, L. C., Bohlin, J., Page, C. M., Holvik, K., Parr, C. L., Magnus, M. C., Håberg, S. E., Joubert, B. R., & Wu, M. C. (2016). 25-hydroxyvitamin d in pregnancy and genome wide cord blood dna methylation in two pregnancy cohorts (moba and alspac). *The Journal of steroid biochemistry and molecular biology*, *159*, 102–109.

Switkowski, K. M., Camargo Jr, C. A., Perron, P., Rifas-Shiman, S. L., Oken, E., & Hivert, M.-F. (2019). Cord blood vitamin d status is associated with cord blood insulin and c-peptide in two cohorts of mother-newborn pairs. *The Journal of Clinical Endocrinology & Metabolism*, *104*(9), 3785–3794.

Tam, T. T. T., Juzeniene, A., Steindal, A. H., Iani, V., & Moan, J. (2009). Photodegradation of 5-methyltetrahydrofolate in the presence of uroporphyrin. *Journal of Photochemistry and Photobiology B: Biology*, *94*(3), 201–204.

Team, R. C. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria.

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., & Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*, *29*(2), 189–196.

Tønnesen, R., Hovind, P. H., Jensen, L. T., & Schwarz, P. (2016). Determinants of vitamin d status in young adults: Influence of lifestyle, sociodemographic and anthropometric factors. *BMC public health*, *16*(1), 385.

Vinkhuyzen, A. A. E., Eyles, D. W., Burne, T. H. J., Blanken, L. M. E., Kruithof, C. J., Verhulst, F., Jaddoe, V. W., Tiemeier, H., & McGrath, J. J. (2016). Gestational vitamin d deficiency and autism-related traits: The generation r study. *Molecular psychiatry*.

Voortman, T., van den Hooven, E. H., Heijboer, A. C., Hofman, A., Jaddoe, V. W. V., & Franco, O. H. (2015). Vitamin d deficiency in school-age children is associated with sociodemographic and lifestyle factors. *The Journal of nutrition*, *145*(4), 791–798.

Wang, H., Xiao, Y., Zhang, L., & Gao, Q. (2018). Maternal early pregnancy vitamin d status in relation to low birth weight and small-for-gestational-age offspring. *The Journal of steroid biochemistry and molecular biology*, *175*, 146–150.

Webb, A. R. (2006). Who, what, where and when—influences on cutaneous vitamin d synthesis. *Progress in biophysics and molecular biology*, *92*(1), 17–25.

Wei, Z., Zhang, J., & Yu, X. (2016). Maternal vitamin d status and childhood asthma, wheeze, and eczema: A systematic review and meta-analysis. *Pediatric Allergy and Immunology*, *27*(6), 612–619.

Welles, C. C., Whooley, M. A., Karumanchi, S. A., Hod, T., Thadhani, R., Berg, A. H., Ix, J. H., & Mukamal, K. J. (2014). Vitamin d deficiency and cardiovascular events in patients with coronary heart disease: Data from the heart and soul study. *American journal of epidemiology*, *179*(11), 1279–1287.

# Chapter 3

# Mendelian randomization approaches to the study of prenatal exposures: a systematic review

Elizabeth W. Diemer, Jeremy A. Labrecque, Alexander Neumann, Henning Tiemeier, Sonja A. Swanson

## 3.1 Abstract

**Background**: Mendelian randomization (MR) designs apply instrumental variable techniques using genetic variants to study causal effects. MR is increasingly used to evaluate the role of maternal exposures during pregnancy on offspring health.

**Objectives**: We review the application of MR to prenatal exposures and describe reporting of methodologic challenges in this area.

**Data sources**: We searched Pubmed, Embase, Medline Ovid, Cochrane Central, Web of Science, and Google Scholar.

**Study selection and data extraction**: Eligible studies met the following criteria: (1) a maternal pregnancy exposure ; (2) an outcome assessed in offspring of the pregnancy; and (3) a genetic variant or score proposed as an instrument or proxy for an exposure.

**Synthesis**: We quantified the frequency of reporting of MR conditions stated, techniques used to examine assumption plausibility, and reported limitations.

**Results**: 43 eligible studies were identified. When discussing challenges or limitations, the most common issues described were known potential biases in the broader MR literature, including population stratification (n=29), weak instrument bias (n=18), and certain types of pleiotropy (n=30). Of 22 studies presenting point estimates for the effect of exposure, four defined their causal estimand. Twenty-four studies discussed issues unique to prenatal MR, including selection on pregnancy (n=1) and pleiotropy via postnatal exposure (n=10) or offspring genotype (n=20).

**Conclusions**: Prenatal MR studies frequently discuss issues that affect all MR studies, but rarely discuss problems specific to the prenatal context, including selection on pregnancy and effects of postnatal exposure. Future prenatal MR studies should report and attempt to falsify their assumptions, with particular attention to issues specific to prenatal MR. Further research is needed to evaluate the impacts of biases unique to prenatal MR in practice.

## 3.2   Background

Many pregnancy exposures, including maternal nutrition, substance use, and chronic health conditions, are associated with offspring adverse birth outcomes and health across the life course (Fleming et al., 2015; Linnet et al., 2003; Marques et al., 2013; Sacks et al., 2016). However, mothers who differ in specific prenatal behaviors and traits are also likely to differ in socioeconomic status and many other health behaviors, including substance use, exercise habits, diet, social support, and engagement with medical professionals, that could likewise affect or be associated with offspring outcomes (Smith, 2008). These confounders of the relationship between pregnancy exposures and offspring outcomes are complex constructs that are difficult to measure, as they often relate to an individual's latent tendency to engage in healthy behaviors or to be exposed to risk factors associated with socioeconomic position. Therefore, estimates of causal effects of exposures during pregnancy using more traditional analytic techniques that require measuring and adjusting for confounders may be biased.

Instrumental variable analysis proposing genetic variants as instruments, also known as Mendelian randomization (MR), is an alternative approach to estimate causal effects of exposures on outcomes. In prenatal MR designs, the mothers' genetic variants (e.g., single nucleotide polymorphisms [SNPs]) are proposed as instruments to examine the effect of an exposure during pregnancy on an offspring outcome. Under specific conditions, MR allows for unbiased estimation of an average causal effect of an exposure on an outcome, even in the presence of unmeasured confounding of the exposure-outcome relationship (Hernan and Robins, 2018). An MR study requires an instrument, defined as a variable that meets the following conditions:

1. The instrument $Z$ (i.e., the genetic variant) must be associated with the exposure $X$

2. The instrument $Z$ does not affect the outcome $Y$ except through its possible effect on the exposure $X$ (also known as the exclusion restriction)

3. Individuals at different levels of the instrument $Z$ are exchangeable (i.e., comparable) with regard to counterfactual outcome.

One important implication of condition 3 is that the instrument $Z$ and the outcome $Y$ cannot share any unmeasured causes. A causal structure that

meets these requirements is portrayed in Figure 3.1.
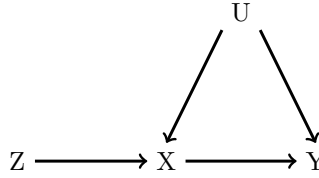


Figure 3.1: Causal DAG representing an MR study where $Z$ is a valid instrument for the effect of $X$ on $Y$.

Under these three conditions, investigators can test whether there is an effect of the exposure on the outcome for at least one individual in the study population (Swanson et al., 2018), and can estimate bounds for the average causal effect (Balke and Pearl, 1997; Robins, 1989). In order to obtain a point estimate of an average causal effect, investigators must assume one of a set of additional conditions holds. These conditions vary in strength and plausibility, and some choices of weaker conditions will produce estimates of average causal effects in unidentifiable subgroups of the study population (see Appendix for further detail). This choice of condition alters the population to which the estimated effect applies, and a subgroup average causal effect can differ dramatically from the population average causal effect. Therefore, guidelines for MR analyses recommend explicit reporting of this "fourth" point-identifying condition and the targeted effect estimand (Didelez et al., 2010; Didelez and Sheehan, 2005; Swanson and Hernán, 2013). Of further note, there are several estimators allowing for relaxation of MR conditions 2 and 3, although these require some alternative assumptions and often the availability of multiple possible instruments (Bowden et al., 2015; Bowden et al., 2016; Kang et al., 2016; Tchetgen et al., 2017).

Although the application of MR to pregnancy exposures is growing, to our knowledge, no existing study has examined the frequency of this design, or the assumptions and analytic strategies commonly employed in such applications. As guidelines for MR suggest that the key conditions need to be assessed on a case-by-case basis relative to the study design and research question (Glymour et al., 2012; Holmes et al., 2017; Swanson, 2017; Swanson and Hernán, 2013; VanderWeele et al., 2014), and prenatal MR studies present several unique challenges relative to other types of MR designs (Diemer et al., 2020; Lawlor et al., 2017), it is important to understand how prenatal MR studies report on both study-specific and general challenges to the validity and interpretation

of MR results. In addition, by identifying key areas of concern reported by researchers, we may be able to determine which sources of bias in prenatal MR are in most need of further research. Therefore, the aim of this study was to review the use of MR designs to study the effect of the prenatal environment on offspring outcomes, and to describe the nature and reporting of key potential strengths and weaknesses of the design in this context.

## 3.3   Methods

To investigate the use of MR in studies of pregnancy exposures, we searched Pubmed, Embase, Medline Ovid, Cochrane Central, Web of Science, and Google Scholar. Each database was searched from its start date to May 14, 2019. Inclusion in our review required the study met the following criteria: 1) the exposure of interest was a characteristic of the maternal environment that occurred during or proximate to pregnancy, 2) the outcome was assessed in the offspring of the pregnancy and 3) a genetic variant or genetic variant score was proposed as an instrument and used either as a proxy for an exposure or to conduct an instrumental variable analysis of the effect of exposure on outcome. The inclusion of proxy approaches is especially important for a review of prenatal MR designs, because some early studies did not conceptualize this approach as an application of previously established instrumental variable methods, but rather viewed genetic variants as unconfounded proxies for the exposure of interest. Testing the association between such a genetic variant and an outcome is equivalent to sharp null hypothesis testing in MR, and requires the same MR conditions hold (Swanson, 2017). Because birthweight is used both as a characteristic of the offspring and as a proxy for a broad range of characteristics of the prenatal environment, which complicate comparisons to MR analyses of other specific prenatal exposures, we excluded studies using birthweight as an exposure from this review. We also required that the study include analysis of real data and we eliminated any duplicate analyses. All studies were independently reviewed by two coauthors (ED & AN), and any disagreements between coauthors were resolved by third author (JL) review and discussion (see Figure 3.2). Details of the search terms and identified studies are available in the Appendix.

Authors extracted data from each included study using a form with open response fields for each data point. Data collected from eligible studies included the study exposure, study outcome, sample size, methodologic approach used,

```
┌─────────────────────────────────────────────────────────┐
│                 772 total records found                  │
│ (embase, Medline Ovid, Web of Science, Cochrane Central, │
│                     Google Scholar                       │
└─────────────────────────────────────────────────────────┘

                    ┌──────────────────────────────────────┐
                    │ 680 rejected on title/abstract review │
                    └──────────────────────────────────────┘

     ┌───────────────────────────────────────────┐
     │ 92 full text articles reviewed for eligibility │
     └───────────────────────────────────────────┘

                    ┌──────────────────────────────────────┐
                    │ 49 articles rejected on full text review │
                    │       7 did not use IV methods        │
                    │    13 exposure not at pregnancy       │
                    │       6 non-genetic instrument        │
                    │        2 non-offspring outcome        │
                    │ 16 review/pure theory (no data analysis) │
                    │  5 only abstract available (conference │
                    │              presentation)            │
                    └──────────────────────────────────────┘

          ┌─────────────────────┐
          │  43 eligible articles │
          └─────────────────────┘
```

Figure 3.2: Flowchart depicting article eligibility

falsification tests and sensitivity analyses performed, and limitations mentioned. For each of the MR conditions, rather than pre-specifying a list of possible types of violations and noting whether a particular article described said violation, reviewers listed all sources of bias described by the article under review that would violate the MR conditions. Although this approach relies on the ability of the reviewer to correctly identify sources of violation that are not explicitly described in the language of instrumental variables (particularly with regard to the fourth assumption), it allows for identification of novel and subject-specific approaches and potential sources of bias, rather than restricting responses to a predefined set of possible violations of the MR conditions. Data was extracted by the first author (ED); to assess accuracy in extraction, 5 included studies were randomly chosen for independent extraction by a coauthor (JL) (see Appendix for details of extraction comparison procedure). Both authors agreed on 56/60 data points (93%) across 5 articles.

## 3.4 Results

Initial searches resulted in 772 potentially eligible articles. Of these, 680 articles were excluded based on review of the abstract. Of the 92 articles that underwent full manuscript review, 43 articles met eligibility criteria and were included in this review (Figure 3.2).

**Study settings**

The included studies covered a wide range of exposures, including alcohol or tobacco use (n=12, 28%), caffeine use (n=1, 2%), C-reactive protein (n=2, 5%), diabetes (n=4, 9%), thyroid hormone levels (n=1, 2%), anthropometric traits (n=8, 19%), placental methylation (n=1, 2%), hemoglobin levels (n=3, 7%), blood lipid levels (n=2 , 5%), blood pressure (n=1, 2%), and micronutrient levels (n=13, 30%) (Table 3.1 Column 3). Of the micronutrient studies, 6 focused on folate, 2 on vitamin B-12, 2 on homocysteine, 2 on vitamin D, and 1 on polyunsaturated fatty acids. Outcomes of interest included DNA methylation, autoimmune conditions, cognitive development, anthropometric measures (e.g. adiposity-related outcomes), birthweight, bone density, behavioral disorders, smoking initiation, adverse birth outcomes, orofacial cleft, wheezing, and blood pressure (Table 3.1 Column 4). The majority (n=34, 79%) of the studies used data from a birth cohort, with a few studies using case-control designs (n=4) or cross-sectional data (n=5). Three studies (7%) used a 2-sample design, in which the association between the proposed instrument and exposure,

and between the proposed instrument and outcome, are estimated in independent samples.

Table 3.1: Included Studies

| First Author | Proposed Instrument(s) | Exposure | Outcome |
| --- | --- | --- | --- |
| Allard et al., 2015 | 2 step*: glucose genetic risk score (GRS), methylation GRS | 2 step: maternal fasting glucose, methylation | 2 step: methylation, cord blood leptin |
| Alwan et al., 2012 | C282Y | Iron | blood pressure, waist circumference, body mass index (BMI) |
| Bech et al., 2006 | NAT2, CYP1A2, GSTA1 | caffeine | stillbirth |
| Bédard et al., 2018 | maternal 12 SNP weighted GRS | hemoglobin | wheezing, asthma, atopy, low lung function |
| Bernard et al., 2018 | 8 FADS variants | omega 3 and omega 6 polyunsaturated fatty acids | gestational duration, birthweight, birth length |
| Binder and Michels, 2013 | MTHFR rs1801133, rs1801131 | Folate | genome-wide methylation |
| Bonilla, Lawlor, Ben-Shlomo, et al., 2012 | GRS | fasting glucose, type 2 diabetes | Intelligence quotient (IQ) at age 8 |
| Bonilla, Lawlor, Taylor, et al., 2012 | rs492602, rs1801198, rs96-6756 | vitamin B12 | IQ at age 8 |

Table 3.1: Included Studies *(continued)*

| First Author | Proposed Instrument(s) | Exposure | Outcome |
| --- | --- | --- | --- |
| Caramaschi et al., 2017 | 2 step: rs492602 + rs1047781 for vitamin b12, rs5750236, rs1890131 for methylation | 2 step: vitamin B12, methylation | 2 step: methylation, IQ |
| Caramaschi et al., 2018 | rs1051730 | smoking heaviness | autism spectrum disorder |
| Evans et al., 2019 | 403 SNP GRS | maternal type 2 diabetes | birthweight |
| Geng and Huang, 2018 | 35, 25, and 41 SNP GRS | waist-to-hip ratio adjusted for BMI, hip circumference adjusted for BMI, waist circumference adjusted for BMI | birthweight, birth length, head circumference, |
| Granell et al., 2008 | MTHFR C677T | Folate | atopy, asthma |
| Howe et al., 2019 | rs1229984 | alcohol | facial morphology |
| Humphriss et al., 2013 | ADH1B rs1229984 | alcohol | 3 composite balance scores (dynamic balance, static balance eyes open, static balance eyes closed) |

Table 3.1: Included Studies *(continued)*

| First Author | Proposed Instrument(s) | Exposure | Outcome |
|---|---|---|---|
| Hwang et al., 2019 | 96, 82, and 60 SNP GRS | High density lipoprotein (HDL) cholesterol, low density lipoprotein (LDL) cholesterol, triglycerides | birthweight |
| Korevaar et al., 2014 | GRS | Thyroid stimulating hormone (TSH), free thyroxine (FT4) | Soluble fms-like tyrosine kinase-1 (sFlt1), placental growth factor (PlGF) |
| Lawlor et al., 2008 | FTO | BMI | fat mass at age 9-11 |
| Lawlor et al., 2017 | GRS | BMI | BMI, fat mass index |
| Lee et al., 2013 | MTHFR C677T | homocysteine | birthweight |
| Lewis et al., 2009 | MTHFR C677T | folate intake | total weight,total body fat mass, total lean mass |
| Lewis et al., 2012 | 10 SNP in ADH4, ADH1A, AHD1B, ADH7 (rs4699714, rs3763894, rs4148884, rs2866151, rs975833, rs1229966, rs2066701, rs4147536, rs1229984, rs284779) | alcohol | cognitive score (IQ at age 8) |
| Lewis et al., 2014 | GRS based on rs1799945, rs1800562, rs4820268 | Iron | IQ at age 8 |

Table 3.1: Included Studies *(continued)*

| First Author | Proposed Instrument(s) | Exposure | Outcome |
| --- | --- | --- | --- |
| Mamasoula et al., 2013 | MTHFR rs1801133 | folate | congenital heart disease |
| Morales et al., 2011 | rs1205 | c-reactive protein (CRP) | wheezing, lower respiratory tract infection |
| Morales et al., 2016 | rs1983204, rs344008, rs6795327, rs7637701, rs11929637 | methylation at top-ranked cpg site for placental methylation in smokers | birthweight |
| Murray et al., 2016 | GRS ADH1A rs2866151, rs975833, AHD1B rs4147536, ADH7 rs284779 | alcohol | conduct problem trajectories (6 measures of strengths and difficulties questionnaire) |
| Richmond et al., 2016 | GRS | BMI | HIF3A methylation |
| Richmond et al., 2017 | GRS | BMI | BMI, fat mass index |
| Scholder et al., 2014 | ADH1B rs1229984 | alcohol | academic achievement (KS1,KS2,KS3, GCSE) |
| Shaheen et al., 2014 | ADH1B rs1229984 | alcohol | childhood atopic disease |
| Graaff and Roza, 2012 | MTHFR C677T | folate | emotional and behavioral score (child behavior checklist) |

Table 3.1: Included Studies *(continued)*

| First Author | Proposed Instrument(s) | Exposure | Outcome |
| --- | --- | --- | --- |
| Steer and Tobias, 2011 | MTHFR C677T | folate | Bone mineral content, bone mineral density, bone area |
| Taylor et al., 2014 | rs1051730 | smoking | latent class of offspring smoking initiation |
| Thompson et al., 2019 | separate 7 SNP GRS | vitamin D, calcium | Birthweight |
| Tyrrell et al., 2016 | GRS | BMI, fasting glucose, diabetes, triglycerids, HDL, blood pressure, vitamin D, adiponectin | birthweight |
| Wehby, Fletcher, et al., 2011 | 14 SNPs | smoking | Birthweight |
| Wehby, Jugessur, et al., 2011 | 4 SNPS (rs1435252, rs1930139, rs1547272, rs2743467 | smoking | orofacial cleft |
| Wehby and Scholder, 2013 | smoking: rs12914385, rs1051730, alcohol: ADH1B rs1229984, BMI: rs8050136 | smoking. alcohol use, obesity | birthweight |
| Yajnik et al., 2014 | MTHFR rs1801133 | homocysteine | birthweight |
| Zerbo et al., 2016 | rs3116656, rs2794520 | CRP | autism spectrum disorder |
| Zhang et al., 2015 | GRS | maternal height | birth length, birth weight |

Table 3.1: Included Studies *(continued)*

| First Author | Proposed Instrument(s) | Exposure | Outcome |
|---|---|---|---|
| Zuccolo et al., 2013 | rs1229984 | alcohol (1st trimester) | IQ at age 8, educational attainment |

*Note:*

2 step Mendelian randomization designs are a subtype of Mendelian randomization design proposed to investigate mediation of the relationship between maternal exposures and offspring outcomes by offspring DNA methylation, under additional strong assumptions. In this approach, maternal genetic variants are proposed as instruments for the effect of maternal exposures on offspring methylation across all measured sites. For any methylation sites where a non-null effect was detected for any individual in the population, offspring genetic variants associated with methylation at that site are proposed as instruments for the effect of methylation at that site on offspring outcomes.

The type and number of proposed instruments used varied across included studies. Most (n=31, 72%) studies proposed only maternal genetic factors as instruments, while the remainder used offspring genetic factors either alone or in tandem with maternal genetic factors. Overall, 19 studies (44%) proposed a single SNP as an instrument, while 24 (56%) used multiple genetic loci.

**Studies' discussion of key conditions**

Eighteen studies (42%) mentioned weak instrument bias, with 10 studies (23%) reporting F-statistics as a measure of proposed instrument strength (range: 0.66 to 74) (Appendix Table 3.4 Column 6). Seventeen studies (40%) incorporated methods explicitly to limit weak instrument bias into their analysis by leveraging multiple genetic loci as either a genetic risk score or using limited information maximum likelihood and weak instrument robust confidence intervals (Finlay and Magnusson, 2009; Stock et al., 2002).

Of 15 studies using genetic risk scores, rather than individual SNPs, 2 explicitly removed SNPs with known pleiotropic effects, that is, SNPs known both to be associated with the exposure and to impact the outcome through paths other than the exposure, from the genetic risk scores. Ten studies (23%) used alternative methods - Egger regression, weighted median regression, and sisVive -which allow for specific types of violations of MR condition 2 under alternative conditions (Bowden et al., 2015; Bowden et al., 2016; Kang et al., 2016) (Appendix Table 3.4 Column 11). Ten analyses (23%) controlled for offspring genotype, incorporated offspring genotype into a structural equation model,

or used only non-transmitted haplotypes as assumed instruments to mitigate violations of MR condition 2 by offspring genotype.

Twenty-six of the included studies (61%) used some method to avoid violations of MR condition 3 by population stratification, a type of confounding of the proposed instrument-outcome relationship by ancestry group, primarily (n=19, 44%) via restricting the maternal sample to white European women. Twelve studies (27.9%) included a sensitivity or primary analysis adjusting for GWAS derived principal components, to limit residual confounding by population stratification. Three studies discussed possible violations of MR condition 3 by assortative mating, a bias resulting from parents selecting mates based on particular characteristics that can result in confounding of the proposed instrument-outcome relationship. One study used linear mixed modeling to mitigate bias resulting from relatedness within the sample.

### Causal parameters of interest and reporting of additional key conditions

Twenty-one studies (49%) reported proposed instrument-outcome associations only, and twenty-two (51%) used IV estimation to derive a point estimate of an effect of the exposure on the outcome (Appendix Table 3.4 Column 5). Of the studies that reported such a point estimate, four explicitly reported their estimand of interest (See Appendix for details).

### Reported sensitivity analyses and falsification tests

While MR conditions 2 and 3 cannot be empirically verified, they can be falsified or indirectly assessed using a variety of techniques (Glymour et al., 2012; Jackson and Swanson, 2015). However, some of these techniques only detect extreme biases, and, particularly in the case of covariate balance, can be difficult to interpret (Glymour et al., 2012; Jackson and Swanson, 2015). Three analyses (7%) reported the results of a falsification test (Table 3.2). One study (2%) estimated a weighting function, and two (5%) used overidentification tests (Angrist and Imbens, 1995; Hausman, 1978). No studies reported instrumental inequalities (Diemer et al., 2020; Pearl, 1995) . Twenty-one studies (49%) reported the balance of covariates across levels of their proposed instrument, 17 of which compared this to covariate balance across levels of exposure; no studies used bias or bias component plots to report these comparisons (Jackson and Swanson, 2015) (Appendix Table 3.4 Column 11).

Eleven studies (26%) reported analyses stratified across levels of the exposure or conducted tests of instrument-exposure interaction or interaction between the instrument and a potentially confounded determinant of exposure. One

Table 3.2: Falsification approaches and sensitivity analyses reported by included articles

| Falsification Tests and Sensitivity Analyses | Percent studies reporting (n) |
|---|---|
| **Falsification Technique** | |
| Overidentification Test | 5% (2) |
| Weighting Function | 2% (1) |
| Covariate Balance | 49% (21) |
| **Sensitivity Analysis** | |
| Alternative Methods (MR-Egger, weighted median, nontransmitted haplotype, SisVive, mode-based estimator) | 23% (10) |
| Pruned GRS | 5% (2) |
| Simulations to evaluate impact of specific type of violation | 9% (4) |
| Adjustment for additional factors | 14% (6) |
| Exposure stratification (would only be valid if no unmeasured confounding of exposure and outcome) | 26% (11) |

study (2%) stratified across a level of maternal behavior in which the exposure was expected not to exist, and one study (2%) adjusted for several possible consequences of the proposed instrument and exposure. Because stratifying on or controlling for the exposure or a consequence of the exposure (as in a test of instrument-exposure interaction) can induce collider bias, these analyses will provide a valid falsification test only if there is no confounding of the exposure-outcome relationship, which is extremely unlikely given the typical motivation for conducting an MR analysis (Didelez and Sheehan, 2005).

**Reported limitations**

39 studies (91%) discussed versions of potential violations of the MR condition 2, with 30 (70%) describing pleiotropy, 10 (23%) noting possible postnatal effects of the proposed genetic instrument, 14 (33%) discussing possible exposure measurement error, 3 (7%) noting possible preconceptional effects of the proposed genetic instrument on egg quality or maternal characteristics, and 6 (14%) noting their exposure was assumed constant over the course of the pregnancy (Table 3.3). Thirty four studies (79%) discussed versions of potential violations of MR condition 3, with most (n=29, 67% of total) focusing on population stratification. Twenty eight studies (65%) mentioned low statistical

Table 3.3: Possible sources of violation of the MR conditions reported by the included articles

| Assumption | Percent Studies Reporting (n) |
| --- | --- |
| **Assumption 1** | |
| Weak Instrument Bias | 42% (18) |
| Can't Prove Assumption 1 | 7% (3) |
| Winner's Curse | 2% (1) |
| **Assumption 2** | |
| Pleiotropy | 70% (30) |
| Exposure Measurement Error | 33% (14) |
| Postnatal Effects of Genotype | 23% (10) |
| Preconceptional Effects of Genotype | 7% (3) |
| Exposure Assumed Constant over Pregnancy | 14% (6) |
| Offspring Genotype | 47% (20) |
| **Assumption 3** | |
| Population Stratification | 67% (29) |
| Assortative Mating | 7% (3) |
| Residual Confounding | 16% (7) |
| Relatedness | 2% (1) |
| **Other Concerns** | |
| Modeling Assumptions | 37% (16) |
| Selection Bias - Loss to Followup | 26% (11) |
| Selection on Pregnancy | 2% (1) |
| Outcome Measurement Error | 19% (8) |
| Low Power | 65% (28) |
| Limited Generalizability | 19% (8) |
| Use of GWAS in nonpregnant adults may be inappropriate | 9% (4) |
| Noncomparable cohort populations (2 sample designs only) | 2% (1) |

power. Eleven studies (26%) discussed possible selection bias related to missingness of exposure and outcome data. One study (2%) explicitly mentioned selection bias related to the use of a cohort defined by successful pregnancy completion. Sixteen studies (37%) discussed the vulnerability of their analysis to model misspecification resulting from nonlinearity or heterogeneity or violation of proportional hazards. Four studies (9%) noted that they used genetic risk scores weighted based on large GWAS of men and non-pregnant women, which might result in model misspecification when applied to prenatal MR. Of the three studies using two-sample designs, one discussed bias resulting from non-comparability of the samples.

## 3.5   Comment

**Principal findings**
The use of MR designs is becoming more frequently applied to study a wide range of types of prenatal exposures, and is most often conducted in large, well-characterized birth cohorts. Overall, investigators appear to be aware of possible bias due to pleiotropy and weak associations between proposed instruments and exposures, as well as the low power of MR studies, and demonstrate efforts to address the potential impact of these issues. However, some violations of the MR conditions that are more specific to and perhaps more common in prenatal MR, including violation of MR condition 2 by postnatal or preconceptional exposure status, and selection on pregnancy, are rarely mentioned. The fourth condition used to report point estimates is rarely stated.

**Strengths of the study**

This study is, to our knowledge, the first to investigate the use of the prenatal MR design and the possible violations discussed by applications of this design. The use of prenatal MR is increasing, and a clear evaluation of reported and unreported sources of potential bias is a key consideration for future authors and consumers of prenatal MR studies. By using an open-ended extraction strategy, rather than predefining biases of interest, we were able to identify novel sources of bias specific to this setting. This flexible approach enabled reviewers to identify violations of point-identifying assumptions that were not explicitly described in the language of instrumental variables.

**Limitations of the data**

However, this extraction strategy is, by definition, somewhat subjective. Be-

cause this approach relies on the expertise of the reviewer, reproducibility may be impacted. However, when data from 5 articles were independently extracted by a second coauthor, there was a high degree of agreement between reviewers. As with all systematic reviews, it is also possible that our search algorithm was incomplete, and we did not identify all relevant articles. This limitation is especially relevant to early prenatal MR studies, which did not always use the same language to describe their analysis, or conceptualize their analysis as an application of instrumental variables.

Our study focused exclusively on reporting, and therefore could not determine whether any potential bias meaningfully impacted the results of a particular study. However, key MR conditions are unverifiable, meaning the absence of all potential biases cannot be proven. Given this, MR studies should, whenever possible, attempt to falsify their assumptions, and provide sensitivity analyses quantifying the impact of possible biases. If the impact of particular bias is believed to be minor, justification of this assumption based on subject matter knowledge is vital to the interpretation of study findings.

**Interpretation**
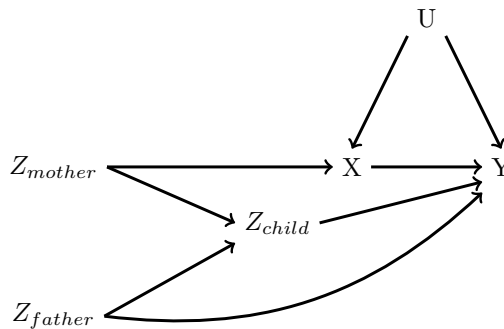


Figure 3.3: Causal DAG depicting a maternal genetic loci that violates the MR conditions. Here, offspring genotype ($Z_{child}$) is an open backdoor path between the proposed instrument ($Z_{mother}$) and the outcome ($Y$), violating MR condition 2. However, conditioning on $Z_{child}$ may induce a collider bias if paternal genotype ($Z_{father}$) is also related to $Y$, potentially via paternal exposure.

Violations of MR condition 2 were some of the most noted problems in this review. Pleiotropy, where genetic loci proposed as an instrument affect both the exposure and another maternal factor associated with the outcome, is a well-recognized problem for all MR studies, and was mentioned by nearly

94

three-quarters of the studies (70%) in this review. However, several types of violations of MR condition 2 are relatively unique to the prenatal MR design, some of which remain rarely acknowledged. When maternal genetic factors are proposed as instruments, MR condition 2 could be violated if the offspring's own genotype has an effect on the outcome (Figure 3.3) (Lawlor et al., 2017). This type of bias may be especially common in settings where the maternal exposure and offspring outcome are similar, including studies of the effect of maternal pregnancy BMI on offspring BMI (Lawlor et al., 2017). However, this type of bias could also occur in any setting where offspring exposure level might impact the outcome, or where the mechanism by which a genetic variant proposed as an instrument impacts exposure might also impact the outcome. In MR studies of the effect of prenatal micronutrient exposures on offspring outcomes in later life, MR condition 2 would be violated if offspring micronutrient levels after birth also affect the outcome, because offspring genotype likely impacts their micronutrient levels after birth. Some approaches to limit this bias have been proposed, including controlling for offspring genotype, the use of non-transmitted haplotypes, and a specific linear structural equation model. However, both the nontransmitted haplotype approach and controlling for offspring genotype can induce collider bias via paternal genotype, as both condition on offspring genotype. The structural equation modeling approach proposed by Warrington et al., 2018 avoids this issue, but require much stronger assumptions regarding linearity and relationships between covariates than conventional MR (VanderWeele, 2012).
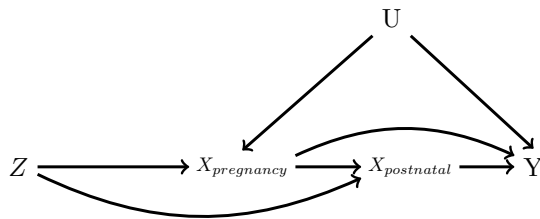


Figure 3.4: Causal DAG depicting a maternal genetic locus proposed as an instrument ($Z$), that violates the MR conditions. Here, $Z$ affects maternal exposure levels both during and after pregnancy, and maternal postnatal exposure also impacts offspring outcomes. Thus, maternal postnatal exposure ($X_{postnatal}$) creates an open backdoor path between $Z$ and the outcome ($Y$), violating MR condition 2.

For maternal proposed genetic instruments, if the outcome of interest occurs after birth, MR condition 2 can be violated if the mother's postnatal exposure

status also affects the offspring (Figure 3.4) (VanderWeele et al., 2014). This is because the mother's genes would logically affect her exposure after birth, and the postnatal effect of the exposure creates an open path between the proposed instrument and the outcome not via prenatal exposure. For example, if the exposure of interest impacts the content of the mother's breastmilk, this would violate MR condition 2. That path is particularly relevant for studies of the effects of obesity, diabetes, substance use, and vitamin $B_{12}$, all of which have been associated with altered breastmilk content (Allen, 2005; Andreas et al., 2014; Giglia and Binns, 2006; Koletzko et al., 2008; Rowe et al., 2013; Soderborg et al., 2016; Young et al., 2017) . In contrast, previous work has not found any association between maternal iron status and breastmilk content (Allen, 2005). Altered social exposures and parenting behaviors resulting from maternal postnatal exposure status (e.g., altered socioeconomic status or attachment style resulting from alcohol consumption) may also violate MR condition 2. For studies proposing offspring genetics as instruments, a similar violation can occur if the offspring's genetic factors continue to impact their exposure after birth. For example, as with biases resulting from the causal effect of maternal genotype on offspring genotype, in studies of the effect of micronutrients that propose offspring genetic factors as instruments will be biased if offspring micronutrient levels after birth impact their outcome, as offspring genotype likely continues to affect micronutrient levels after birth. Further, MR condition 2 can be violated if the mother's preconceptional exposure status affects her offspring, through mechanisms like alterations in oocyte quality.

Although an MR estimate of a maternal exposure with postnatal or preconceptional effects could retrieve a valid estimate of the effect of maternal exposure from oocyte formation to outcome measurement, such an approach implies exposures remain the same over several years (in the case of preconceptional effects, from the mother's own gestation to outcome measurement) and do not change as a result of pregnancy, an unreasonable assumption for many exposures of interest (Labrecque and Swanson, 2018). In addition, if the relationship between the proposed genetic instruments and maternal exposure status varies over the course of pregnancy, prenatal MR will produce biased estimates even if the exposure has no postnatal effect (Labrecque and Swanson, 2018; Swanson et al., 2017). Time-varying gene-exposure relationships were not explicitly mentioned in any of the articles reviewed here, though 10 studies mentioned pleiotropy via postnatal or prepregnancy effects as a possible limitation, and 6 noted the exposure was assumed constant over the course of pregnancy. In settings where postnatal exposure status is believed to substantially affect offspring outcomes, and the gene-exposure relationship varies over time (either

before or after birth), prenatal MR with the usual MR estimators will likely be an inappropriate approach, and investigators should consider alternative methods.
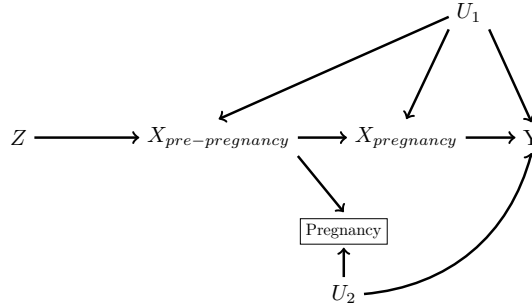


Figure 3.5: Causal DAG depicting a maternal genetic locus proposed as an instrument ($Z$) that violates the MR conditions. Here, the maternal exposure $X_{pre-pregnancy}$ impacts a woman's ability to become pregnant. As outcomes ($Y$) can only be measured in children of women who successfully conceive and carry a pregnancy to term, a prenatal MR study must necessarily select on pregnancy status, which will generate collider bias in this scenario, violating the MR conditions.

Violations of MR condition 3 by population stratification, a problem recognized in the broader MR literature, were well-discussed by studies included in this review (n= 29, 67%) (Davey Smith and Ebrahim, 2003; Didelez and Sheehan, 2007; Frangakis and Rubin, 1999; Lawlor et al., 2017; Palmer et al., 2012). Violations by selection bias related to participant loss to followup, another known problem in MR, were also mentioned by almost a third of studies in this review (n= 11, 26%) (Davey Smith and Ebrahim, 2003; Didelez and Sheehan, 2007; Frangakis and Rubin, 1999; Lawlor et al., 2017; Palmer et al., 2012). However, because many exposures also negatively impact fertility or ability to carry a pregnancy to term, prenatal MR studies are also uniquely vulnerable to bias resulting from selecting on successful pregnancy completion (Figure 3.5), which would result in a violation of the MR condition 3, a limitation mentioned by only 1 study in this review (Canan et al., 2017). This bias could also occur if women with particular substance use and dietary behaviors were less interested in becoming pregnant, or have other lifestyle factors that make it difficult to become pregnant. Previous research suggests that women who are obese are less likely to become pregnant than women who are not obese (Jokela et al., 2008). Folate status, diabetes, alcohol use, and smoking have been associated with worsened fertility, miscarriage, or stillbirth in experimental animal models

and previous observational research 45-56 (Altmäe et al., 2010; Bailey and Sokol, 2011; Boots and Stephenson, n.d.; Dechanet et al., 2010; Elbers et al., 2011; Fan et al., 2017; Feodor Nilsson et al., 2014; Gaskins et al., 2012; Laanpere et al., 2010; Waylen et al., 2008; Whitworth et al., 2011). Research on becoming pregnant in relation to other exposures included in this review, such as iron status, caffeine use, and c-reactive protein, is less conclusive 57-60 (Hanson et al., 2017; Wathes et al., 2007; Wesselink et al., 2016; Yadegari et al., 2016). Importantly, while selection due to miscarriage can also bias studies of prenatal exposures using more conventional confounder adjustment methods, selection bias due to differences in becoming pregnant affect only prenatal MR designs. Because the vast majority of prenatal MR studies are conducted in cohorts recruited based on the presence of a pregnancy, direct correction of estimates using inverse probability weights, a correction approach used in other applications of instrumental variable methods (Canan et al., 2017), will rarely be possible. Under specific conditions, the recently proposed MR GENIUS estimator may retrieve unbiased estimates of the causal effect in the presence of selection bias, though this motivation for applying the estimator has not been thoroughly evaluated (Tchetgen et al., 2017). As an alternative, authors using prenatal MR might consider using sensitivity analyses informed by previous research on their exposure and fertility in similar populations to evaluate the robustness of findings to selection bias (Swanson, 2019). However, to this point, no research has examined the magnitude of bias resulting from selection on pregnancy completion in prenatal MR, or optimal bias mitigation and sensitivity analysis strategies in the context of cohorts recruited based on the presence of a pregnancy. It is therefore unclear to what extent prenatal MR studies are biased by selection on pregnancy, and what measures future studies should take to limit or identify this bias.

Some sources of bias in prenatal MR may be particularly difficult to identify via the types of sensitivity analyses and falsification tests used by articles in this review. Comparisons of covariate balance across levels of the instrument and exposure, used by nearly half of the studies in this review, can be difficult to interpret, as even small differences in balance can result in substantial bias (Jackson and Swanson, 2015). Other methods used in this review, such as overidentification tests, which evaluate the null hypothesis that effect estimates from multiple different instruments are identical, and certain alternative methods allowing for relaxation of MR condition 2, assume that different estimates are not biased in the same way. While this assumption might be reasonable for some forms of horizontal pleiotropy, it will be violated if MR conditions 2 or 3 are violated as a result of a shared mechanism like postnatal effects of the

exposure, or by selection on pregnancy (Swanson, 2019). Two studies in this review attempted to limit pleiotropy by manually removing SNPs proposed as instruments that had known pleiotropic effects from genetic risk scores. This approach is a useful way of leveraging existing research to identify invalid IVs. However, identifying potentially pleiotropic SNPs in this way requires large GWAS of traits on potential pleiotropic pathways, which may be unavailable for many exposures used in prenatal MR.

While over half of the studies presented point estimates for a causal effect of exposure, few analyses explicitly discuss their estimand (n=4, 9% of total) or any form of model misspecification (n=15, 35% of total). Importantly, certain choices of weaker modeling assumptions will identify point estimates in different subsets of the population, and violations of modeling assumptions can also impact the interpretation of alternative methods, as well as falsification techniques like overidentification tests. Thus, explicit reporting of investigator assumptions is crucial to critical evaluation of MR analyses. This is especially true in prenatal MR, where certain subpopulations are not characterized in the same way as conventional MR, and, in the case of certain exposures, including maternal alcohol consumption and smoking, there is evidence that some modeling assumptions are unreasonable (see Appendix).

## 3.6    Conclusion

The use of prenatal MR is especially popular in the study of the effects of adiposity, micronutrient sufficiency, and substance use during pregnancy on offspring health. Because offspring are only directly exposed to maternal genetic factors and certain exposures during gestation, prenatal MR is an appealing method to examine the impact of maternal behaviors on offspring outcomes in the presence of unmeasured exposure-outcome confounding. Authors explicitly discuss and attempt to combat issues that could affect all MR studies, including population stratification, weak instruments, and certain types of pleiotropy, but much less frequently discuss some of the more specific challenges of prenatal MR designs, such as postnatal effects of the exposure and selection bias related to becoming pregnant. The evaluation of prenatal MR point estimates is also complicated by infrequent reporting of the authors' modeling assumptions and effect of interest, although this pattern has been seen in MR studies and even in other types of instrumental variable analyses more generally (Swanson and Hernán, 2013; Swanson et al., 2017).

Importantly, the presence of such potential biases in general in prenatal MR studies does not necessarily invalidate their existence. All causal inference is vulnerable to bias, and even biased studies can provide useful information about the world around us, particularly if the bias is well-understood. Future studies in this area should include explicit reporting and justification of the authors' assumptions, including those specific to the prenatal context, as well as falsification tests and sensitivity analyses to evaluate the impact of violations of those assumptions. Further research is needed to evaluate how selection bias related to fertility affects prenatal MR estimates, and to determine the best choice of analysis in the presence of violations of the MR conditions in studies of prenatal exposures. Altogether, the relatively frequent reporting of non-specific challenges while underreporting challenges specific to prenatal MR designs may also serve as an important lesson to the developers, teachers, and methodologic collaborators of MR analyses: while published MR applications may be increasingly better at reporting "standard" strengths and limitations of MR studies, critical assessment of the unique challenges of an MR study nonetheless needs to be done on a case-by-case basis.

**Acknowledgements**

# Appendix

## Details of search terms used in systematic review

| embase.com | 218 | 215 |
|---|---|---|
| Medline Ovid | 177 | 24 |
| Web of science | 250 | 129 |
| Cochrane CENTRAL | 16 | 1 |
| Google scholar | 200 | 136 |
| Total | 861 | 505 |

embase.com **218** ('instrumental variable analysis'/exp OR ((mendelian* NEAR/3 random*) OR (instrumental* NEAR/3 variab*)):ab,ti) AND ('prenatal exposure'/exp OR 'prenatal drug exposure'/exp OR 'pregnant woman'/exp OR 'pregnancy'/exp OR 'prenatal disorder'/exp OR 'pregnancy disorder'/exp OR 'parameters concerning the fetus, newborn and pregnancy'/exp OR 'prenatal development'/exp OR 'maternal nutrition'/exp OR 'maternal smoking'/exp OR 'Maternal Exposure'/exp OR 'embryonic and placental structures'/exp OR (prenatal* OR perinatal* OR pregnan* OR in*-uter* OR intrauter* OR gestation* OR maternal* OR offspring OR birthweight OR birth-weight OR fetus OR fetal OR foetus OR foetal OR placenta* OR embryo* OR fetomaternal* OR PreEclampsia OR Eclampsia):ab,ti)

Medline Ovid **262** (Mendelian Randomization Analysis/ OR ((mendelian* ADJ3 random*) OR (instrumental* ADJ3 variab*)).ab,ti.) AND (exp Pregnancy Complications/ OR Maternal Exposure/ OR pregnant women/ OR

exp pregnancy/ OR exp Fetal Diseases/ OR exp Pregnancy Complications/ OR exp Birth Weight/ OR exp Infant, Low Birth Weight/ OR Perinatal Mortality/ OR Perinatal Death/ OR Embryology/ OR exp "Embryonic and Fetal Development"/ OR exp Maternal Nutritional Physiological Phenomena/ OR exp Embryonic Structures/ OR (prenatal* OR perinatal* OR pregnan* OR in*-uter* OR intrauter* OR gestation* OR maternal* OR offspring OR birthweight OR birth-weight OR fetus OR fetal OR foetus OR foetal OR placenta* OR embryo* OR fetomaternal* OR PreEclampsia OR Eclampsia).ab,ti.)

Cochrane CENTRAL **16** (((mendelian* NEAR/3 random*) OR (instrumental* NEAR/3 variab*)):ab,ti) AND ((prenatal* OR perinatal* OR pregnan* OR in*-uter* OR intrauter* OR gestation* OR maternal* OR offspring OR birthweight OR birth-weight OR fetus OR fetal OR foetus OR foetal OR placenta* OR embryo* OR fetomaternal* OR PreEclampsia OR Eclampsia):ab,ti)

Web of science **250** TS=((((mendelian* NEAR/2 random*) OR (instrumental* NEAR/2 variab*))) AND ((prenatal* OR perinatal* OR pregnan* OR in*-uter* OR intrauter* OR gestation* OR (maternal* NEAR/3 (exposure* OR smoking OR drinking OR alcohol)) OR offspring OR birthweight OR birth-weight OR fetus OR fetal OR foetus OR foetal OR placenta* OR embryo* OR fetomaternal* OR PreEclampsia OR Eclampsia)))

Google scholar "mendelian randomization|randomisation"| "instrumental variable" prenatal|perinatal|pregnancy| pregnant|"in-uterus"|intrauterine| gestational|maternal| offspring|birthweight| "birth-weight"|fetus|fetal|foetus| foetal|placenta| embryo|fetomaternal

## Details of Extraction Procedure

Data points were extracted by the first author (ED); to ensure accuracy in extraction, 5 included studies were randomly chosen for independent extraction by a coauthor (JL). Data points on discussion of MR assumptions were considered in agreement when both authors agreed on the presence/absence of any discussion of violations of the assumption in question.

Sensitivity analyses and limitations were excluded from the comparison checking due to variability in how specific secondary analyses and limitations were categorized. This is because, rather than prespecifying sets of possible limitations and sensitivity analyses of interest, extraction of data points related to both sensitivity analyses and limitations discussed were open-ended to allow

for unexpected or unknown analyses and perspectives. This approach meant that each independent extractor could generate an arbitrarily large number of reported limitations and sensitivity analyses based on the same article. Thus, it would be difficult to measure the degree to which independent extractors agreed on datapoints related to sensitivity analyses and limitations discussed, as it is not possible to measure the number of datapoints the two authors agreed were not present in the dataset.

## Description of Included Studies

Table 3.4: Description of Included Studies

| First Author | 2. Maternal or Offspring Instrument | 3. Design | 4. Recruited Based on Presence of a Pregnancy | 5. Point Estimation? |
|---|---|---|---|---|
| Allard et al., 2015 | 1st step: maternal | cohort | yes | point estimate |
| Alwan et al., 2012 | maternal | cohort | no | point estimate |
| Bech et al., 2006 | maternal | nested case-control | yes ( nested in recruited that way) | instrument–outcome association |
| Bédard et al., 2018 | maternal | cohort | yes | point estimate |
| Bernard et al., 2018 | maternal and offspring | cohort | yes | instrument–outcome association |
| Binder and Michels, 2013 | maternal | cross-sectional | yes | point estimate |
| Bonilla, Lawlor, Ben-Shlomo, et al., 2012 | maternal and offspring | cohort | yes | instrument–outcome association |
| Bonilla, Lawlor, Taylor, et al., 2012 | maternal | cohort | yes | instrument–outcome association |
| Caramaschi et al., 2017 | maternal, offspring in 2nd step | cohort | yes | point estimate |
| Caramaschi et al., 2018 | maternal | cohort | yes | instrument–outcome association |

104

Table 3.4: Description of Included Studies (*continued*)

| First Author | 2. Maternal or Offspring Instrument | 3. Design | 4. Recruited Based on Presence of a Pregnancy | 5. Point Estimation? |
|---|---|---|---|---|
| Evans et al., 2019 | maternal and offspring | cross-sectional, UK biobank | no | point estimate |
| Geng and Huang, 2018 | maternal | 2 sample cross-sectional | no | point estimate |
| Granell et al., 2008 | maternal | cohort | yes | instrument–outcome association |
| Howe et al., 2019 | maternal | cohort | yes | instrument–outcome association |
| Humphriss et al., 2013 | maternal | cohort | yes | instrument–outcome association |
| Hwang et al., 2019 | maternal, controlled for offspring using SEM | 2 sample summary results, UK biobank and EGG | UK biobank no, EGG mostly yes | point estimate |
| Korevaar et al., 2014 | offspring | cohort | yes | instrument–outcome association |
| Lawlor et al., 2008 | maternal | cohort | yes | point estimate |
| Lawlor et al., 2017 | maternal | cohort | yes | point estimate |
| Lee et al., 2013 | maternal | cohort | yes | point estimate |
| Lewis et al., 2009 | maternal and offspring | cohort | yes | instrument–outcome association |

Table 3.4: Description of Included Studies (*continued*)

| First Author | 2. Maternal or Offspring Instrument | 3. Design | 4. Recruited Based on Presence of a Pregnancy | 5. Point Estimation? |
|---|---|---|---|---|
| Lewis et al., 2012 | both | cohort | yes | instrument-outcome association |
| Lewis et al., 2014 | maternal | cohort | yes | instrument-outcome association |
| Mamasoula et al., 2013 | maternal and offspring | case-control | cases yes, controls no | instrument-outcome association |
| Morales et al., 2011 | maternal | cohort | yes | instrument-outcome association |
| Morales et al., 2016 | offspring | cohort | yes | point estimate |
| Murray et al., 2016 | offspring | cohort | yes | instrument-outcome association |
| Richmond et al., 2016 | maternal | cohort | yes | point estimate |
| Richmond et al., 2017 | maternal | cohort | yes | point estimate |
| Scholder et al., 2014 | maternal, controlling for offspring | cohort | yes | point estimate |
| Shaheen et al., 2014 | maternal | cohort | yes | instrument-outcome association |

Table 3.4: Description of Included Studies (*continued*)

| First Author | 2. Maternal or Offspring Instrument | 3. Design | 4. Recruited Based on Presence of a Pregnancy | 5. Point Estimation? |
| --- | --- | --- | --- | --- |
| Graaff and Roza, 2012 | maternal | cohort | yes | instrument-outcome association |
| Steer and Tobias, 2011 | maternal | cohort | yes | instrument-outcome association |
| Taylor et al., 2014 | maternal | cohort | yes | instrument-outcome association |
| Thompson et al., 2019 | maternal | 2 sample: various GWAS+ UKB cross-sectional, sensitivity analyses: ALSPAC/EFSOCH cohort | ALSPAC/EFSOCH yes, UKB no | point estimate |
| Tyrrell et al., 2016 | maternal | meta-analysis of multiple cohorts | yes for all | point estimate |
| Wehby, Fletcher, et al., 2011 | maternal | 1 cohort, 1 cross-sectional | yes for Norway, no for AddHealth | point estimate |
| Wehby, Jugessur, et al., 2011 | maternal | case-control | yes | point estimate |
| Wehby and Scholder, 2013 | maternal | cohort (multiple) | yes for both | point estimate |
| Yajnik et al., 2014 | maternal | 2 cohorts | yes | point estimate |
| Zerbo et al., 2016 | maternal | nested case-control | yes | instrument-outcome association |

Table 3.4: Description of Included Studies (*continued*)

| First Author | 2. Maternal or Offspring Instrument | 3. Design | 4. Recruited Based on Presence of a Pregnancy | 5. Point Estimation? |
|---|---|---|---|---|
| Zhang et al., 2015 | maternal | 3 case-control nested within cohorts | yes for all | point estimate |
| Zuccolo et al., 2013 | maternal, sensitivity analysis with offspring | cohort | yes | instrument-outcome association |

Table 3.5: Description of Included Studies

| First Author | 6. F-statistic | 7. Assumption 1 Violations Discussed | 8. Assumption 2 Violations Discussed | 9. Assumption 3 Violations Discussed |
|---|---|---|---|---|
| Allard et al., 2015 | | | | Population Stratification |
| Alwan et al., 2012 | 10 | Weak Instrument Bias | Pleiotropy, Exposure Measurement Error, Offspring Genotype Path | |
| Bech et al., 2006 | | Weak Instrument Bias | | |
| Bédard et al., 2018 | | | Pleiotropy, Exposure Measurement Error, Offspring Genotype Path | Population Stratification |
| Bernard et al., 2018 | | Weak Instrument Bias | Pleiotropy ,Exposure Measurement Error | Population Stratification |
| Binder and Michels, 2013 | 4.88 | Weak Instrument Bias | 2nd or 3rd Assumption - General | Population Stratification |
| Bonilla, Lawlor, Ben-Shlomo, et al., 2012 | | | Offspring Genotype Path | Population Stratification |
| Bonilla, Lawlor, Taylor, et al., 2012 | | | Exposure Measurement Error, Offspring Genotype Path | Population Stratification |
| Caramaschi et al., 2017 | | | Pleiotropy, Exposure Measurement Error | Population Stratification |

Table 3.5: Description of Included Studies (*continued*)

| First Author | 6. F-statistic | 7. Assumption 1 Violations Discussed | 8. Assumption 2 Violations Discussed | 9. Assumption 3 Violations Discussed |
|---|---|---|---|---|
| Caramaschi et al., 2018 | | | Pleiotropy, Exposure Measurement Error | |
| Evans et al., 2019 | | Weak Instrument Bias | Pleiotropy, Offspring Genotype Path | Population Stratification |
| Geng and Huang, 2018 | | Weak Instrument Bias | Pleiotropy, Postnatal Effects of Genotype, Offspring Genotype Path, | Population Stratification |
| Granell et al., 2008 | | Weak Instrument Bias | | |
| Howe et al., 2019 | | | Pleiotropy ,Offspring Genotype Path | Assortative Mating |
| Humphriss et al., 2013 | | Weak Instrument Bias, Cannot Prove Assumption 1 (Reduced Form) | Pleiotropy | Population Stratification |
| Hwang et al., 2019 | | | Pleiotropy, Postnatal Effects of Genotype, Exposure Assumed Constant Over Pregnancy, Offspring Genotype Path | Population Stratification |
| Korevaar et al., 2014 | | | Exposure Assumed Constant Over Pregnancy | |

110

Table 3.5: Description of Included Studies (*continued*)

| First Author | 6. F-statistic | 7. Assumption 1 Violations Discussed | 8. Assumption 2 Violations Discussed | 9. Assumption 3 Violations Discussed |
|---|---|---|---|---|
| Lawlor et al., 2008 | 12.9, 10.1 after adjustment for offspring genotype | | Pleiotropy, Postnatal Effects of Genotype, Offspring Genotype Path | Population Stratification |
| Lawlor et al., 2017 | >45 | Weak Instrument Bias | Pleiotropy, Postnatal Effects of Genotype, Offspring Genotype Path | Population Stratification, Assortative Mating |
| Lee et al., 2013 | | | Pleiotropy, 2nd or 3rd Assumption - General | |
| Lewis et al., 2009 | | | Exposure Measurement Error | |
| Lewis et al., 2012 | | Weak Instrument Bias | Pleiotropy | Population Stratification |
| Lewis et al., 2014 | | | Pleiotropy, Offspring Genotype Path | Population Stratification |
| Mamasoula et al., 2013 | | | 2nd or 3rd Assumption - General | |
| Morales et al., 2011 | | Weak Instrument Bias | Pleiotropy, Exposure Assumed Constant Over Pregnancy, 2nd or 3rd Assumption - General | |

Table 3.5: Description of Included Studies *(continued)*

| First Author | 6. F-statistic | 7. Assumption 1 Violations Discussed | 8. Assumption 2 Violations Discussed | 9. Assumption 3 Violations Discussed |
|---|---|---|---|---|
| Morales et al., 2016 | | Winner's Curse | Pleiotropy ,Exposure Measurement Error, | |
| Murray et al., 2016 | | Cannot Prove Assumption 1 (Reduced Form) | Pleiotropy | Population Stratification |
| Richmond et al., 2016 | 45.7 in offspring | | Pleiotropy, Exposure Measurement Error, Offspring Genotype Path | Population Stratification |
| Richmond et al., 2017 | minimum 45 | Weak Instrument Bias | Pleiotropy, Exposure Measurement Error,Postnatal Effects of Genotype, Offspring Genotype Path | Population Stratification |
| Scholder et al., 2014 | 1.38-24.76 | Weak Instrument Bias | Pleiotropy ,Postnatal Effects of Genotype, Exposure Assumed Constant Over Pregnancy, Offspring Genotype Path, 2nd or 3rd Assumption - General | Population Stratification |
| Shaheen et al., 2014 | | | Offspring Genotype Path | Population Stratification |

Table 3.5: Description of Included Studies (*continued*)

| First Author | 6. F-statistic | 7. Assumption 1 Violations Discussed | 8. Assumption 2 Violations Discussed | 9. Assumption 3 Violations Discussed |
|---|---|---|---|---|
| Graaff and Roza, 2012 | | | Exposure Assumed Constant Over Pregnancy ,2nd or 3rd Assumption - General | Population Stratification |
| Steer and Tobias, 2011 | | | Offspring Genotype Path | |
| Taylor et al., 2014 | | | Pleiotropy, Exposure Measurement Error, Postnatal Effects of Genotype | Population Stratification |
| Thompson et al., 2019 | | Weak Instrument Bias | Pleiotropy, Exposure Measurement Error, Exposure Assumed Constant Over Pregnancy, Offspring Genotype Path | Population Stratification |
| Tyrrell et al., 2016 | | Weak Instrument Bias | Pleiotropy, Postnatal Effects of Genotype, Offspring Genotype Path | Population Stratification |
| Wehby, Fletcher, et al., 2011 | 3.3-4.4 | Weak Instrument Bias | Pleiotropy | Population Stratification |
| Wehby, Jugessur, et al., 2011 | 3.33 | Weak Instrument Bias | Pleiotropy, Exposure Measurement Error | |

Table 3.5: Description of Included Studies *(continued)*

| First Author | 6. F-statistic | 7. Assumption 1 Violations Discussed | 8. Assumption 2 Violations Discussed | 9. Assumption 3 Violations Discussed |
|---|---|---|---|---|
| Wehby and Scholder, 2013 | 0.66-35.49 | Weak Instrument Bias | Pleiotropy, Postnatal Effects of Genotype | Population Stratification |
| Yajnik et al., 2014 | | | Pleiotropy, Offspring Genotype Path | Population Stratification |
| Zerbo et al., 2016 | | | | Population Stratification |
| Zhang et al., 2015 | | | Pleiotropy, Offspring Genotype Path | Assortative Mating |
| Zuccolo et al., 2013 | | Cannot Prove Assumption 1 (Reduced Form) | Pleiotropy, Exposure Measurement Error, Postnatal Effects of Genotype, 2nd or 3rd Assumption - General | Population Stratification |

Table 3.6: Description of Included Studies

| First Author | 10. Other Limitations Discussed | 11. Falsification Techniques | 12. Exposure Stratification/Testing |
|---|---|---|---|
| Allard et al., 2015 | Outcome Measurement Error,Low Power | Alternative MR Methods | |
| Alwan et al., 2012 | Low Power | Covariate Balance | |
| Bech et al., 2006 | Low Power | Covariate Balance | Exposure Stratification/Testing |
| Bédard et al., 2018 | Selection Bias,Low Power | Covariate Balance,Alternative MR Methods | |
| Bernard et al., 2018 | Outcome Measurement Error,Low Power | | |
| Binder and Michels, 2013 | Modeling Assumptions,Outcome Measurement Error,Low Power, | | |
| Bonilla, Lawlor, Ben-Shlomo, et al., 2012 | Modeling Assumptions,Low Power | Covariate Balance | |
| Bonilla, Lawlor, Taylor, et al., 2012 | ,Selection Bias,Low Power,Limited Generalizability | Covariate Balance | Exposure Stratification/Testing |
| Caramaschi et al., 2017 | Low Power,Limited Generalizability | | |
| Caramaschi et al., 2018 | Low Power | Covariate Balance | Exposure Stratification/Testing |

115

Table 3.6: Description of Included Studies (*continued*)

| First Author | 10. Other Limitations Discussed | 11. Falsification Techniques | 12. Exposure Stratification/Testing |
|---|---|---|---|
| Evans et al., 2019 | Modeling Assumptions | Alternative MR Methods | |
| Geng and Huang, 2018 | Modeling Assumptions, Limited Generalizability | Alternative MR Methods | |
| Granell et al., 2008 | Modeling Assumptions,Selection Bias,Outcome Measurement Error | Covariate Balance | Exposure Stratification/Testing |
| Howe et al., 2019 | Outcome Measurement Error | Covariate Balance | |
| Humphriss et al., 2013 | Outcome Measurement Error,Low Power,Limited Generalizability | Covariate Balance | |
| Hwang et al., 2019 | Limited Generalizability | Alternative MR Methods | |
| Korevaar et al., 2014 | Low Power | | |
| Lawlor et al., 2008 | Selection Bias, Low Power | Covariate Balance | |
| Lawlor et al., 2017 | Low Power | Covariate Balance,Alternative MR Methods | |
| Lee et al., 2013 | Selection on Pregnancy,Low Power,Limited Generalizability | Covariate Balance | Exposure Stratification/Testing |
| Lewis et al., 2009 | | | |

Table 3.6: Description of Included Studies (*continued*)

| First Author | 10. Other Limitations Discussed | 11. Falsification Techniques | 12. Exposure Stratification/Testing |
|---|---|---|---|
| Lewis et al., 2012 | Selection Bias | | Exposure Stratification/Testing |
| Lewis et al., 2014 | Modeling Assumptions,Low Power | Covariate Balance | |
| Mamasoula et al., 2013 | Modeling Assumptions | | |
| Morales et al., 2011 | Low Power | | |
| Morales et al., 2016 | Low Power | Alternative MR Methods | |
| Murray et al., 2016 | Selection Bias, Low Power | Covariate Balance | Exposure Stratification/Testing |
| Richmond et al., 2016 | Modeling Assumptions,Selection Bias,Low Power | Covariate Balance | |
| Richmond et al., 2017 | Modeling Assumptions,Selection Bias, Low Power | Covariate Balance,Alternative MR Methods | |
| Scholder et al., 2014 | Modeling Assumptions | Weight Function,Covariate Balance | |
| Shaheen et al., 2014 | Selection Bias,Low Power | Covariate Balance | Exposure Stratification/Testing |

117

Table 3.6: Description of Included Studies (*continued*)

| First Author | 10. Other Limitations Discussed | 11. Falsification Techniques | 12. Exposure Stratification/Testing |
|---|---|---|---|
| Graaff and Roza, 2012 | Modeling Assumptions,Selection Bias | | Exposure Stratification/Testing |
| Steer and Tobias, 2011 | | | |
| Taylor et al., 2014 | Low Power,Limited Generalizability | | Exposure Stratification/Testing |
| Thompson et al., 2019 | Outcome Measurement Error | Alternative MR Methods | |
| Tyrrell et al., 2016 | Modeling Assumptions,Low Power | Covariate Balance | |
| Wehby, Fletcher, et al., 2011 | Modeling Assumptions | Overidentification Tests | |
| Wehby, Jugessur, et al., 2011 | Modeling Assumptions | Overidentification Tests | |
| Wehby and Scholder, 2013 | Modeling Assumptions,Low Power,Limited Generalizability | Covariate Balance | |
| Yajnik et al., 2014 | Low Power | | |
| Zerbo et al., 2016 | Outcome Measurement Error,Low Power | | |
| Zhang et al., 2015 | Selection Bias | Alternative MR Methods | |
| Zuccolo et al., 2013 | Low Power | Covariate Balance | Exposure Stratification/Testing |

## Assumptions required for point estimation

Investigators can test whether there is a non-null effect of the exposure on the outcome for at least one individual in the study population, and can estimate bounds for the average causal effect using only the 3 instrumental variable assumptions discussed in the main text (Hernán and Robins, 2006). To estimate the average causal effect of an exposure X on an outcome Y, using an instrument Z in the total study population, investigators must assume one of the following conditions hold (Hernan and Robins, 2018; Hernán and Robins, 2006; Tchetgen et al., 2017).

4a. The effect of $X$ on $Y$ is identical (constant) for all individuals in the population: $E(Y^{x=x} - Yx = 0|X = x) = E(Y^{x=x} - Y^{x=0}|X = 0)$

4b. No effect modification by the instrument $Z$ in all levels of the exposure $X$:
$E(Y^{x=x} - Y^{x=0}|X = x, Z = z) = E(Y^{x=x} - Y^{x=0}|X = x, Z = z')$
Or equivalently
$E(Y^{x=x} - Y^{x=0}|X = x, Z = z) = E(Y^{x=x} - Y^{x=0}|X = x', Z = z)$

Recent research has found that the average causal effect can also be identified, even in the presence of violations of the second and third assumption, under one of two alternative assumptions by an additional variable $J$, as can be seen in Figure 3.6 (Tchetgen et al., 2017). In this case, the usual 3 IV conditions are replaced by the following:

1. $Z \not\!\perp\!\!\!\perp Y|J$

2. $Z \perp\!\!\!\perp U \mid X$

3. $Z \perp\!\!\!\perp Y \mid (J,\ U,\ X)$

4. $Y^x \perp\!\!\!\perp (Z,X) \mid (J,U)$

Under these conditions, point estimation of the average causal effect is possible if one of the two following conditions hold:

4c. No additive $U - Z$ interaction on $E(X|Z, J, U)$: $E(X|Z = z, J, U)$ - $E(X|Z = 0, J, U) = E(X|Z = z, J)$ - $E(X|Z = 0, J)$

4d. No additive $U - X$ interaction on the average causal effect of $X$ on $Y$:
$\mathrm{E}(Y^{x=x} - Y^{X=0}|J, U) = \mathrm{E}(Y^{X=x} - Y^{X=0}|J)$

If the above assumptions are not plausible for a particular analysis, researchers can estimate the average causal effect within the compliers, those individuals for whom $X^{Z=a} > X^{Z=b}$ for all $a > b$ (Hernán and Robins, 2006). This value is also known as the local average treatment effect, or LATE. In order to estimate this quantity, researchers must assume:

4e. The causal effect of $Z$ on $X$ is monotonic, that is, it only works in one direction for every individual in the study population. Formally, $X^z$ is a non-decreasing function of $z$ on the support of $Z$.
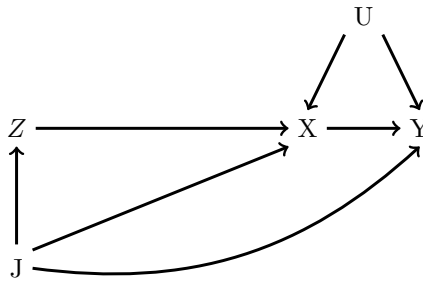


Figure 3.6: Causal DAG representing an instrumental variable model with violation of assumption 3 by $J$. Under this model, valid estimation of $\mathrm{E}(Y^{X=x} - Y^{X=0})$ is still possible using the assumptions presented by Tchetgen et al., 2017.

## Interpretation of certain additional point-estimating assumptions in prenatal MR

Four studies in this review reported additional point-estimating assumptions and their targeted estimand. Of these, 3 assumed monotonicity (condition 4e) in order to estimate the average causal effect among the compliers, and 1 assumed no effect modification by the instrument in all levels of the exposure (condition 4b) to estimate the average causal effect in the total study population.

In the context of certain pregnancy exposures, there is evidence that conditions 4a, 4b, 4c, and 4d are unreasonable. When genes related to alcohol metabolism are used as instruments for maternal drinking during pregnancy, fetal exposure

to alcohol and alcohol metabolites will depend on maternal intake and the speed at which the mother can metabolize alcohol, as well as other environmental factors. For the same level of maternal alcohol intake, offspring of slow metabolizers will have a longer exposure to alcohol, and would be at greater risk of negative health outcomes (Smith, 2010). This means that the average causal effect of alcohol exposure on offspring outcomes will be modified by the level of the maternal genetic variant proposed as an instrument, violating conditions 4a, 4b, 4c, and 4d. For this reason, most studies of alcohol use during pregnancy in this review focused on a testing approach, rather than point estimation. The same logic applies to other metabolism-related genetic variants proposed as instruments for substance use behaviors, like smoking and caffeine use. In these cases, studies may choose to focus on approaches with weaker assumptions, such as the complier average treatment effect, testing approaches, or bounds.

It is important to note that, in prenatal MR proposing maternal genetic factors as instruments, the interpretation of "compliers" and the complier average causal effect (described above) are different than the usual interpretation in MR studies or most studies using instrumental variable analyses (Swanson, 2017). This is because a mother-child pair's compliance status is determined by the relationship between a mother's genetics and exposure, while the average causal effect of interest occurs in the offspring of those mothers. In typical MR and instrumental variable studies, the proposed instrument, exposure, and outcome are all measured within the same individual. In those cases, under condition 4e, researchers can estimate the average causal effect among the compliers. In contrast, in pregnancy MR designs, under condition 4e, researchers can estimate the average causal effect among the offspring of mothers who are compliers, although the offspring themselves would not necessarily be compliers.

# References

Allard, C., Desgagné, V., Patenaude, J., Lacroix, M., Guillemette, L., Battista, M. C., Doyon, M., Ménard, J., Ardilouze, J. L., Perron, P., Bouchard, L., & Hivert, M. F. (2015). Mendelian randomization supports causality between maternal hyperglycemia and epigenetic regulation of leptin gene in newborns. *Epigenetics*, *10*(4), 342–351. https://doi.org/10.1080/15592294.2015.1029700

Allen, L. H. (2005). Multiple micronutrients in pregnancy and lactation: An overview. *The American journal of clinical nutrition*, *81*(5), 1206S–1212S.

Altmäe, S., Stavreus-Evers, A., Ruiz, J. R., Laanpere, M., Syvänen, T., Yngve, A., Salumets, A., & Nilsson, T. K. (2010). Variations in folate pathway genes are associated with unexplained female infertility. *Fertility and sterility*, *94*(1), 130–137.

Alwan, N. A., Lawlor, D. A., McArdle, H. J., Greenwood, D. C., & Cade, J. E. (2012). Exploring the relationship between maternal iron status and offspring's blood pressure and adiposity: A mendelian randomization study. *Clin Epidemiol*, *4*(1), 193–200.

Andreas, N. J., Hyde, M. J., Gale, C., Parkinson, J. R. C., Jeffries, S., Holmes, E., & Modi, N. (2014). Effect of maternal body mass index on hormones in breast milk: A systematic review. *PloS one*, *9*(12), e115043.

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, *90*(430), 431–442.

Bailey, B. A., & Sokol, R. J. (2011). Prenatal alcohol exposure and miscarriage, stillbirth, preterm delivery, and sudden infant death syndrome. *Alcohol Research and Health*, *34*(1), 86.

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176.

Bech, B. H., Autrup, H., Nohr, E. A., Henriksen, T. B., & Olsen, J. (2006). Stillbirth and slow metabolizers of caffeine: Comparison by genotypes. *Int J Epidemiol*, *35*(4), 948–953. https://doi.org/10.1093/ije/dyl116

Bédard, A., Lewis, S. J., Burgess, S., John Henderson, A., & Shaheen, S. O. (2018). Maternal iron status during pregnancy and respiratory and atopic outcomes in the offspring: A mendelian randomisation study. *BMJ Open Respir Res*, *5*(1). https://doi.org/10.1136/bmjresp-2018-000275

Bernard, J. Y., Pan, H., Aris, I. M., Moreno-Betancur, M., Soh, S. E., Yap, F., Tan, K. H., Shek, L. P., Chong, Y. S., Gluckman, P. D., Calder, P. C., Godfrey, K. M., Chong, M. F. F., Kramer, M. S., Karnani, N., & Lee, Y. S. (2018). Long-chain polyunsaturated fatty acids, gestation duration, and birth size: A mendelian randomization study using fatty acid desaturase variants. *Am J Clin Nutr*, *108*(1), 92–100. https://doi.org/10.1093/ajcn/nqy079

Binder, A. M., & Michels, K. B. (2013). The causal effect of red blood cell folate on genome-wide methylation in cord blood: A mendelian randomization approach. *BMC Bioinformatics*, *14*, 353. https://doi.org/10.1186/1471-2105-14-353

Bonilla, C., Lawlor, D. A., Ben-Shlomo, Y., Ness, A. R., Gunnell, D., Ring, S. M., Smith, G. D., & Lewis, S. J. (2012). Maternal and offspring fasting glucose and type 2 diabetes-associated genetic variants and cognitive function at age 8: A mendelian randomization study in the avon longitudinal study of parents and children. *BMC Med Genet*, *13*. https://doi.org/10.1186/1471-2350-13-90

Bonilla, C., Lawlor, D. A., Taylor, A. E., Gunnell, D. J., Ben-Shlomo, Y., Ness, A. R., Timpson, N. J., Pourcain, B. S., Ring, S. M., Emmett, P. M., Smith, A. D., Refsum, H., Pennell, C. E., Brion, M. J., Smith, G. D., & Lewis, S. J. (2012). Vitamin b-12 status during pregnancy and child's iq at age 8: A mendelian randomization study in the avon longitudinal study of parents and children. *PLoS ONE*, *7*(12). https://doi.org/10.1371/journal.pone.0051084

Boots, C., & Stephenson, M. D. (n.d.). Does obesity increase the risk of miscarriage in spontaneous conception: A systematic review. *Seminars in reproductive medicine*, *29*, 507–513.

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International journal of epidemiology*, *44*(2), 512–525.

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

Canan, C., Lesko, C., & Lau, B. (2017). Instrumental variable analyses and selection bias. *Epidemiology (Cambridge, Mass.)*, *28*(3), 396.

Caramaschi, D., Sharp, G. C., Nohr, E. A., Berryman, K., Lewis, S. J., Davey Smith, G., & Relton, C. L. (2017). Exploring a causal role of dna methylation in the relationship between maternal vitamin b12 during pregnancy and child's iq at age 8, cognitive performance and educa-

tional attainment: A two-step mendelian randomization study. *Hum Mol Genet*, *26*(15), 3001–3013.

Caramaschi, D., Taylor, A. E., Richmond, R. C., Havdahl, K. A., Golding, J., Relton, C. L., Munafò, M. R., Davey Smith, G., & Rai, D. (2018). Maternal smoking during pregnancy and autism: Using causal inference methods in a birth cohort study. *Transl Psychiatry*, *8*(1). https://doi.org/10.1038/s41398-018-0313-5

Davey Smith, G., & Ebrahim, S. (2003). 'mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, *32*(1), 1–22.

Dechanet, C., Anahory, T., Mathieu Daude, J. C., Quantin, X., Reyftmann, L., Hamamah, S., Hedon, B., & Dechaud, H. (2010). Effects of cigarette smoking on reproduction. *Human reproduction update*, *17*(1), 76–95.

Didelez, V., Meng, S., & Sheehan, N. A. (2010). Assumptions of iv methods for observational epidemiology. *Statistical Science*, 22–40.

Didelez, V., & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, *16*(4), 309–330.

Didelez, V., & Sheehan, N. A. (2005). Mendelian randomisation and instrumental variables: What can and what can't be done. *University of Leicester, Department of Health Science, Technical Report 05*, *2*.

Diemer, E. W., Labrecque, J., Tiemeier, H., & Swanson, S. A. (2020). Application of the instrumental inequalities to a mendelian randomization study with multiple proposed instruments. *Epidemiology*, *31*(1), 65–74.

Elbers, C. C., Onland-Moret, N. C., Eijkemans, M. J. C., Wijmenga, C., Grobbee, D. E., & van der Schouw, Y. T. (2011). Low fertility and the risk of type 2 diabetes in women. *Human reproduction*, *26*(12), 3472–3478.

Evans, D. M., Moen, G. H., Hwang, L. D., Lawlor, D. A., & Warrington, N. M. (2019). Elucidating the role of maternal environmental exposures on offspring health and disease using two-sample mendelian randomization.

Fan, D., Liu, L., Xia, Q., Wang, W., Wu, S., Tian, G., Liu, Y., Ni, J., Wu, S., & Guo, X. (2017). Female alcohol consumption and fecundability: A systematic review and dose-response meta-analysis. *Scientific Reports*, *7*(1), 13815.

Feodor Nilsson, S., Andersen, P. K., Strandberg-Larsen, K., & Nybo Andersen, A.-M. (2014). Risk factors for miscarriage from a prevention perspec-

tive: A nationwide follow-up study. *BJOG: An International Journal of Obstetrics and Gynaecology*, *121*(11), 1375–1385.

Finlay, K., & Magnusson, L. M. (2009). Implementing weak-instrument robust tests for a general class of instrumental-variables models. *Stata Journal*, *9*(3), 398.

Fleming, T. P., Velazquez, M. A., & Eckert, J. J. (2015). Embryos, dohad and david barker. *Journal of developmental origins of health and disease*, *6*(5), 377–383.

Frangakis, C. E., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, *86*(2), 365–379.

Gaskins, A. J., Mumford, S. L., Chavarro, J. E., Zhang, C., Pollack, A. Z., Wactawski-Wende, J., Perkins, N. J., & Schisterman, E. F. (2012). The impact of dietary folate intake on reproductive function in premenopausal women: A prospective cohort study. *PloS one*, *7*(9), e46276.

Geng, T. T., & Huang, T. (2018). Maternal central obesity and birth size: A mendelian randomization analysis. *Lipids Health Dis*, *17*(1). https: //doi.org/10.1186/s12944-018-0831-4

Giglia, R., & Binns, C. (2006). Alcohol and lactation: A systematic review. *Nutrition and Dietetics*, *63*(2), 103–116.

Glymour, M. M., Tchetgen Tchetgen, E. J., & Robins, J. M. (2012). Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American journal of epidemiology*, *175*(4), 332–339.

Graaff, J. S.-d., & Roza, S. J. (2012). Maternal folate status in early pregnancy and child emotional and behavioral problems: The generation r study. *The American journal . . .*

Granell, R., Heron, J., Lewis, S., Smith, G. D., Sterne, J. A. C., & Henderson, J. (2008). The association between mother and child mthfr c677t polymorphisms, dietary folate intake and childhood atopy in a population-based, longitudinal birth cohort. *Clin. Exp. Allergy*, *38*(2), 320–328.

Hanson, B., Johnstone, E., Dorais, J., Silver, B., Peterson, C. M., & Hotaling, J. (2017). Female infertility, infertility-associated diagnoses, and comorbidities: A review. *Journal of assisted reproduction and genetics*, *34*(2), 167–177.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, 1251–1271.

Hernan, M. A., & Robins, J. M. (2018). *Causal inference*. Chapman; Hall/CRC.

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 360–372.

Holmes, M. V., Ala-Korpela, M., & Smith, G. D. (2017). Mendelian randomization in cardiometabolic disease: Challenges in evaluating causality. *Nature Reviews Cardiology*, *14*(10), 577.

Howe, L. J., Sharp, G. C., Hemani, G., Zuccolo, L., Richmond, S., & Lewis, S. J. (2019). Prenatal alcohol exposure and facial morphology in a uk cohort. *Drug Alcohol Depend*, *197*, 42–47. https://doi.org/10.1016/j. drugalcdep.2018.11.031

Humphriss, R., Hall, A., May, M., Zuccolo, L., & Macleod, J. (2013). Prenatal alcohol exposure and childhood balance ability: Findings from a uk birth cohort study. *BMJ Open*, *3*(6). https://doi.org/10.1136/ bmjopen-2013-002718

Hwang, L. D., Lawlor, D. A., Freathy, R. M., Evans, D. M., & Warrington, N. M. (2019). Using a two-sample mendelian randomization design to investigate a possible causal effect of maternal lipid concentrations on offspring birth weight. *Int J Epidemiol*. https://doi.org/10.1093/ije/ dyz160

Jackson, J. W., & Swanson, S. A. (2015). Toward a clearer portrayal of confounding bias in instrumental variable applications. *Epidemiology (Cambridge, Mass.)*, *26*(4), 498.

Jokela, M., Elovainio, M., & Kivimäki, M. (2008). Lower fertility associated with obesity and underweight: The us national longitudinal survey of youth. *The American journal of clinical nutrition*, *88*(4), 886–893.

Kang, H., Zhang, A., Cai, T. T., & Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, *111*(513), 132–144.

Koletzko, B., Lien, E., Agostoni, C., Böhles, H., Campoy, C., Cetin, I., Decsi, T., Dudenhausen, J. W., Dupont, C., & Forsyth, S. (2008). The roles of long-chain polyunsaturated fatty acids in pregnancy, lactation and infancy: Review of current knowledge and consensus recommendations. *Journal of perinatal medicine*, *36*(1), 5–14.

Korevaar, T. I. M., Steegers, E. A. P., Schalekamp-Timmermans, S., Ligthart, S., de Rijke, Y. B., Visser, W. E., Visser, W., de Muinck Keizer-Schrama, S. M. P. F., Hofman, A., & Hooijkaas, H. (2014). Soluble flt1 and placental growth factor are novel determinants of newborn thyroid (dys) function: The generation r study. *The Journal of Clinical Endocrinology and Metabolism*, *99*(9), E1627–E1634.

Laanpere, M., Altmäe, S., Stavreus-Evers, A., Nilsson, T. K., Yngve, A., & Salumets, A. (2010). Folate-mediated one-carbon metabolism and its effect on female fertility and pregnancy viability. *Nutrition reviews*, *68*(2), 99–113.

Labrecque, J., & Swanson, S. A. (2018). Understanding the assumptions underlying instrumental variable analyses: A brief review of falsification strategies and related tools. *Current Epidemiology Reports*, 1–7.

Lawlor, D. A., Timpson, N. J., Harbord, R. M., Leary, S., Ness, A., McCarthy, M. I., Frayling, T. M., Hattersley, A. T., & Smith, G. D. (2008). Exploring the developmental overnutrition hypothesis using parental-offspring associations and fto as an instrumental variable. *PLoS Med*, *5*(3), 0484–0493. https://doi.org/10.1371/journal.pmed.0050033

Lawlor, D., Richmond, R., Warrington, N., McMahon, G., Smith, G. D., Bowden, J., & Evans, D. M. (2017). 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*, *2*.

Lee, H. A., Park, E. A., Cho, S. J., Kim, H. S., Kim, Y. J., Lee, H., Gwak, H. S., Kim, K. N., Chang, N., Ha, E. H., & Park, H. (2013). Mendelian randomization analysis of the effect of maternal homocysteine during pregnancy, as represented by maternal mthfr c677t genotype, on birth weight. *J Epidemiol*, *23*(5), 371–375. https://doi.org/10.2188/jea.JE20120219

Lewis, S. J., Bonilla, C., Brion, M. J., Lawlor, D. A., Gunnell, D., Ben-Shlomo, Y., Ness, A., & Smith, G. D. (2014). Maternal iron levels early in pregnancy are not associated with offspring iq score at age 8, findings from a mendelian randomization study. *Eur J Clin Nutr*, *68*(4), 496–502. https://doi.org/10.1038/ejcn.2013.265

Lewis, S. J., Leary, S., Smith, G. D., & Ness, A. (2009). Body composition at age 9 years, maternal folate intake during pregnancy and methyl-tetrahydrofolate reductase (mthfr) c677t genotype. *British journal of nutrition.*

Lewis, S. J., Zuccolo, L., Smith, G. D., Macleod, J., Rodriguez, S., Draper, E. S., Barrow, M., Alati, R., Sayal, K., & Ring, S. (2012). Fetal alcohol exposure and iq at age 8: Evidence from a population-based birth-cohort study. *PloS one*, *7*(11), e49407.

Linnet, K. M., Dalsgaard, S., Obel, C., Wisborg, K., Henriksen, T. B., Rodriguez, A., Kotimaa, A., Moilanen, I., Thomsen, P. H., & Olsen, J. (2003). Maternal lifestyle factors in pregnancy risk of attention deficit

hyperactivity disorder and associated behaviors: Review of the current evidence. *American Journal of Psychiatry*, *160*(6), 1028–1040.

Mamasoula, C., Prentice, R. R., Pierscionek, T., Pangilinan, F., Mills, J. L., Druschel, C., Pass, K., Russell, M. W., Hall, D., Topf, A., Brown, D. L., Zelenika, D., Bentham, J., Cosgrove, C., Bhattacharya, S., Riveron, J. G., Setchfield, K., Brook, J. D., Bu'Lock, F. A., . . . Keavney, B. D. (2013). Association between c677t polymorphism of methylene tetrahydrofolate reductase and congenital heart disease: Meta-analysis of 7697 cases and 13 125 controls. *Circ.-Cardiovasc. Genet.*, *6*(4), 347–353.

Marques, A. H., O'Connor, T. G., Roth, C., Susser, E., & Bjørke-Monsen, A.-L. (2013). The influence of maternal prenatal and early childhood nutrition and maternal prenatal stress on offspring immune system development and neurodevelopmental disorders. *Frontiers in neuroscience*, *7*.

Morales, E., Guerra, S., & García-Esteban, R. (2011). Maternal c-reactive protein levels in pregnancy are associated with wheezing and lower respiratory tract infections in the offspring. *American journal of . . .*

Morales, E., Vilahur, N., Salas, L. A., Motta, V., Fernandez, M. F., Murcia, M., Llop, S., Tardon, A., Fernandez-Tardon, G., Santa-Marina, L., Gallastegui, M., Bollati, V., Estivill, X., Olea, N., Sunyer, J., & Bustamante, M. (2016). Genome-wide dna methylation study in human placenta identifies novel loci associated with maternal smoking during pregnancy. *Int J Epidemiol*, *45*(5), 1644–1655.

Murray, J., Burgess, S., Zuccolo, L., Hickman, M., Gray, R., & Lewis, S. J. (2016). Moderate alcohol drinking in pregnancy increases risk for children's persistent conduct problems: Causal effects in a mendelian randomisation study. *Journal of child psychology and psychiatry*, *57*(5), 575–584.

Palmer, T. M., Lawlor, D. A., Harbord, R. M., Sheehan, N. A., Tobias, J. H., Timpson, N. J., Smith, G. D., & Sterne, J. A. C. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical methods in medical research*, *21*(3), 223–242.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 435–443.

Richmond, R. C., Sharp, G. C., Ward, M. E., Fraser, A., Lyttleton, O., McArdle, W. L., Ring, S. M., Gaunt, T. R., Lawlor, D. A., Smith, G. D., & Relton, C. L. (2016). Dna methylation and bmi: Investigating identified methylation sites at hif3a in a causal framework. *Diabetes*, *65*(5), 1231–1244. https://doi.org/10.2337/db15-0996

Richmond, R. C., Timpson, N. J., Felix, J. F., Palmer, T., Gaillard, R., McMahon, G., Davey Smith, G., Jaddoe, V. W., & Lawlor, D. A. (2017). Using genetic variation to explore the causal effect of maternal pregnancy adiposity on future offspring adiposity: A mendelian randomisation study. *PLoS Med*, *14*(1). https://doi.org/10.1371/journal.pmed.1002221

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

Rowe, H., Baker, T., & Hale, T. W. (2013). Maternal medication, drug use, and breastfeeding. *Pediatric Clinics*, *60*(1), 275–294.

Sacks, K. N., Friger, M., Shoham-Vardi, I., Abokaf, H., Spiegel, E., Sergienko, R., Landau, D., & Sheiner, E. (2016). Prenatal exposure to gestational diabetes mellitus as an independent risk factor for long-term neuropsychiatric morbidity of the offspring. *American journal of obstetrics and gynecology*, *215*(3), 380. e1–380. e7.

Scholder, S. V. K., Wehby, G. L., Lewis, S., & Zuccolo, L. (2014). Alcohol exposure in utero and child academic achievement. *Econ. J.*, *124*(576), 634–667.

Shaheen, S. O., Rutterford, C., Zuccolo, L., Ring, S. M., Davey Smith, G., Holloway, J. W., & Henderson, A. J. (2014). Prenatal alcohol exposure and childhood atopic disease: A mendelian randomization approach. *J Allergy Clin Immunol*, *133*(1), 225–232.e5. https://doi.org/10.1016/j.jaci.2013.04.051

Smith, G. D. (2008). Assessing intrauterine influences on offspring health outcomes: Can epidemiological studies yield robust findings? *Basic Clin Pharmacol Toxicol*, *102*(2), 245–256. https://doi.org/10.1111/j.1742-7843.2007.00191.x

Smith, G. D. (2010). Mendelian randomization for strengthening causal inference in observational studies: Application to gene× environment interactions. *Perspectives on Psychological Science*.

Soderborg, T. K., Borengasser, S. J., Barbour, L. A., & Friedman, J. E. (2016). Microbial transmission from mothers with obesity or diabetes to infants: An innovative opportunity to interrupt a vicious cycle. *Diabetologia*, *59*(5), 895–906.

Steer, C. D., & Tobias, J. H. (2011). Insights into the programming of bone development from the avon longitudinal study of parents and children (alspac). *The American journal of clinical nutrition*.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*, *20*(4), 518–529.

Swanson, S. A. (2017). Commentary: Can we see the forest for the ivs? mendelian randomization studies with multiple genetic variants. *Epidemiology*, *28*(1), 43–46.

Swanson, S. A. (2019). A practical guide to selection bias in instrumental variable analyses. *Epidemiology*, *30*(3), 345–349.

Swanson, S. A., & Hernán, M. A. (2013). Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, *24*(3), 370–374.

Swanson, S. A., Labrecque, J., & Hernán, M. A. (2018). Causal null hypotheses of sustained treatment strategies: What can be tested with an instrumental variable? *European journal of epidemiology*, 1–6.

Swanson, S. A., Tiemeier, H., Ikram, M. A., & Hernán, M. A. (2017). Nature as a trialist? *Epidemiology*, *28*(5), 653–659.

Taylor, A. E., Davies, N. M., Ware, J. J., VanderWeele, T., Smith, G. D., & Munafo, M. R. (2014). Mendelian randomization in health research: Using appropriate genetic variants and avoiding biased estimates. *Econ. Hum. Biol.*, *13*, 99–106.

Tchetgen, E. J. T., Sun, B., & Walter, S. (2017). The genius approach to robust mendelian randomization inference. *arXiv preprint arXiv:1709.07779*.

Thompson, W. D., Tyrrell, J., Borges, M. C., Beaumont, R. N., Knight, B. A., Wood, A. R., Ring, S. M., Hattersley, A. T., Freathy, R. M., & Lawlor, D. A. (2019). Association of maternal circulating 25(oh)d and calcium with birth weight: A mendelian randomisation analysis. *PLoS Med*, *16*(6). https://doi.org/10.1371/journal.pmed.1002828

Tyrrell, J., Richmond, R. C., Palmer, T. M., Feenstra, B., Rangarajan, J., Metrustry, S., Cavadino, A., Paternoster, L., Armstrong, L. L., De Silva, N. M. G., Wood, A. R., Horikoshi, M., Geller, F., Myhre, R., Bradfield, J. P., Kreiner-Møller, E., Huikari, I., Painter, J. N., Hottenga, J. J., ... Freathy, R. M. (2016). Genetic evidence for causal relationships between maternal obesity-related traits and birth weight. *JAMA*, *315*(11), 1129–1140. https://doi.org/10.1001/jama.2016.1975

VanderWeele, T. J. (2012). Invited commentary: Structural equation models and epidemiologic analysis. *American journal of epidemiology*, *176*(7), 608–612.

VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., & Kraft, P. (2014). Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.)*, *25*(3), 427.

Warrington, N. M., Freathy, R. M., Neale, M. C., & Evans, D. M. (2018). Using structural equation modelling to jointly estimate maternal and fetal effects on birthweight in the uk biobank. *International journal of epidemiology*, *47*(4), 1229–1241.

Wathes, D. C., Abayasekara, D. R. E., & Aitken, R. J. (2007). Polyunsaturated fatty acids in male and female reproduction. *Biology of reproduction*, *77*(2), 190–201.

Waylen, A. L., Metwally, M., Jones, G. L., Wilkinson, A. J., & Ledger, W. L. (2008). Effects of cigarette smoking upon clinical outcomes of assisted reproduction: A meta-analysis. *Human reproduction update*, *15*(1), 31–44.

Wehby, G. L., Fletcher, J. M., Lehrer, S. F., Moreno, L. M., Murray, J. C., Wilcox, A., & Lie, R. T. (2011). A genetic instrumental variables analysis of the effects of prenatal smoking on birth weight: Evidence from two samples. *Biodem Soc Biol*, *57*(1), 3–32. https://doi.org/10.1080/19485565.2011.564468

Wehby, G. L., Jugessur, A., Murray, J. C., Moreno, L. M., Wilcox, A., & Lie, R. T. (2011). Genes as instruments for studying risk behavior effects: An application to maternal smoking and orofacial clefts. *Health Serv Outcomes Res Methodol*, *11*(1-2), 54–78. https://doi.org/10.1007/s10742-011-0071-9

Wehby, G. L., & Scholder, S. V. H. K. (2013). Genetic instrumental variable studies of effects of prenatal risk factors. *Biodem Soc Biol*, *59*(1), 4–36. https://doi.org/10.1080/19485565.2013.774615

Wesselink, A. K., Wise, L. A., Rothman, K. J., Hahn, K. A., Mikkelsen, E. M., Mahalingaiah, S., & Hatch, E. E. (2016). Caffeine and caffeinated beverage consumption and fecundability in a preconception cohort. *Reproductive Toxicology*, *62*, 39–45.

Whitworth, K. W., Baird, D. D., Stene, L. C., Skjaerven, R., & Longnecker, M. P. (2011). Fecundability among women with type 1 and type 2 diabetes in the norwegian mother and child cohort study. *Diabetologia*, *54*(3), 516–522.

Yadegari, M., Khazaei, M., Anvari, M., & Eskandari, M. (2016). Prenatal caffeine exposure impairs pregnancy in rats. *International journal of fertility and sterility*, *9*(4), 558.

Yajnik, C. S., Chandak, G. R., Joglekar, C., Katre, P., Bhat, D. S., Singh, S. N., Janipalli, C. S., Refsum, H., Krishnaveni, G., Veena, S., Osmond, C., & Fall, C. H. D. (2014). Maternal homocysteine in pregnancy and offspring birthweight: Epidemiological associations and mendelian ran-

domization analysis. *Int J Epidemiol*, *43*(5), 1487–1497. https://doi.org/10.1093/ije/dyu132

Young, B. E., Patinkin, Z., Palmer, C., de la Houssaye, B., Barbour, L. A., Hernandez, T., Friedman, J. E., & Krebs, N. F. (2017). Human milk insulin is related to maternal plasma insulin and bmi: But other components of human milk do not differ by bmi. *European Journal of Clinical Nutrition*.

Zerbo, O., Traglia, M., Yoshida, C., & Heuer, L. S. (2016). *Maternal mid-pregnancy c-reactive protein and risk of autism spectrum disorders: The early markers for autism study*. ncbi.nlm.nih.gov.

Zhang, G., Bacelis, J., Lengyel, C., Teramo, K., Hallman, M., Helgeland, Ø., Johansson, S., Myhre, R., Sengpiel, V., Njølstad, P. å., Jacobsson, B., & Muglia, L. (2015). Assessing the causal relationship of maternal height on birth size and gestational age at birth: A mendelian randomization analysis. *PLoS Med*, *12*(8). https://doi.org/10.1371/journal.pmed.1001865

Zuccolo, L., Lewis, S. J., Smith, G. D., Saya, K., Draper, E. S., Fraser, R., Barrow, M., Alati, R., Ring, S., Macleod, J., Golding, J., Heron, J., & Gray, R. (2013). Prenatal alcohol exposure and offspring cognition and school performance. a mendelian randomization natural experiment. *Int J Epidemiol*, *42*(5), 1358–1370. https://doi.org/10.1093/ije/dyt172

# Chapter 4

# Application of the Instrumental Inequalities to a Mendelian Randomization Study With Multiple Proposed Instruments

Elizabeth W. Diemer, Jeremy A. Labrecque, Henning Tiemeier, Sonja A. Swanson

## 4.1 Abstract

**Background**: Investigators often support the validity of Mendelian randomization (MR) studies, an instrumental variable approach proposing genetic variants as instruments, via subject matter knowledge. However, the instrumental variable model implies certain inequalities, offering an empirical method of falsifying (but not verifying) the underlying assumptions. While these inequalities are said to detect only extreme assumption violations in practice, to our knowledge they have not been used in settings with multiple proposed instruments.
**Methods**: We applied the instrumental inequalities to an MR analysis of the effect of maternal pregnancy Vitamin D on offspring psychiatric outcomes, proposing four independent maternal genetic variants as instruments. We assessed whether the proposed instruments satisfied the instrumental inequalities separately and jointly and explored the instrumental inequalities' properties via simulations.
**Results**: The instrumental inequalities were satisfied (i.e., we did not falsify the MR model) when considering each variant separately. However, the inequalities were violated when considering four variants jointly and for some combinations of two or three variants (2 of 36 two-variant combinations and 18 of 24 three-variant combinations). In simulations, the inequalities detected structural biases more often when assessing proposed instruments jointly, while falsification in the absence of structural bias remained rare.
**Conclusions**: The instrumental inequalities detected violations of the MR assumptions for genetic variants jointly proposed as instruments in our study, though the instrumental inequalities were satisfied when considering each proposed instrument separately. We discuss how investigators can assess instrumental inequalities to eliminate clearly invalid analyses in settings with many proposed instruments.

## 4.2 Introduction

Mendelian randomization (MR), an increasingly popular tool for studying causal effects even when unmeasured confounding appears insurmountable, is a type of instrumental variable (IV) model where genetic variants are proposed as instruments. Briefly, a valid MR analysis with one genetic variant requires:

1. The genetic variant $Z$ is associated with the exposure $X$

2. The genetic variant $Z$ does not affect the outcome $Y$ except through its effect on the exposure $X$

3. Individuals at different levels of the genetic variant $Z$ are exchangeable (i.e., comparable) with regard to counterfactual outcome

Conditions 2 and 3 are unverifiable. Forms of these conditions are necessary but not usually sufficient for all versions of MR analyses: obtaining point estimates of an average causal effect requires additional assumptions (Hernán and Robins, 2006), although these three conditions suffice for estimating bounds and sharp causal null testing (Balke and Pearl, 1997; Manski, 1990; Robins, 1989).
Frequently, MR analyses propose that multiple single nucleotide polymorphisms (SNPs) act as instruments and therefore that those SNPs jointly satisfy the MR assumptions. Leveraging multiple proposed instruments mitigates issues with power and weak instrument biases that can arise in analyses with a single proposed instrument (Burgess and Thompson, 2013; Pierce et al., 2011), though investigators are then challenged to support that the MR assumptions are satisfied for each SNP and for all SNPs jointly. As many genetic loci jointly proposed as instruments are derived from genome-wide association studies and the exact biologic mechanisms are often poorly understood, it is likely that these required assumptions do not hold for many MR analyses. Given this, several recently developed estimators allow for specific relaxations in exchange for additional, different assumptions (Bowden et al., 2015; Bowden et al., 2016; Kang et al., 2016; Tchetgen et al., 2017; Verbanck et al., 2018; Zhu et al., 2018) . For example, some approaches only require a subset of proposed instruments are true instruments (Bowden et al., 2016; Hartwig et al., 2017).

Often missing from the MR literature, however, is any discussion of whether the data are consistent with the MR model proposed. Over two decades ago, Pearl showed that the IV assumptions imply the following inequality for discrete proposed instruments, exposures, and outcomes:

$max_{i=1..n} \sum_{j=1}^{m} max_{k=1..l} P(X = i, Y = j | Z = k) \leq 1$ (Pearl, 1995),

which is equivalent to the set of inequalities resulting from

$\sum_{j=1}^{m} P(X = i, Y = j | Z = k_{i,j}) \leq 1$

for all $1 \leq i \leq n, k_{i,j} \in 1, ..., l$ (Balke and Pearl, 1997).

For an IV model with an exposure with $n$ levels, an outcome with $m$ levels, and a proposed instrument with $l$ levels, this equation will result in a set of $nl^m$ inequalities. Later, Bonet proved the IV model also implies additional constraints, and that such inequalities can be generalized to settings in which the proposed instrument and outcome, but not the exposure, are continuous (Bonet, 2001). Although Bonet's additional constraints are often difficult to state with straightforward equations, he did provide one expression for the case of a trichotomous instrument, dichotomous exposure, and dichotomous outcome:

$P(X = 1, Y = 2 | Z = 2) + P(X = 1, Y = 1 | Z = 3) + P(X = 1, Y = 2 | Z = 1) + P(X = 2, Y = 2 | Z = 2) + P(X = 2, Y = 1 | Z = 1) \leq 2$ (Bonet, 2001).

If the inequalities presented by Pearl and Bonet, known as instrumental inequalities, do not hold, the IV model cannot hold (Bonet, 2001; Pearl, 1995). This means that investigators can attempt to falsify the IV model with their data alone when they have a dataset with measures of the proposed instrument, exposure, and outcome: if the instrumental inequalities are not satisfied, the data tell us that one or more of our assumptions are not satisfied. Recognizing the importance of falsification strategies (when available) for causal inference, multiple reporting guidelines recommend assessing the instrumental inequalities in all IV analyses (Glymour et al., 2012; Labrecque and Swanson, 2018; Swanson and Hernán, 2013). Despite this, few MR analyses use them, perhaps because, for dichotomous proposed instruments, it has been suggested that only extreme assumption violations will be detected in practice (Glymour et al., 2012; Swanson and Hernán, 2013). No study has applied the instrumental inequalities to investigate the validity of multiple genetic loci jointly proposed as instruments. Here, we aim to explore the utility of the instrumental inequalities in identifying violations of the assumptions required for MR with multiple proposed instruments in real and simulated data, and to provide adaptable software for the implementation and visualization of the instrumental inequalities. We begin by describing how to interpret the results of the instrumental inequalities when applied to a specific MR model and dataset.

## 4.3 Interpretation of the instrumental inequalities

**A**
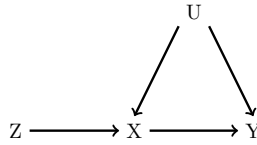


**B**
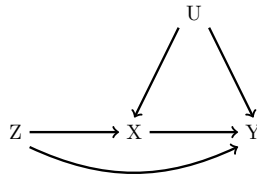


Figure 4.1: DAGs representing an MR study with one genetic variant $Z$, proposed as an instrument for the effect of $X$ on $Y$. In **A**, $Z$ is a valid instrument. In **B**, the MR assumptions are violated by a direct effect of $Z$ on $Y$.

Because such falsification tests are relatively uncommon, let us begin by considering for illustrative purposes a scenario in which we believe that the two causal diagrams in Figure 4.1 are the only possible relationships between a particular SNP, exposure, and outcome. If the instrumental inequalities failed to hold, 4.1A could not be true, meaning that 4.1B must be true and the SNP has a direct effect on the outcome. However, if the instrumental inequalities hold, the data are consistent with the SNP having a direct effect or having no direct effect on the outcome, as we have failed to falsify 4.1A.

The same logic applies where multiple SNPs are believed to be instruments. Figure 4.2 presents a causal diagram in which four independent SNPs are valid instruments both individually and as a single joint variable. When multiple SNPs are available, MR analyses using different subsets of SNPs, and thus slightly different assumptions, can be proposed. As such, the instrumental inequalities can be applied to each SNP individually, to any combination of two, three, or four of the SNPs, or to a summary score derived from these SNPs (e.g., an allele score) to evaluate the validity of each subset as a (jointly) proposed instrument. For example, one could propose all four SNPs jointly as

instruments by combining the SNPs into a $3^4 = 81$ level variable, where each level represents a different possible combination of alleles for the four SNPs. Violations of the instrumental inequalities when proposing this combination variable as an instrument provide evidence against the causal diagram in Figure 4.2. Likewise, violations of the instrumental inequalities when considering any SNP individually or any subset of SNPs would also provide evidence against this particular causal diagram.
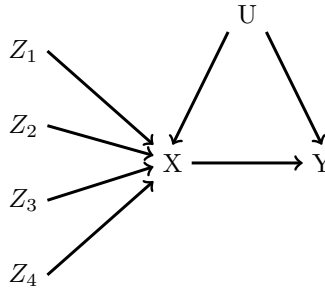


Figure 4.2: DAG representing an MR study four independent genetic variants $Z_1$, $Z_2$, $Z_3$, and $Z_4$, proposed as instruments for the effect of $X$ on $Y$. Here, all four genetic variants are valid instruments individually and jointly.

It is possible to apply the instrumental inequalities directly to an allele score. Violations of the instrumental inequalities when proposing this allele score as an instrument could also provide evidence against the causal diagram in Figure 4.2. However, allele scores imply additional linearity and additivity assumptions, which are not required for the use of MR or the instrumental inequalities, and may result in loss of power (Pierce et al., 2011), though this approach may be useful to investigators considering using the allele score in their particular MR analysis.

Importantly, the instrumental inequalities do not actually require us to specify an alternative causal diagram like we did in Figure 4.1. The instrumental inequalities simply show us whether a proposed MR model is false. In fact, without additional assumptions, the instrumental inequalities do not give evidence as to how the MR assumptions are violated, only that the MR model cannot be true in the dataset.

In practice, the usefulness of the instrumental inequalities for evaluating many proposed instruments may be hindered by sample size. As the number of SNPs jointly proposed as instruments increases, the number of individuals within a

138

given stratum of the proposed joint instrument becomes increasingly small, and it becomes more likely that the instrumental inequalities will fail to hold by random chance. The concept of random violations of MR assumptions is similar to that of "random confounding" (Greenland and Mansournia, 2015; Hernan and Robins, 2018): in randomized trials, although randomization implies we expect balance of covariates across trial arms on average, it does not guarantee balance within a particular study. If there are imbalances in the distribution of a risk factor for the outcome in a study, adjustment for the imbalanced risk factor is recommended to produce unbiased causal effect estimates. Analogously, even if the MR assumptions for a proposed joint instrument are met in a theoretical super-population, the distribution of the proposed instrument, exposure, and outcome within a particular sample might deviate substantially from the expected distribution in the super-population, especially in small samples, which are more prone to notable deviations from what is expected. As a result, the MR assumptions, and thus, the instrumental inequalities could fail to hold by chance. Such violations may occur more often in small samples. As in a randomized trial with "random confounding", an MR analysis in a sample where the assumptions were violated by chance is expected to produce biased estimates of causal effects. Furthermore, in contrast to testing for chance imbalances of risk factors in randomized trials, the source of a violation of the instrumental inequalities cannot be determined without further assumptions. Thus, any evidence of a violation of the MR assumptions should be considered as important evidence about the validity of an MR analysis for that specific dataset. It remains important to understand the impact of sample size on the ability to detect structural violations of the MR assumptions, as it would otherwise remain unclear whether a violation found in one dataset provides evidence against a similar MR model in another dataset.

The application of the instrumental inequalities to multiple proposed instruments allows for many layers of falsification strategies: we can attempt to falsify the model for any proposed instrument individually, any combination of proposed instruments jointly, and any summary score. A potential advantage of applying the instrumental inequalities to each of these is that they might be used to identify subsets of SNPs for which the MR assumptions definitely do not hold, and subsets of SNPs where an MR analysis could be pursued with caution.

In the next section, we explore this possibility in a study of the effects of maternal prenatal Vitamin D levels on childhood behavioral health outcomes, and introduce a new visualization for the instrumental inequalities. We follow this application with a simulation study in order to better understand

the impact of sample size on the instrumental inequalities. All analyses were conducted in R 3.4.1 (Team, 2020). We provide adaptable R functions that allow the user to calculate the instrumental inequalities for multiple proposed instruments and display the results in a novel graph format. These functions, which appear in the supplement to the published version of this article (https://journals.lww.com/epidem/Fulltext/2020/01000/Application_of_the_Instrumental_Inequalities_to_a.7.aspx), are omitted from this dissertation for the sake of brevity.

## 4.4 Data example: Estimating the effects of maternal pregnancy vitamin D on childhood behavioral health outcomes in Generation R

**Study population**

Generation R is a population-based cohort from fetal life to young adulthood, based in Rotterdam, the Netherlands. Mothers with a delivery date between April 2002 and January 2006 who lived in the study area were eligible for participation. Further information about the study is available elsewhere (Jaddoe et al., 2010). In total, 8,880 mothers were enrolled during pregnancy. To avoid overt violation of the MR assumptions by population stratification or relatedness, we restrict our analysis to the 3,188 mother-child pairs for which mothers were of self-reported Dutch ancestry and the child was the first offspring of the mother included in the cohort. For each MR model investigated, analysis was restricted to individuals with complete data available on exposure, outcome, and all proposed instruments, resulting in analytic samples of 1,970 (pervasive developmental problems[PDP]), 1,971(mother-reported attention deficit-hyperactivity disorder [ADHD] symptoms), and 1,146 (teacher-reported ADHD symptoms) for each outcome studied, respectively (see Appendix for descriptive statistics). This complete case analysis approach aligns with common practices in MR analyses, but it can violate the MR assumptions (and in fact may be the reason for violations of the instrumental inequalities in these samples) (Canan et al., 2017; Swanson, 2019). Future studies might mitigate this issue by conducting the instrumental inequalities and MR models in samples weighted by the inverse probability of selection (Canan et al., 2017). The study was approved by the Medical Ethics Committee of Erasmus Medical Center and was

in accordance with the World Medical Association Declaration of Helsinki.

**Proposed Instruments**

Maternal genotyping was performed using Taqman allelic discrimination assay (Applied Biosystems, Foster City, CA), with an error rate of less than 1% confirmed in a random subsample (n=276) (Kruithof et al., 2014). Based on existing literature, we proposed four independent maternal SNPs (rs2282679, rs12785878, rs6013897, rs10741657) as instruments. These SNPs have been associated genome-wide with serum vitamin D in a sample of 42,274 individuals (Wang et al., 2010), and are often used in MR studies of vitamin D (Mokry et al., 2015; Ong et al., 2016; Vimaleswaran et al., 2013). For all models, we coded SNPs trichotomously, based on the presence of 0,1, or 2 risk alleles.

**Exposure**

Pregnancy serum vitamin D status was defined using the storage form of vitamin D, total 25OHD, measured in venous blood taken between 18.1 and 24.9 weeks gestation (Vinkhuyzen et al., 2016). We defined exposure dichotomously and trichotomously, based on established clinical cutoffs at which treatment for vitamin D is recommended (Holick, 2009; Holick et al., 2011; Vieth, 2011). Total serum 25OHD was dichotomized at 75 nmol/L based on sufficiency; and trichotomized as deficiency (0-50 nmol/L), insufficiency (50-74.99 nmol/L), and sufficiency ($\geq$ 75 nmol/L). While these categorizations imply strong assumptions about a step-function relationship between vitamin D and offspring behavioral health, it is important to recognize that modeling vitamin D continuously in MR typically makes a likewise strong and potentially inaccurate assumption of a linear relationship.

**Outcomes**

Maternal-reported pervasive developmental disorder (PDD) and attention deficit hyperactivity disorder (ADHD) symptoms at age 5 years were assessed from the Persistent Developmental Problems and the Attention Deficit-Hyperactivity subscales, respectively, of the Dutch translation of the Child Behavior Checklist (Achenbach and Rescorla, 2000; Tick et al., 2007). The former subscale has been used as a screening tool to identify children with autism spectrum disorder (Sikora et al., 2008), while the latter has shown good convergent validity with clinician ratings (Hudziak et al., 2004; Soma et al., 2009). We used the 98th percentile of each subscale's T-scores (PDD: $T \geq 8.98$; ADHD: $T \geq 9$) as cutoffs to classify children with mother-reported PDP and ADHD symptoms in the clinical range. Teacher-reported ADHD symptoms at age 7 were defined as a T-score above the 98th percentile on

the Teacher Report Form Attention Problems subscale (T $\geq$ 15) (Achenbach, 1991; de Groot et al., 1996; Verhulst et al., 1985).

### 4.4.1 Analysis

We assessed whether the instrumental inequalities would identify violations of MR models for the causal effect of maternal serum vitamin D during pregnancy on offspring PDP and ADHD symptoms, using the above-mentioned four SNPs proposed as instruments. For each possible combination of SNPs, we applied the instrumental inequalities to MR models for the causal effect of maternal vitamin D on an outcome. We then extracted the maximum value of the instrumental inequalities, along with the number of strata of the proposed instrument with exactly 0 or fewer than 10 individuals. For binary exposure models, we also applied the Bonet inequality for trichotomous instruments to each SNP marginally. Although in any plausible scenario where an allele score satisfies the MR assumptions, each contributing SNP would also individually and jointly satisfy those assumptions (Burgess and Thompson, 2013), we also applied the instrumental inequalities to MR models with a categorical, unweighted allele score proposed as an instrument.

Though the instrumental inequalities cannot be applied to continuous measures of exposures, evaluating models based on categorized measures could still be informative. However, the MR assumptions can be violated if the exposure is inappropriately categorized (VanderWeele et al., 2014), implying the instrumental inequalities might be detecting this mismeasurement rather than another MR assumption violation. If that were the case, we may expect to see decreasing instances in which the instrumental inequalities were violated as the number of categories of the exposure increases, though evaluating this property might require prohibitively large samples. To see if coding of the exposure variable altered the conclusions, we evaluated the instrumental inequalities using dichotomous and trichotomous exposure definitions, as described above.

### 4.4.2 Results

For all definitions of exposures and outcomes, the instrumental inequalities, including the stronger inequalities developed by Bonet, held for each SNP individually, indicating that there was no evidence in the data alone against each specific proposed instrument being valid. However, as the number of

SNPs jointly proposed as instruments increased, the instrumental inequalities increasingly failed to hold (Figure 4.3).
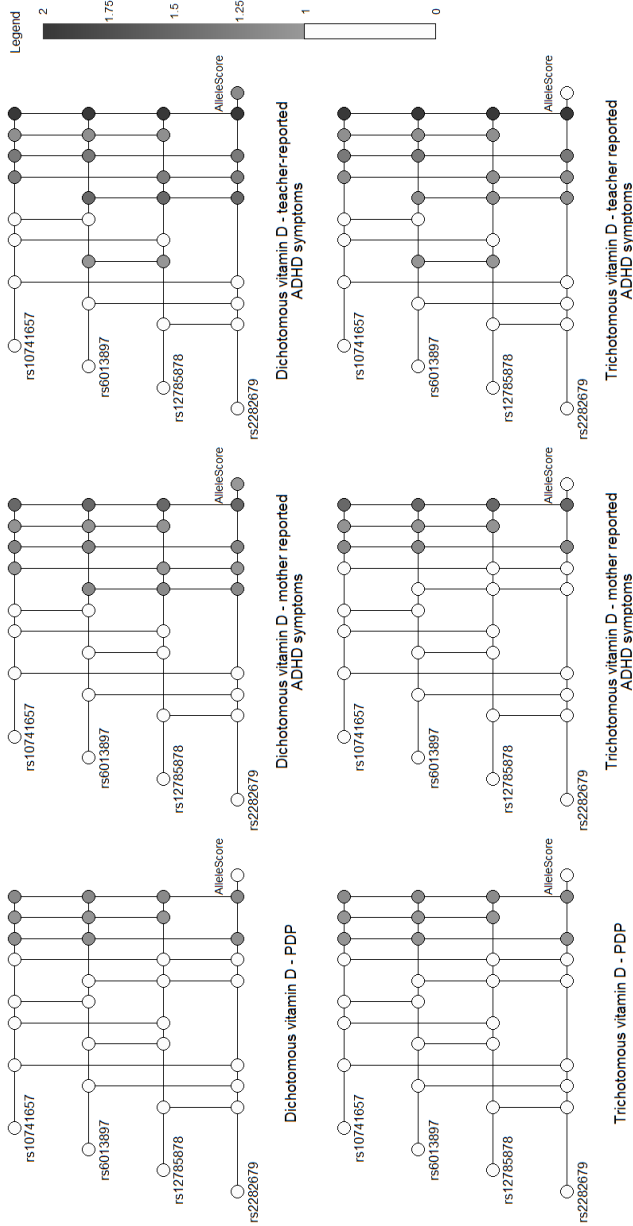
Figure 4.3: In these visualizations, each horizontal line represents a SNP, and each horizontal line connects a set of SNPs proposed as instruments (with the number of included SNPs increasing from left to right). The color of each node represents the maximum value of the instrumental inequalities, with white indicating a value less than one and darker colors indicating larger values represent violations (see Legend). See Appendix for further details of visualization technique.

144

When the instrumental inequalities were applied to MR models for the causal effect of maternal vitamin D coded dichotomously on mother-reported PDP symptoms, the instrumental inequalities failed to hold for half of the combinations of three SNPs jointly proposed as instruments and the combination of all four SNPs (Tables 4.1-4.3). When applied to MR models for the causal effect of maternal vitamin D on mother-reported ADHD symptoms, the instrumental inequalities failed to hold for all three SNP and four SNP combinations, as well as the allele score. For teacher-reported ADHD symptoms, the instrumental inequalities failed to hold for the allele score, all three SNP and four SNP combinations, and one two-SNP combination.

When we coded maternal vitamin D trichotomously, the maximum value of the instrumental inequalities for each possible combination of SNPs proposed as instruments was less than or equal to the maximum value of the inequalities in models with a dichotomized measure of maternal vitamin D. For some models, the instrumental inequalities held in the trichotomous exposure case but not the dichotomous exposure case, including two settings in which the allele score was the proposed instrument.

Table 4.1: Summary of the instrumental inequalities for studying the effect of maternal vitamin D on mother-reported pervasive development problems symptoms with varying combinations of proposed instruments and definitions of exposure.

| Proposed Instrument(s) | Nonzero Cell Count/Total Cell Count | Number Cells $\leq 10$ | Instrumental inequalities hold for a binary exposure? | Instrumental inequalities hold for a 3-level exposure? |
| --- | --- | --- | --- | --- |
| rs2282679 | 3/3 | 3 | yes (0.75) | yes (0.46) |
| rs12785878 | 3/3 | 3 | yes (0.63) | yes (0.41) |
| rs6013897 | 3/3 | 3 | yes (0.63) | yes (0.41) |
| rs10741657 | 3/3 | 3 | yes (0.65) | yes (0.43) |
| {rs2282679, rs12785878} | 9/9 | 9 | yes (0.83) | yes (0.57) |
| {rs2282679, rs6013897} | 9/9 | 8 | yes (0.82) | yes (0.52) |
| {rs2282679, rs10741657} | 9/9 | 9 | yes (0.87) | yes (0.54) |
| {rs12785878, rs6013897} | 9/9 | 8 | yes (0.70) | yes (0.58) |
| {rs12785878, rs10741657} | 9/9 | 9 | yes (0.73) | yes (0.47) |
| {rs6013897, rs10741657} | 9/9 | 9 | yes (0.71) | yes (0.58) |
| {rs2282679, rs12785878, rs6013897} | 26/27 | 21 | yes (0.90) | yes (0.73) |
| {rs2282679, rs12785878, rs10741657} | 27/27 | 22 | yes (1.00) | yes (0.83) |
| {rs2282679, rs6013897, rs10741657} | 27/27 | 21 | no (1.11) | no (1.06) |
| {rs12785878, rs6013897, rs10741657} | 27/27 | 22 | no (1.04) | no (1.04) |
| {rs2282679, rs12785878, rs6013897, rs10741657} | 73/81 | 35 | no (1.17) | no (1.14) |
| Allele Score | 8/8 | 7 | yes (0.81) | yes (0.54) |

Table 4.2: Summary of the instrumental inequalities for studying the effect of maternal vitamin D on mother-reported attention deficit hyperactivity disorder symptoms with varying combinations of proposed instruments and definitions of exposure

| Proposed Instrument(s) | Nonzero Cell Count/Total Cell Count | Number Cells ≤ 10 | Instrumental inequalities hold for a binary exposure? | Instrumental inequalities hold for a 3-level exposure? |
|---|---|---|---|---|
| rs2282679 | 3/3 | 3 | yes (0.75) | yes (0.46) |
| rs12785878 | 3/3 | 3 | yes (0.63) | yes (0.41) |
| rs6013897 | 3/3 | 3 | yes (0.63) | yes (0.41) |
| rs10741657 | 3/3 | 3 | yes (0.65) | yes (0.44) |
| {rs2282679, rs12785878} | 9/9 | 9 | yes (0.83) | yes (0.59) |
| {rs2282679, rs6013897} | 9/9 | 8 | yes (0.84) | yes (0.52) |
| {rs2282679, rs10741657} | 9/9 | 9 | yes (0.93) | yes (0.59) |
| {rs12785878, rs6013897} | 9/9 | 8 | yes (0.79) | yes (0.57) |
| {rs12785878, rs10741657} | 9/9 | 9 | yes (0.71) | yes (0.49) |
| {rs6013897, rs10741657} | 9/9 | 9 | yes (0.71) | yes (0.59) |
| {rs2282679, rs12785878, rs6013897} | 26/27 | 21 | no (1.17) | yes (1.00) |
| {rs2282679, rs12785878, rs10741657} | 27/27 | 22 | no (1.04) | yes (0.88) |
| {rs2282679, rs6013897, rs10741657} | 27/27 | 21 | no (1.22) | no (1.17) |
| {rs12785878, rs6013897, rs10741657} | 27/27 | 22 | no (1.06) | no (1.06) |
| {rs2282679, rs12785878, rs6013897, rs10741657} | 73/81 | 35 | no (1.50) | no (1.50) |
| Allele Score | 8/8 | 7 | no (1.02) | yes (0.62) |

Table 4.3: Summary of the instrumental inequalities for studying the effect of maternal vitamin D on teacher-reported attention deficit hyperactivity disorder symptoms with varying combinations of proposed instruments and definitions of exposure

| Proposed Instrument(s) | Nonzero Cell Count/Total Cell Count | Number Cells ≤ 10 | Instrumental inequalities hold for a binary exposure? | Instrumental inequalities hold for a 3-level exposure? |
|---|---|---|---|---|
| rs2282679 | 3/3 | 3 | yes (0.73) | yes (0.46) |
| rs12785878 | 3/3 | 3 | yes (0.63) | yes (0.42) |
| rs6013897 | 3/3 | 3 | yes (0.66) | yes (0.43) |
| rs10741657 | 3/3 | 3 | yes (0.64) | yes (0.42) |
| {rs2282679, rs12785878} | 9/9 | 8 | yes (0.79) | yes (0.48) |
| {rs2282679, rs6013897} | 9/9 | 8 | yes (0.92) | yes (0.60) |
| {rs2282679, rs10741657} | 9/9 | 9 | yes (0.88) | yes (0.54) |
| {rs12785878, rs6013897} | 9/9 | 8 | no (1.03) | no (1.03) |
| {rs12785878, rs10741657} | 9/9 | 9 | yes (0.71) | yes (0.49) |
| {rs6013897, rs10741657} | 9/9 | 9 | yes (0.74) | yes (0.50) |
| {rs2282679, rs12785878, rs6013897} | 25/27 | 19 | no (1.50) | no (1.12) |
| {rs2282679, rs12785878, rs10741657} | 27/27 | 19 | no (1.29) | no (1.12) |
| {rs2282679, rs6013897, rs10741657} | 27/27 | 19 | no (1.33) | no (1.33) |
| {rs12785878, rs6013897, rs10741657} | 26/27 | 20 | no (1.11) | no (1.11) |
| {rs2282679, rs12785878, rs6013897, rs10741657} | 68/81 | 25 | no (2.00) | no (2.00) |
| Allele Score | 8/8 | 7 | no (1.15) | yes (0.82) |

## 4.5   Simulation study

### 4.5.1   Methods

We simulated four independent binary genetic variants $Z_1 - Z_4$ with causal effects on the exposure $X$. While $Z_2$, $Z_3$, and $Z_4$ were true causal instruments, $Z_1$ also had a direct causal effect on the outcome Y, thereby violating the MR assumptions. We then applied the instrumental inequalities in scenarios with varying sample sizes (n=1,000; 10,000; 100,000), proposed instrument strengths, and strengths of the direct effect of $Z_1$ on $Y$. Details of simulated parameters are available in the Appendix. Code for the simulations were published in the supplement to the published version of this article, and were omitted from the dissertation for brevity.

### 4.5.2   Results

The instrumental inequalities were increasingly violated for combinations of proposed instruments including $Z_1$ as the strength of violation and number of proposed instruments included in a combination increased (Figure 4). When the strength of violation was relatively weak, the instrumental inequalities were more often violated for combinations including $Z_1$ in the smaller (n=1,000) samples.
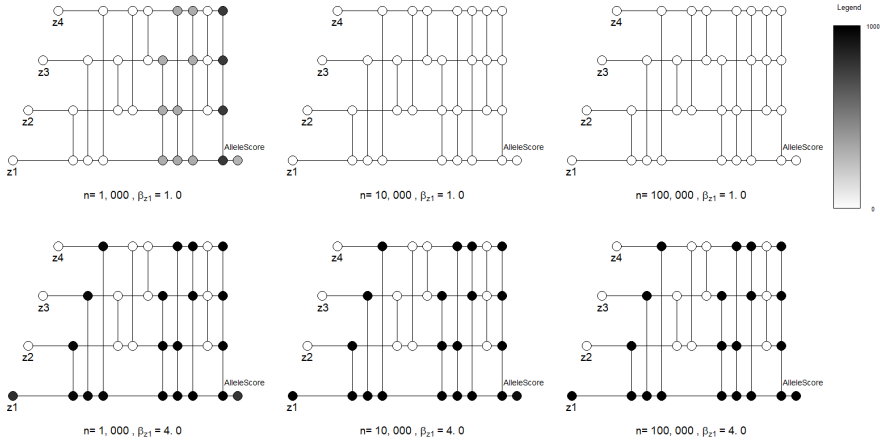
Figure 4.4: Results of six simulations with four dichotomous proposed instruments $Z_1$, $Z_2$, $Z_3$, and $Z_4$, a dichotomous exposure $X$, dichotomous outcome $Y$, and continuous exposure-outcome confounder $U$. For each setting, we simulated 1,000 samples such that $Z_{1i} \sim bernoulli(0.5)$, $Z_{2i} \sim bernoulli(0.5)$, $Z_{3i} \sim bernoulli(0.5)$, $Z_{4i} \sim bernoulli(0.5)$, $U_i \sim norm(0,1)$, $X_i \sim bernoulli(expit[0.6 + 0.1 * U_i + 0.1 * Z_{1i} + 0.1 * Z_{2i} + 0.1 * Z_{3i} + Z_{4i}])$. We varied sample sizes (n= 1,000, 10,000, 100,000) across simulations. In addition, in each of the six simulations, $Z_1$ violated the MR conditions, with $Y_i \sim bernoulli(expit[0.02 + 0.1 * U_i + \beta_{z1} * Z_{1i}])$. Here, each horizontal line represents a single variable, and each vertical line connects a set of variables proposed as joint instruments (with the number of included variables increasing from left to right). Unlike Figure 4.3, here the color of each node represents the number of samples where the instrumental inequalities were violated, out of 1,000 total samples for each setting. This is in contrast to Figure 4.3 where the color of each node represented the maximum value of the inequalities for each set of proposed instruments within a particular dataset. See Appendix for further details.

In samples of 100,000 individuals, the instrumental inequalities were never violated for combinations not including $Z_1$, regardless of instrument strength or strength of violation (Appendix). In simulated samples of 10,000 and 1,000 individuals, the instrumental inequalities were occasionally violated for some combinations not including $Z_1$ (i.e., for combinations when no structural bias was present), though this occurred in less than 1% of simulations for each true instrument marginally (Appendix). This was especially likely when considering the three valid instruments jointly in the smallest sample size and the strongest proposed instrument strength simulated, in which 90% of the time

the inequalities were violated. In all cases in which the inequalities were violated for a combination that did not include $Z_1$, the instrumental inequalities were also always violated for combinations including $Z_1$. When we proposed $Z_1 - Z_4$ jointly as instruments in these settings, the instrumental inequalities were violated in more than 95% of simulations.

## 4.6   Discussion

Our results indicate that, for studies of the causal effect of maternal pregnancy vitamin D on offspring PDP and ADHD within Generation R, there are clear violations of the MR assumptions when proposing four SNPs (rs2282679, rs12785878, rs6013897, rs10741657) jointly as instruments, as well as for several combinations of three of the four SNPs. We did not detect violations of the MR assumptions when each SNP was proposed as an instrument marginally, or for most combinations of two of the four SNPs. The results of our simulations suggest that the instrumental inequalities will be increasingly violated as the magnitude of the violation of the MR assumptions grows, are more sensitive to violations of the MR assumptions when multiple instruments are proposed jointly, and that, within our simulations, small sample sizes appear to increase the probability of finding a true structural violation with limited risk of incorrectly detecting a structural violation when none existed.

Because a violation of the instrumental inequalities for any of the sets of SNPs proposed as instruments would indicate that the four SNPs are not jointly valid instruments, our results clearly demonstrate that certain MR analyses would be biased if conducted in our dataset. Moreover, for teacher-reported and mother-reported ADHD using a dichotomous exposure, the MR assumptions fail to hold when every possible overlapping combination of three of the four SNPs are proposed jointly as instruments, which for independent SNPs logically implies that the MR assumptions cannot hold for at least two of the included SNPs individually. Altogether, our results then suggest that MR analyses which require all four SNPs to be jointly instruments (e.g., analyses proposing an allele score) are inappropriate in our dataset, and also that MR analyses that only require a subset of SNPs to be instruments (e.g., the median-based approach [Bowden et al., 2016]) should be pursued with extreme caution. Our dataset found no particular pattern suggestive of a specific problematic SNP, and thus is not helpful in pruning clearly invalid instruments. On the other hand, our simulations suggest that a pattern consistent with one "bad apple" is possible to detect and may aid in pruning clearly invalid instruments: investigators might

consider removing the offending SNP from their proposed instrument set and continuing with an MR analysis. It is also possible for investigators to consider MR estimators that allow for all proposed instruments to be invalid in specific ways, although these methods require alternative assumptions beyond those considered here (Bowden et al., 2015; Tchetgen et al., 2017) and the results of the instrumental inequalities would only be informative if coupled with a strong biologic rationale for these alternative assumptions. Finally, it is worth reiterating two important points on interpretation. First, the instrumental inequalities falsify but do not verify the MR model. Thus, if an application of the inequalities detects no violation it is still possible for the MR analysis to be biased. Investigators should still weigh subject matter knowledge, perform other falsification strategies and sensitivity analyses, and choose an appropriate method if they decide to pursue an MR analysis, as outlined in prior guidelines (Swanson and Hernán, 2013). The relevance of this point is underscored by our simulations, in which a bias was always structurally present but remained undetected in several simulated samples. Second, the instrumental inequalities are a falsification strategy for the core MR assumptions but do not assess the additional point-identifying assumptions (Glymour et al., 2012).
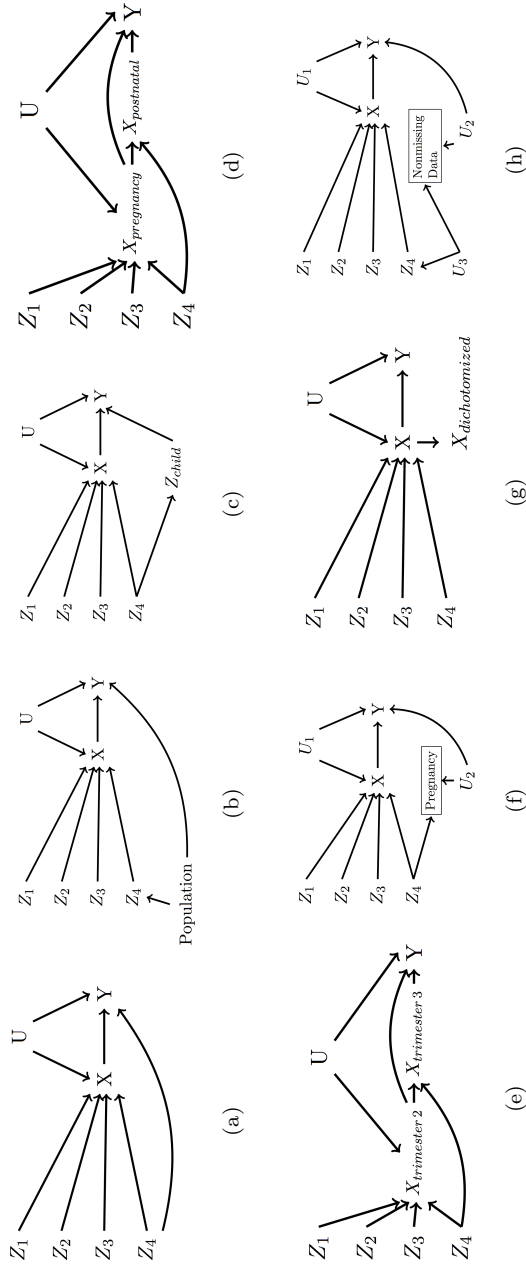
Figure 4.5: DAGs depicting some reasons for possible violations of the MR assumptions. For simplicity, in each causal diagram, $Z_4$ alone (and therefore any combination involving $Z_4$) violates the MR assumptions: (a) pleiotropy, (b) population stratification, (c) offspring genotype, (d) postnatal effects of exposure, (e) changing exposure-instrument association over pregnancy, (f) selection on fertility, (g) exposure dichotomization, (h) missing data. See Appendix for further details.

Finding the instrumental inequalities are not satisfied, however, does not tell us why they are not satisfied. In our data example, there are several structural reasons why the MR assumptions could be violated, some of which are depicted in Figure 5 and described in the Appendix (Bowden et al., 2015; Bowden et al., 2016; Lawlor et al., 2017; Swanson et al., 2018; VanderWeele et al., 2014; Verbanck et al., 2018). It is also possible that the falsification of the MR model indicated by our findings are specific to our dataset, which motivated our simulations. As previously discussed, as sample size decreases and the number of proposed instruments increases, the MR assumptions, and thus the instrumental inequalities, can be more readily violated by chance. In the simple scenario constructed in our simulations, the instrumental inequalities appear to be violated for combinations excluding the invalid proposed instrument only when the bias for the invalid instrument is very strong and the sample is relatively small, in which cases the instrumental inequalities also indicate that the set of four jointly proposed instruments violate the MR conditions. The frequency of this type of sample-specific violation appears to decline with sample size, and there was no evidence of finding violations for combinations with no structural bias in simulated samples of 100,000 participants. Overall, the results of our simple simulations suggest that, even in settings with small samples and strong instruments, where it is possible detected violations are sample-specific, the instrumental inequalities still provide strong evidence regarding the validity of MR analyses within a particular dataset. However, in such settings, it may be difficult, if not impossible, to determine the source of said violations if it is truly limited to a subset of the proposed instruments. It is unclear how this property of the inequalities will be affected when larger numbers of SNPs are proposed as instruments. Although the instrumental inequalities may be impacted by sample size, outside of the all-binary case, statistical inference procedures have not been fully developed (Ramsahai and Lauritzen, 2011; Wang et al., 2017). Critically, this means that it is not yet possible to differentiate random violations from structural ones. In addition to development of such statistical inference methods, further research is needed to more thoroughly evaluate the ways in which sample size might impact the ability of the instrumental inequalities to correctly detect structural violations of the MR conditions in a range of realistic settings.

In our data example, the fact that violations by SNPs jointly proposed as instruments were detected by some of the instrumental inequalities applied to allele scores, which have a smaller number of strata, as well as the relative weakness of the proposed instruments, suggests that not all the violations in our dataset are attributable to sample size. If the violations detected are not

sample specific, but rather indicative of structural biases related to the SNPs proposed as instruments, this might suggest these 4 SNPs should not be used as instruments for the effect of maternal vitamin D on offspring behavioral outcomes.

More broadly, our data example provides a concrete case in which the instrumental inequalities falsified a model proposing multiple variables jointly as instruments, underscoring previous calls for the use of the instrumental inequalities in all IV analyses (Glymour et al., 2012; Labrecque and Swanson, 2018; Swanson and Hernán, 2013). Like all observational research, MR requires strong, unverifiable assumptions. However, in the context of one-sample MR with multiple proposed instruments, the instrumental inequalities may allow us to eliminate clearly invalid analyses and focus efforts on more potentially informative studies.

**Acknowledgements**

# Appendix

## Characteristics of Eligible Mother-Child Pairs

| | Total Eligible Sample (n=3,188) | Vitamin D-PDP (n=1,971) | Vitamin D-mother-reported ADHD (n=1,970) | Vitamin D-teacher-reported ADHD (n=1,146) |
|---|---|---|---|---|
| | n (%) | n (%) | n (%) | n (%) |
| **Maternal Characteristics** | | | | |
| Serum 25OHD | | | | |
| <50 nmol/L | 711 (22.3) | 580 (29.4) | 582 (29.5) | 333 (29.1) |
| 50-75 nmol/L | 748 (23.5) | 604 (30.7) | 604 (30.6) | 365 (31.8) |
| >= 75 nmol/L | 971 (30.5) | 786 (39.9) | 785 (39.8) | 448 (39.1) |
| Education Completed | | | | |
| Primary or less | 79 (2.5) | 37 (1.9) | 37 (1.9) | 31 (2.7) |
| Secondary | 1167 (36.6) | 699 (35.5) | 697 (35.4) | 423 (36.9) |
| Higher | 1841 (57.7) | 1200 (60.9) | 1203 (61) | 675 (58.9) |
| Maternal Age at Intake - mean(sd) | 31.7 (4.4) | 31.6 (4.1) | 31.6 (4.1) | 31.7 (4.2) |
| Drinking during pregnancy | | | | |
| None | 849 (26.6) | 559 (28.4) | 558 (28.3) | 329 (28.7) |
| Until pregnancy was known | 434 (13.6) | 319 (16.2) | 319 (16.2) | 185 (16.1) |
| After pregnancy was known | 1321 (41.4) | 923 (46.9) | 924 (46.9) | 539 (47) |
| Offspring Characteristics | | | | |
| Mother-reported PDP symptoms | 57 (1.8) | 38 (1.9) | N/A | N/A |
| Mother-reported ADHD symptoms | 110 (3.5) | N/A | 69 (3.5) | N/A |
| Teacher-reported ADHD symptoms | 64 (2) | N/A | N/A | 40 (3.5) |
| Female | 1586 (49.7) | 992 (50.4) | 994 (50.4) | 563 (49.1) |

# Novel visualization methods for the instrumental inequalities

One disadvantage of other methods of representing the instrumental inequalities, like forest plots, heatmaps, and tables, is that the ordering in which SNP combinations appear is relatively arbitrary, and it can be difficult to identify consistent patterns, such as single SNP appearing in all sets which violate the instrumental inequalities. While traditional network graphs can somewhat improve this issue, when the number of included SNPs grows large, these graphs begin to resemble "hairballs" and become increasingly difficult to interpret (Longabaugh, 2012). To ease interpretation, we developed a new visualization method for the instrumental inequalities, roughly based on BioFabric (Longabaugh, 2012). In these visualizations, each horizontal line represents a SNP, and each vertical line connects a set of SNPs proposed as instruments (with the number of included SNPs increasing from left to right). Each node thus represents a particular set of SNPs. In real data, the color of each node represents the value of the instrumental inequalities for a particular set of SNPs proposed jointly as instruments, with white indicating values $\leq 1$, meaning the instrumental inequalities held, and darker colors indicating increasing maximum values of the instrumental inequalities.

In simulation studies, this same visualization can be used to visualize the number of simulations in which the instrumental inequalities failed to hold for a given set of simulated proposed instruments. In that setting, the color of the nodes would represent the number of simulations in which the instrumental inequalities were violated for each set of variables jointly proposed as instruments, with darker colors indicating increasing numbers of simulations in which the instrumental inequalities were violated, rather than the value of the instrumental inequalities for a particular set of SNPs jointly proposed as instruments. One benefit of these visualizations is that they provide a simpler and less dense means of representing the values of the instrumental inequalities for large numbers of SNPs than tables. For very large numbers of SNPs, future research in this area might consider reducing computational burden by eliminating calculations of the inequalities for sets of SNPs containing subsets that had already violated the instrumental inequalities and marking such sets with a unique color on the resulting visualization.

One notable advantage of this visualization technique is that it allows for easier identification of a consistent pattern of violations of the MR assumptions originating from a single SNP. As we can see in Figure 4.4D, when all violations are of sufficient magnitude, and originate from a single SNP ($Z_1$), we see a single

dark horizontal line (a SNP where the instrumental inequalities were violated for most or all sets of SNPs jointly proposed as instruments including that particular SNP), and inconsistent dark patterns across the other SNPs (showing violations only in sets of SNPs jointly proposed as instruments including the problem SNP). This contrasts with Figure 4.4C, where we only see violations of the instrumental inequalities when $Z_1$, $Z_2$, $Z_3$, and $Z_4$ are all jointly proposed as instruments. In Figure 4.4C, we do not have enough evidence to suggest that violations of the MR assumptions arise from a single SNP, only that the MR conditions cannot hold for all 4 variables jointly proposed as instruments in the sample.

## Details of the Simulation Parameters

We conducted simulations of a relationship between 4 binary proposed instruments ($Z_1$, $Z_2$, $Z_3$, and $Z_4$), a binary exposure $X$, and a binary outcome $Y$, where the relationship between $X$ and $Y$ was confounded by a continuous variable $U$, and the proposed instrument $Z_1$ was an invalid instrument with a direct effect ($\beta_2$) on the outcome $Y$. Each simulation was constructed such that $Z_{1i} \sim bernoulli(0.5)$, $Z_{2i} \sim bernoulli(0.5)$, $Z_{3i} \sim bernoulli(0.5)$, $Z_{4i} \sim bernoulli(0.5)$, $U_i \sim norm(0,1)$, $X_i \sim bernoulli(expit(0.6 + 0.1 * U_i + \beta_1 * Z_{1i} + \beta_1 * Z_{2i} + \beta_1 * Z_{3i} + \beta_1 * Z_{4i}))$, and $Y_i \sim bernoulli(expit(0.02 + 0.1 * U_i + \beta_2 * Z_{1i}))$. In order to examine the effects of changing sample size and varying magnitudes of violation of the MR assumptions on the instrumental inequalities, we varied simulations across 3 samples sizes (1,000 individuals, 10,000 individuals, 100,000 individuals), 4 possible instrument strengths ($\beta_1$= 0.01, 0.1, 0.5, and 1.0, corresponding roughly to risk differences of 0.003, 0.021, 0.071, 0.079), and 7 possible strengths of violations of the MR assumptions ($\beta_2 = 0.01, 0.1, 0.5, 1, 1.5, 2, 4$, resulting in violation strengths on the risk difference scale of 0.001, 0.025, 0.121, 0.189, 0.230, 0.315, 0.377, and 0.478). For each combination of sample size, instrument strength, and magnitude of direct path violation, we conducted 1,000 simulations.

**Results of simulations with varying proposed sample sizes, proposed instrument strength, and size of violation of the MR assumptions**



Figure 4.6: Results of the instrumental inequalities for 1,000 simulations of samples of 1,000 individuals with effect of each proposed instrument on exposure 0.003 (risk difference). Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments (Y axis) and across increasing size of effect of $Z_1$ on $Y$ (X axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.

Figure content (heatmap): Proposed Instrument(s) vs expit(b_zy)

Color Key — Value: 0, 0.2, 0.4, 0.6, 0.8

| Proposed Instrument(s) | 1E-3 | 0.1 | 0.5 | 1 | 1.5 | 2 | 4 |
|---|---|---|---|---|---|---|---|
| **Violates MR conditions** | | | | | | | |
| z1 | 0 | 0 | 0 | 0 | 0.003 | 0.064 | 0.759 |
| q("z1", "z2") | 0 | 0 | 0 | 0 | 0.069 | 0.364 | 0.955 |
| q("z1", "z3") | 0 | 0 | 0 | 0.002 | 0.08 | 0.36 | 0.948 |
| q("z1", "z4") | 0 | 0 | 0 | 0.003 | 0.074 | 0.389 | 0.959 |
| q("z1", "z2", "z3") | 0 | 0 | 0.007 | 0.095 | | 0.943 | 0.998 |
| q("z1", "z2", "z4") | 0 | 0 | 0.003 | 0.11 | | 0.85 | 0.998 |
| q("z1", "z3", "z4") | 0 | 0 | 0.002 | 0.096 | | 0.854 | 0.997 |
| q("z1", "z2", "z3", "z4") | 0.039 | 0.045 | 0.208 | 0.698 | 0.962 | 0.998 | 1 |
| AlleleScore | 0.002 | 0 | 0.002 | 0.057 | 0.175 | 0.351 | 0.721 |
| **Does not violate MR conditions** | | | | | | | |
| z2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| z3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| z4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| q("z2", "z3") | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| q("z2", "z4") | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| q("z3", "z4") | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| q("z2", "z3", "z4") | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

expit(b_zy)

Figure 4.7: Results of the instrumental inequalities for 1,000 simulations of samples of 1,000 individuals with effect of each proposed instrument on exposure 0.003 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments (Y axis) and across increasing size of effect of $Z_1$ on $Y$ (X axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.

Figure 4.8: Results of the instrumental inequalities for 1,000 simulations of samples of 1,000 individuals with effect of each proposed instrument on exposure 0.071 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments ($Y$ axis) and across increasing size of effect of $Z_1$ on $Y$ ($X$ axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.

Figure 4.9: Results of the instrumental inequalities for 1,000 simulations of samples of 1,000 individuals with effect of each proposed instrument on exposure 0.079 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments (Y axis) and across increasing size of effect of $Z_1$ on $Y$ (X axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.
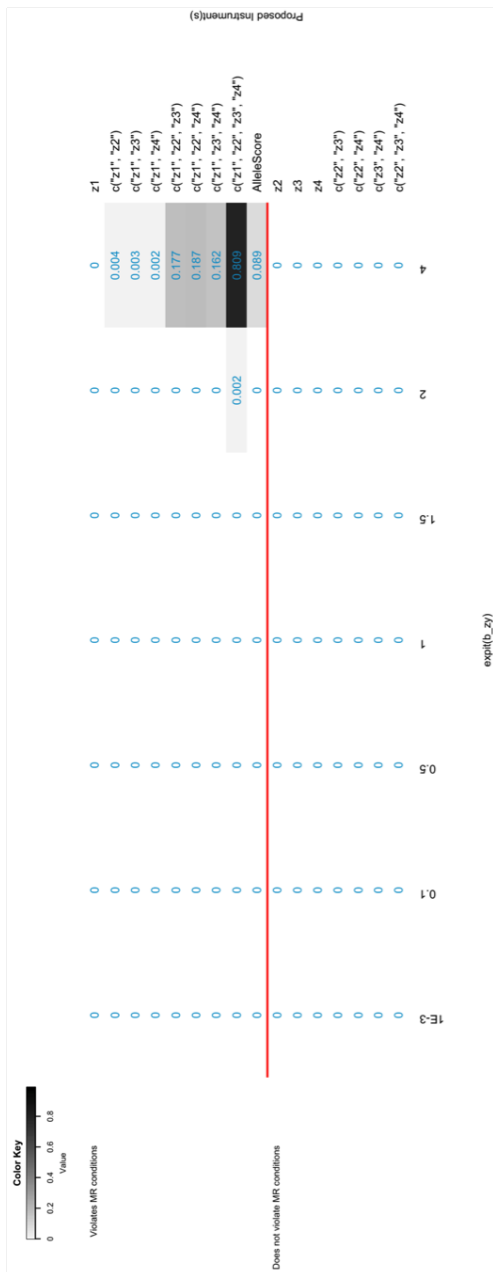
Figure 4.10: Results of the instrumental inequalities for 1,000 simulations of samples of 10,000 individuals with effect of each proposed instrument on exposure 0.003 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments (Y axis) and across increasing size of effect of $Z_1$ on $Y$ (X axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.
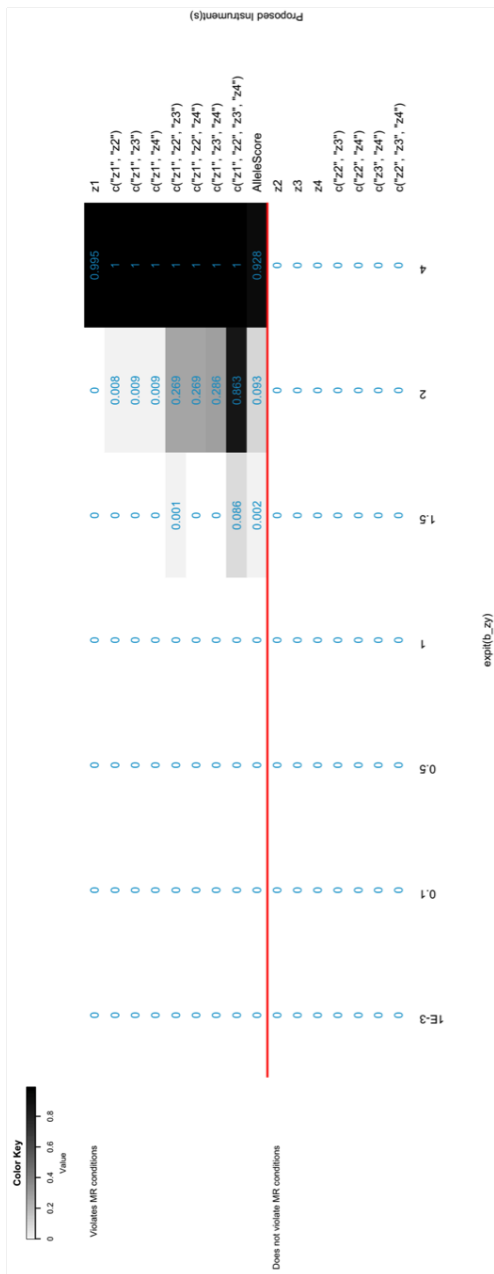
Figure 4.11: Results of the instrumental inequalities for 1,000 simulations of samples of 10,000 individuals with effect of each proposed instrument on exposure 0.021 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments ($Y$ axis) and across increasing size of effect of $Z_1$ on $Y$ ($X$ axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.
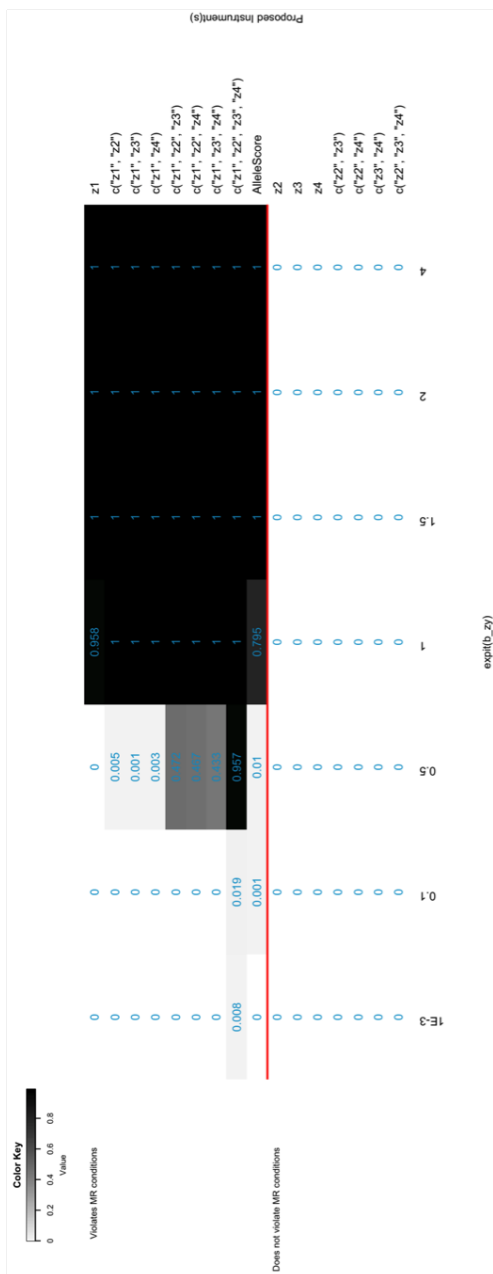
Figure 4.12: Results of the instrumental inequalities for 1,000 simulations of samples of 10,000 individuals with effect of each proposed instrument on exposure 0.071 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments (Y axis) and across increasing size of effect of $Z_1$ on $Y$ (X axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.
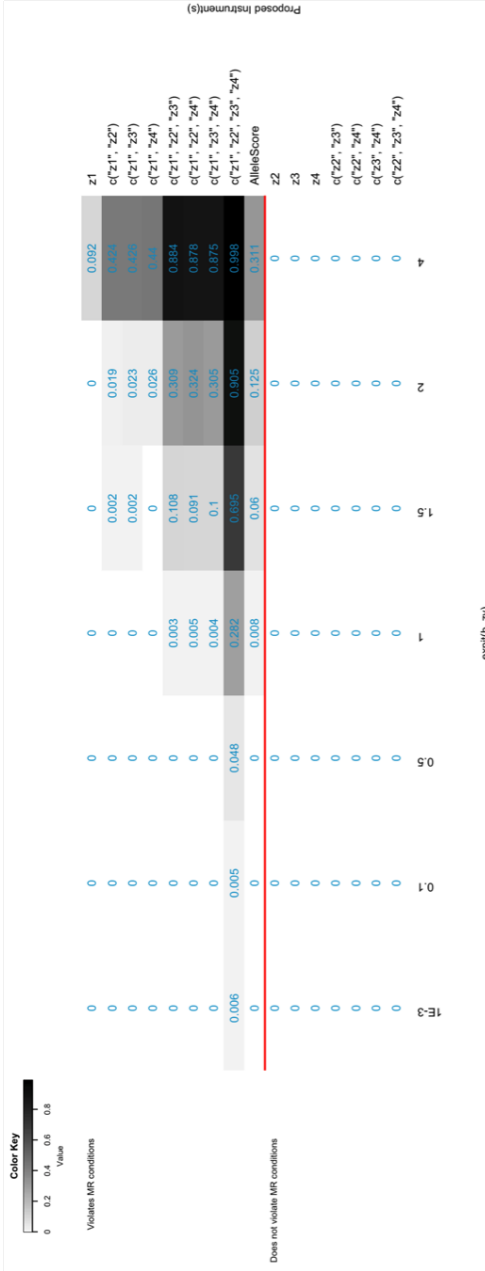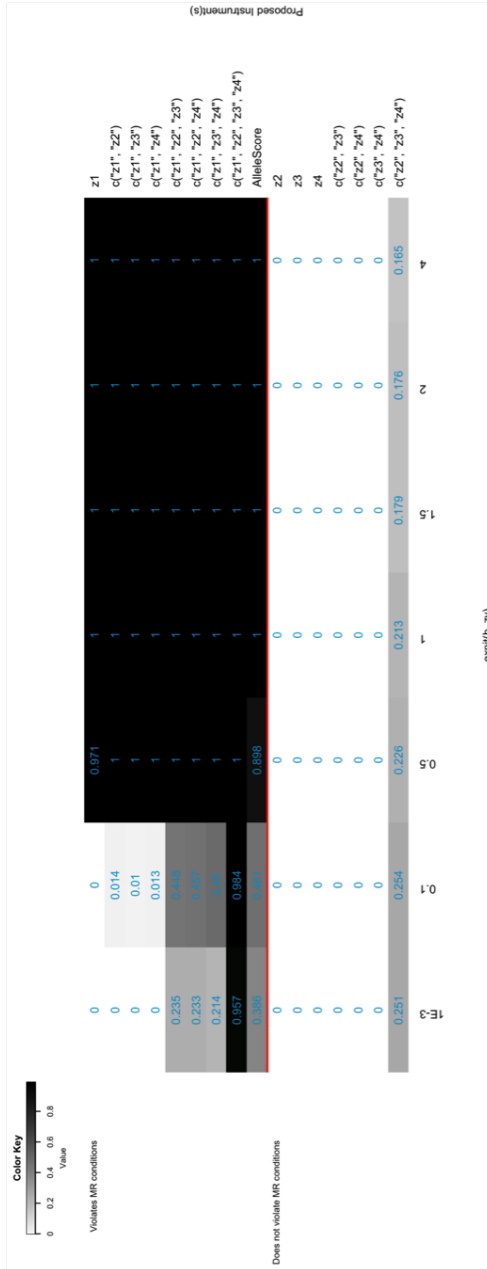
Figure 4.13: Results of the instrumental inequalities for 1,000 simulations of samples of 10,000 individuals with effect of each proposed instrument on exposure 0.079 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments ($Y$ axis) and across increasing size of effect of $Z_1$ on $Y$ ($X$ axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.
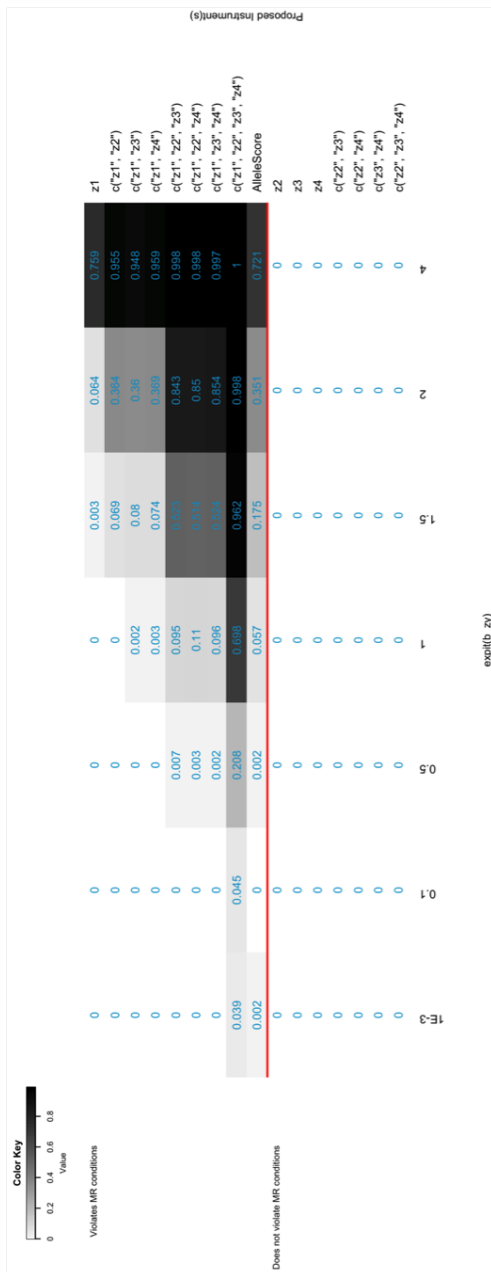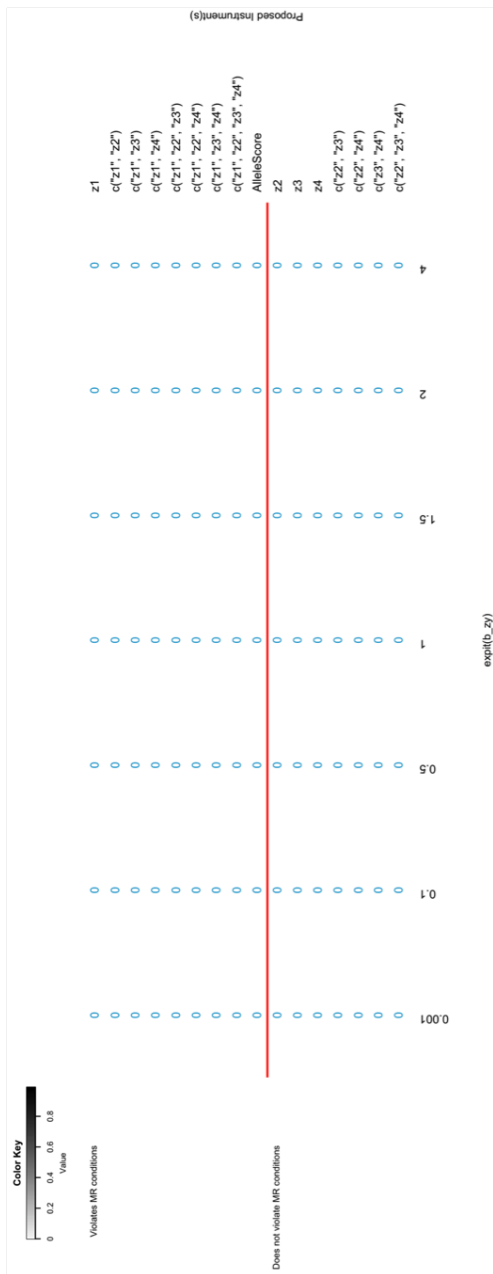
Figure 4.14: Results of the instrumental inequalities for 1,000 simulations of samples of 100,000 individuals with effect of each proposed instrument on exposure 0.003 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments ($Y$ axis) and across increasing size of effect of $Z_1$ on $Y$ ($X$ axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.
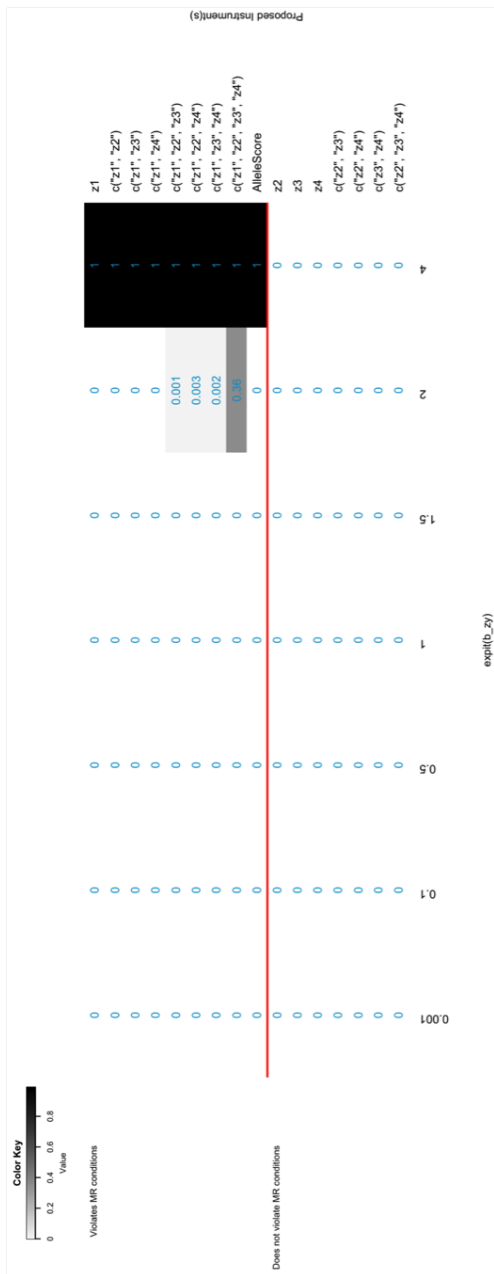
Figure 4.15: Results of the instrumental inequalities for 1,000 simulations of samples of 100,000 individuals with effect of each proposed instrument on exposure 0.021 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments (Y axis) and across increasing size of effect of $Z_1$ on $Y$ (X axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.
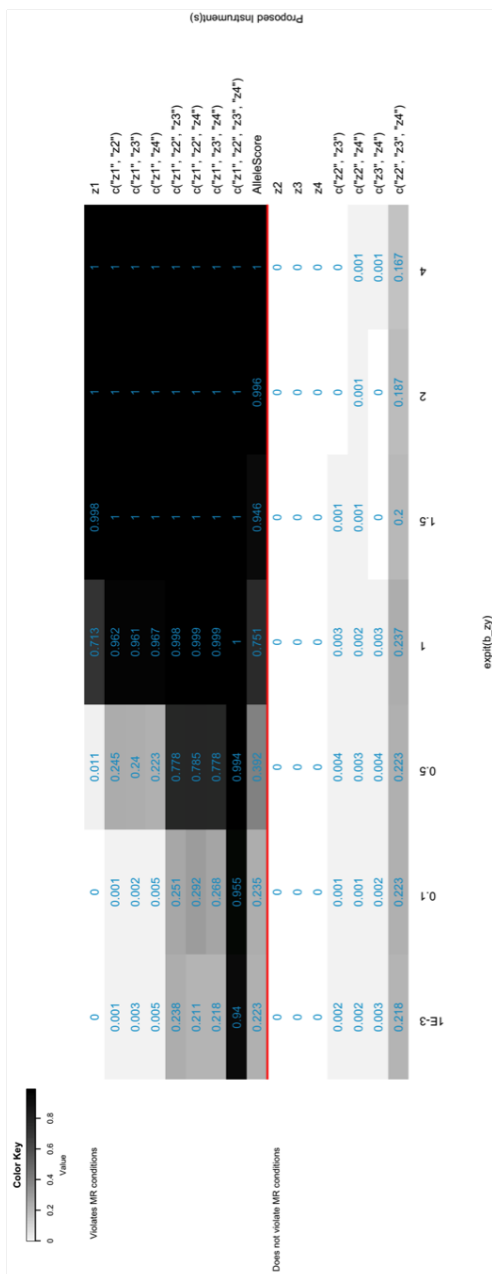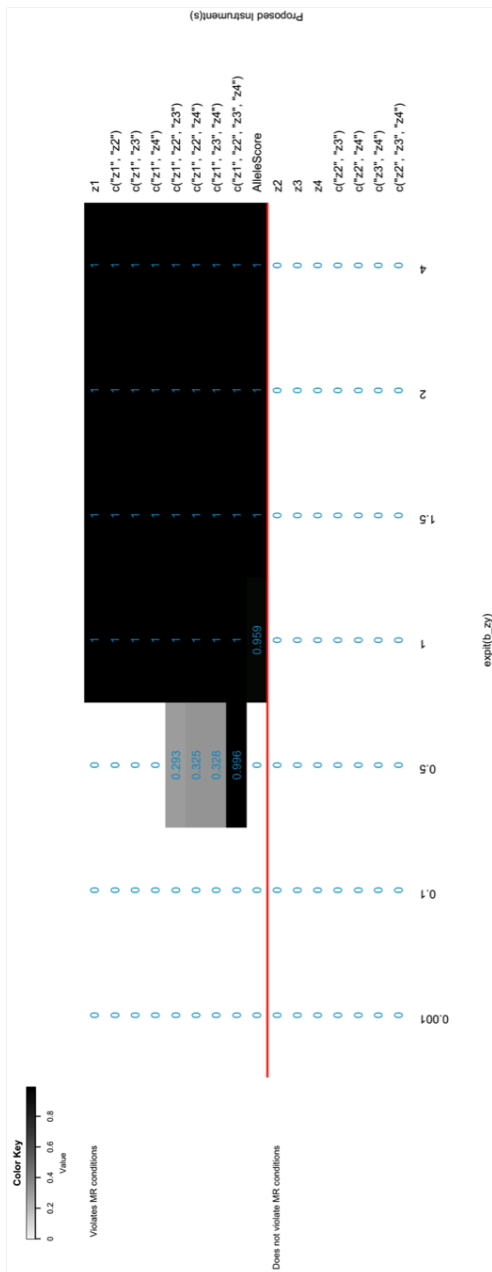
Figure 4.16: Results of the instrumental inequalities for 1,000 simulations of samples of 100,000 individuals with effect of each proposed instrument on exposure 0.071 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments ($Y$ axis) and across increasing size of effect of $Z_1$ on $Y$ (X axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.
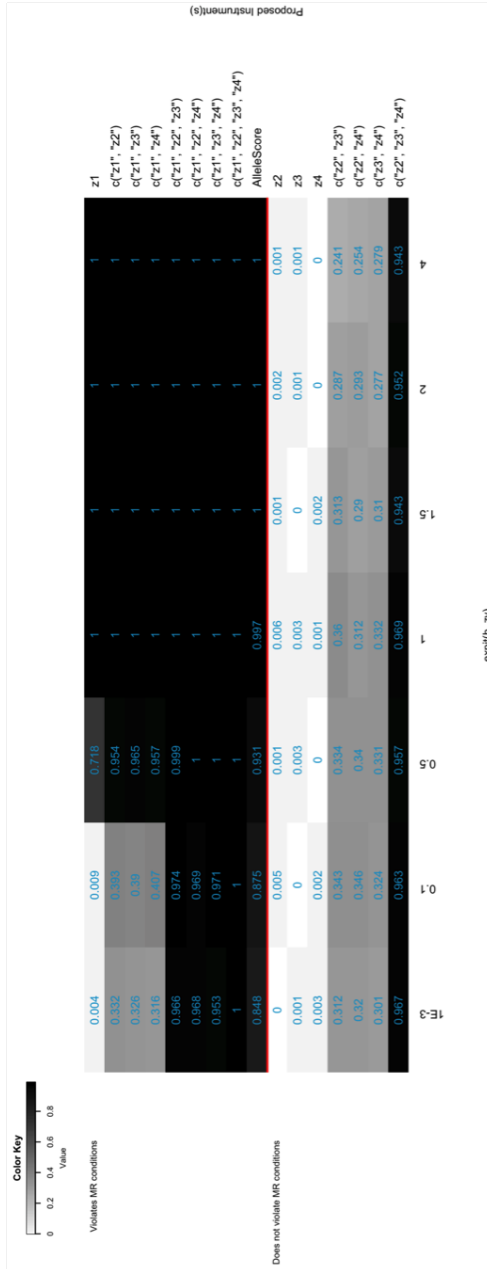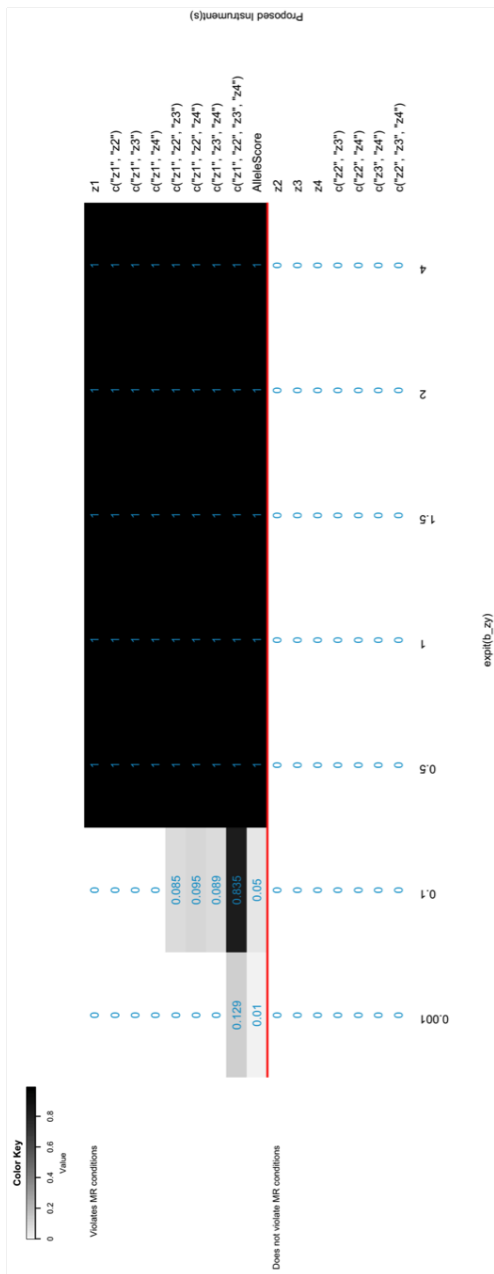
Figure 4.17: Results of the instrumental inequalities for 1,000 simulations of samples of 1,000 individuals with effect of each proposed instrument on exposure 0.003 (risk difference): Heatmap showing proportion of 1,000 simulations for which the instrumental inequalities failed to hold across possible combinations of the four proposed instruments (Y axis) and across increasing size of effect of $Z_1$ on $Y$ (X axis). Proposed instruments containing $Z_1$, which violate the MR conditions, are shown above the red line. Proposed instruments for which the MR conditions hold are shown below the red line.

## Possible sources of structural violations of the MR conditions within the data example

Pleiotropy, in which genetic loci affect multiple traits, violating the 2nd assumption, is one of the most commonly noted sources of potential bias in MR (Figure 4.5a) (Bowden et al., 2015; Bowden et al., 2016; Verbanck et al., 2018). Although we restricted our sample to mothers of European ancestry, it is possible that this strategy did not adequately control for population stratification, or that our sample contained substantial cryptic relatedness, both of which could result in assumption violations (Figure 4.5b). Previous research has also found that the required assumptions can be violated for MR analyses proposing maternal genetic factors as instruments for the effect of pregnancy exposures on offspring outcomes if the offspring's own genotype has a causal effect on the outcome, the mother's exposure status continues to affect the offspring after birth, or if the association between maternal genotype and vitamin D status changed over the course of pregnancy (Figures 4.5c, 4.5d, 4.5e) (de Groot et al., 1996; VanderWeele et al., 2014; Verhulst et al., 1985). In addition, if Vitamin D exposure impacted fertility or ability to carry a pregnancy to term, the MR assumptions could be violated by selection bias resulting from conditioning on live birth (Figure 4.5f). As previously mentioned, categorization of a truly continuous exposure, which is necessary for the use of the instrumental inequalities, can also violate the assumptions of an MR analysis (Figure 4.5g) (VanderWeele et al., 2014). If maternal genotype is related to missingness of exposure or outcome data, the MR assumptions could be violated by our use of complete case analysis (Figure 4.5h). These possible sources of bias are not mutually exclusive, and all could be present in our data at some level. In addition, further research is needed to evaluate the potential magnitude of bias resulting from these sources in plausible scenarios.

# References

Achenbach, T. M. (1991). *Integrative guide for the 1991 cbcl/4-18, ysr, and trf profiles.* Department of Psychiatry, University of Vermont.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the aseba preschool forms and profiles.* Burlington, VT: University of Vermont, Research center for children, youth, families.

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176.

Bonet, B. (2001). Instrumentality tests revisited. *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 48–55.

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International journal of epidemiology*, *44*(2), 512–525.

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

Burgess, S., & Thompson, S. G. (2013). Use of allele scores as instrumental variables for mendelian randomization. *International journal of epidemiology*, *42*(4), 1134–1144.

Canan, C., Lesko, C., & Lau, B. (2017). Instrumental variable analyses and selection bias. *Epidemiology (Cambridge, Mass.)*, *28*(3), 396.

de Groot, A., Koot, H. M., & Verhulst, F. C. (1996). Cross-cultural generalizability of the youth self-report and teacher's report form cross-informant syndromes. *Journal of Abnormal Child Psychology*, *24*(5), 651–664.

Glymour, M. M., Tchetgen Tchetgen, E. J., & Robins, J. M. (2012). Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American journal of epidemiology*, *175*(4), 332–339.

Greenland, S., & Mansournia, M. A. (2015). Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *European journal of epidemiology*, *30*(10), 1101–1110.

Hartwig, F. P., Davey Smith, G., & Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology*, *46*(6), 1985–1998.

Hernan, M. A., & Robins, J. M. (2018). *Causal inference*. Chapman; Hall/CRC.

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 360–372.

Holick, M. F. (2009). Vitamin d status: Measurement, interpretation, and clinical application. *Annals of epidemiology*, *19*(2), 73–78.

Holick, M. F., Binkley, N. C., Bischoff-Ferrari, H. A., Gordon, C. M., Hanley, D. A., Heaney, R. P., Murad, M. H., & Weaver, C. M. (2011). Evaluation, treatment, and prevention of vitamin d deficiency: An endocrine society clinical practice guideline. *The Journal of Clinical Endocrinology and Metabolism*, *96*(7), 1911–1930.

Hudziak, J. J., Copeland, W., Stanger, C., & Wadsworth, M. (2004). Screening for dsm-iv externalizing disorders with the child behavior checklist: A receiver-operating characteristic analysis. *Journal of child psychology and psychiatry*, *45*(7), 1299–1307.

Jaddoe, V. W. V., van Duijn, C. M., van der Heijden, A. J., Mackenbach, J. P., Moll, H. A., Steegers, E. A. P., Tiemeier, H., Uitterlinden, A. G., Verhulst, F. C., & Hofman, A. (2010). The generation r study: Design and cohort update 2010. *European journal of epidemiology*, *25*(11), 823–841.

Kang, H., Zhang, A., Cai, T. T., & Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, *111*(513), 132–144.

Kruithof, C. J., Kooijman, M. N., van Duijn, C. M., Franco, O. H., de Jongste, J. C., Klaver, C. C. W., Mackenbach, J. P., Moll, H. A., Raat, H., & Rings, E. H. H. M. (2014). The generation r study: Biobank update 2015. *European journal of epidemiology*, *29*(12), 911–927.

Labrecque, J., & Swanson, S. A. (2018). Understanding the assumptions underlying instrumental variable analyses: A brief review of falsification strategies and related tools. *Current Epidemiology Reports*, 1–7.

Lawlor, D., Richmond, R., Warrington, N., McMahon, G., Smith, G. D., Bowden, J., & Evans, D. M. (2017). 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*, *2*.

Longabaugh, W. J. R. (2012). Combing the hairball with biofabric: A new approach for visualization of large networks. *BMC bioinformatics*, *13*(1), 275.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, *80*(2), 319–323.

Mokry, L. E., Ross, S., Ahmad, O. S., Forgetta, V., Smith, G. D., Leong, A., Greenwood, C. M. T., Thanassoulis, G., & Richards, J. B. (2015). Vitamin d and risk of multiple sclerosis: A mendelian randomization study. *PLoS medicine*, *12*(8), e1001866.

Ong, J.-S., Cuellar-Partida, G., Lu, Y., Australian Ovarian Cancer, S., Fasching, P. A., Hein, A., Burghaus, S., Beckmann, M. W., Lambrechts, D., & Van Nieuwenhuysen, E. (2016). Association of vitamin d levels and risk of ovarian cancer: A mendelian randomization study. *International journal of epidemiology*, *45*(5), 1619–1630.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 435–443.

Pierce, B. L., Ahsan, H., & VanderWeele, T. J. (2011). Power and instrument strength requirements for mendelian randomization studies using multiple genetic variants. *International journal of epidemiology*, *40*(3), 740–752.

Ramsahai, R. R., & Lauritzen, S. L. (2011). Likelihood analysis of the binary instrumental variable model. *Biometrika*, *98*(4), 987–994.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

Sikora, D. M., Hall, T. A., Hartley, S. L., Gerrard-Morris, A. E., & Cagle, S. (2008). Does parent report of behavior differ across ados-g classifications: Analysis of scores from the cbcl and gars. *Journal of autism and developmental disorders*, *38*(3), 440–448.

Soma, Y., Nakamura, K., Oyama, M., Tsuchiya, Y., & Yamamoto, M. (2009). Prevalence of attention-deficit/hyperactivity disorder (adhd) symptoms in preschool children: Discrepancy between parent and teacher evaluations. *Environmental health and preventive medicine*, *14*(2), 150.

Swanson, S. A. (2019). A practical guide to selection bias in instrumental variable analyses. *Epidemiology*, *30*(3), 345–349.

Swanson, S. A., & Hernán, M. A. (2013). Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, *24*(3), 370–374.

Swanson, S. A., Labrecque, J., & Hernán, M. A. (2018). Causal null hypotheses of sustained treatment strategies: What can be tested with an instrumental variable? *European journal of epidemiology*, 1–6.

Tchetgen, E. J. T., Sun, B., & Walter, S. (2017). The genius approach to robust mendelian randomization inference. *arXiv preprint arXiv:1709.07779*.

Team, R. C. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Tick, N. T., Koot, H. M., & Verhulst, F. C. (2007). 14-year changes in emotional and behavioral problems of very young dutch children. *Journal of the American Academy of Child and Adolescent Psychiatry*, *46*(10), 1333–1340.

VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., & Kraft, P. (2014). Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.)*, *25*(3), 427.

Verbanck, M., Chen, C.-Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nature genetics*, *50*(5), 693.

Verhulst, F. C., Akkerhuis, G. W., & Althaus, M. (1985). Mental health in dutch children:(i) a cross-cultural comparison. *Acta psychiatrica scandinavica*, *72*(s323), 1–108.

Vieth, R. (2011). Why the minimum desirable serum 25-hydroxyvitamin d level should be 75 nmol/l (30 ng/ml). *Best Practice and Research Clinical Endocrinology and Metabolism*, *25*(4), 681–691.

Vimaleswaran, K. S., Berry, D. J., Lu, C., Tikkanen, E., Pilz, S., Hiraki, L. T., Cooper, J. D., Dastani, Z., Li, R., & Houston, D. K. (2013). Causal relationship between obesity and vitamin d status: Bi-directional mendelian randomization analysis of multiple cohorts. *PLoS medicine*, *10*(2), e1001383.

Vinkhuyzen, A. A. E., Eyles, D. W., Burne, T. H. J., Blanken, L. M. E., Kruithof, C. J., Verhulst, F., Jaddoe, V. W., Tiemeier, H., & McGrath, J. J. (2016). Gestational vitamin d deficiency and autism-related traits: The generation r study. *Molecular psychiatry*.

Wang, L., Robins, J. M., & Richardson, T. S. (2017). On falsification of the binary instrumental variable model. *Biometrika*, *104*(1), 229–236.

Wang, T. J., Zhang, F., Richards, J. B., Kestenbaum, B., Van Meurs, J. B., Berry, D., Kiel, D. P., Streeten, E. A., Ohlsson, C., & Koller, D. L. (2010). Common genetic determinants of vitamin d insufficiency: A genome-wide association study. *The Lancet*, *376*(9736), 180–188.

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., Robinson, M. R., McGrath, J. J., Visscher, P. M., & Wray, N. R. (2018). Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature communications*, *9*(1), 224.

# Chapter 5

# Falsification of the instrumental variable conditions in Mendelian randomization studies in the UK Biobank

Kelly Guo, Elizabeth W. Diemer, Jeremy A. Labrecque, Sonja A. Swanson

## 5.1 Abstract

**Background**: Mendelian randomization (MR) is an increasingly popular approach to estimating causal effects. Although the assumptions underlying MR cannot be verified, they imply certain constraints, the instrumental inequalities, which can be used to falsify the MR assumptions. However, the instrumental inequalities are rarely applied in MR.

**Methods**: Using 132 single nucleotide polymorphisms (SNPs), we applied the instrumental inequalities to MR models for the effects of vitamin D concentration, alcohol consumption, C-reactive protein (CRP), triglycerides, high-density lipoprotein (HDL) cholesterol, and low-density lipoprotein (LDL) cholesterol on coronary artery disease in the UK Biobank. For their relevant exposure, we applied the instrumental inequalities to MR models proposing each SNP as an instrument individually, and to MR models proposing unweighted allele score deciles as an instrument.

**Results**: We did not identify any violations of the MR assumptions when proposing each SNP as an instrument individually. When proposing allele scores as instruments, we detected violations of the MR assumptions for 2 of 6 exposures.

**Conclusions**: The instrumental inequalities are useful to identify violations of the MR assumptions when proposing multiple SNPs as instruments, but may be less useful in determining which SNPs are not instruments. This work demonstrates how incorporating the instrumental inequalities into MR analyses can help researchers to identify and mitigate potential bias.

## 5.2 Introduction

Mendelian randomization (MR) has become a popular approach for estimating causal effects and is increasingly popular as new large genetic databases become available (Johansson et al., 2010; Lor et al., 2019). Like any causal inference approach using observational data, MR requires assumptions. In brief, these methods require that genetic variants are (i) associated with the exposure of interest, (ii) only cause the outcome via the exposure, and (iii) share no causes with the outcome. We will refer to these conditions collectively as the instrumental conditions. Notably, these conditions are sufficient only for sharp null testing and bounding. Additional assumptions are necessary for point estimation (Hernán and Robins, 2006).

Though the instrumental conditions (ii) and (iii) are not verifiable, there are methods that can be used to falsify them. In particular, the instrumental conditions imply a set of mathematical constraints, known as the instrumental inequalities, which, if violated, show that that particular dataset's observed data distribution is inconsistent with the instrumental conditions (Balke and Pearl, 1997; Bonet, 2001; Diemer et al., 2020; Pearl, 1995; Richardson and Robins, 2010). Investigators can thus check these instrumental inequalities as one approach for falsifying the instrumental conditions. An application and description of how to do this in the setting of MR, especially in settings with multiple genetic variants being proposed as instruments, was presented in Chapter 4.

The question remains whether the instrumental inequalities are violated in other research settings, including in commonly studied exposures or commonly studied databases. Here, we aimed to apply the instrumental inequalities in the context of analysis of the effect of vitamin D concentration, alcohol consumption, C-reactive protein (CRP), triglycerides, high-density lipoprotein (HDL) cholesterol and low-density lipoprotein (LDL) cholesterol in the UK Biobank.

## 5.3 Methods

*Study Design*
The UK Biobank is a large prospective study of 502,648 adults, aged 40-69, recruited from across the United Kingdom between 2006 and 2010. Details of the cohort, including recruitment, assessment procedures, and quality control have been described in detail elsewhere (Bycroft et al., 2018; Sudlow et al.,

2015). The UK Biobank received ethical approval for the National Health Service National Research Ethics Service North West (Research Ethics Committee [REC] reference: 11/NW/0382), and all participants provided written informed consent. For each exposure under study, we restricted the eligible population to individuals with complete data on that exposure, the outcome of coronary artery disease, and that exposure's proposed genetic instruments, resulting in final sample sizes between 424,978 and 486,195 individuals (see Appendix for details). While this complete case analysis may result in selection bias (Swanson, 2019), it is consistent with common approaches in MR studies within UK Biobank.

*Exposures*
We selected 6 exposures whose relationship to cardiovascular disease were previously studied using MR: vitamin D concentration, alcohol consumption, C-reactive protein (CRP), triglycerides, high-density lipoprotein (HDL) cholesterol, and low-density lipoprotein (LDL) cholesterol (Beasley et al., 2019; Funck-Brentano et al., 2019; Havdahl et al., 2019).

Vitamin D, triglyceride, CRP, HDL-, and LDL- cholesterol levels were measured in blood samples collected at either the initial assessment visit or a repeat assessment visit conducted between 2012 and 2013. Details of biomarker measurements and assay performance in UK Biobank have been described in detail elsewhere (Fry et al., 2019). Briefly, vitamin D concentration was assessed based on total 25-hydroxyvitamin D (25(OH)D) levels measured using the Diasorin Liason, a chemoluminescent immunoassay. CRP levels were measured using an immunoturbidimetric assay in a Beckman Coulter analyzer (AU5800 Analyzer, Beckman Coulter, CA). Triglyceride concentrations were measured using an enzymatic analysis on said Beckman Coulter analyzer. HDL cholesterol levels were measured by enzyme inhibition analysis, and LDL cholesterol levels were measured using enzymatic protective selection analysis on said Beckman Coulter analyzer. Because the instrumental inequalities can only be used with categorical exposures, all these exposures were categorized into deciles. Frequency of alcohol consumption was assessed based on self-report questionnaire. Participants were asked "About how often do you drink alcohol?" with response options "never", "special occasions only", "1 to 3 times a month", "once or twice a week", "3 to 4 times a week", or "daily or almost daily". If participants felt the answer varied, they were instructed to give an average over the past year. This exposure was categorized using all available response categories.

*Outcome*

Participant electronic health records, including International Classification of Disease (ICD-10) diagnosis codes and Office of Population and Censuses Surveys (OPCS-4) procedure codes, have been integrated into UK Biobank. Additionally, patients were asked to report diagnoses of cardiovascular disease using questionnaires, which were subsequently checked during a verbal interview with a trained nurse. Participants were considered to have coronary artery disease (CAD) if they had experienced angina pectoris and acute or subsequent myocardial infarction or other acute or chronic ischemic heart disease (ICD-10 I20X, I21X, I123X, I24X, I25.5, I25.6, I25.8, I25.9) or previously underwent coronary procedure (OPCS-4 K40, K41, K43, K46, K49, K75, K45, K50.1-3) (see Appendix for details).

*Genetic Variants*

In order to identify genetic variants that had previously been used in MR studies with the UK Biobank, we conducted a systematic review of PubMed and the UK Biobank archive using the search terms "Mendelian randomization", "Mendelian random*", and each of the six exposures. Studies were eligible for inclusion in the review if they (1) explicitly reported using an MR approach, (2) studied either vitamin D, alcohol use, triglyceride levels, CRP, LDL-, or HDL-cholesterol as an exposure, and (3) conducted the analysis within a UK Biobank sample. This resulted in 30 articles, of which 13 were rejected based on full text review. After review, 8 articles on vitamin D concentration, 3 articles on alcohol use, 2 articles on CRP, and 4 articles on lipoproteins (LDL-cholesterol, HDL-cholesterol, or triglycerides) met the criteria and were included in the review (see Appendix for details).

We proposed single nucleotide polymorphisms (SNPs) as instruments for one of the six exposures if they had been proposed as instruments in at least 2 previous MR studies in UK Biobank. SNPs were not included if previous studies indicated that they were associated with another phenotype on possible pleiotropic pathway. For all exposures with more than one proposed instrument, we also constructed an unweighted categorical allele score. In total, we proposed 132 independent SNPs as instruments, including 11 SNPs as instruments for vitamin D concentration, 2 SNPs as instruments for alcohol consumption, 1 SNP as an instrument for CRP, 8 SNPs as instruments for triglyceride levels, 43 SNPs as instruments for HDL-cholesterol, and 67 SNPs as instruments for LDL-cholesterol.

*Statistical Analysis*

The properties and use of the instrumental inequalities have been described in detail elsewhere (Diemer et al., 2020; Richardson and Robins, 2010). Es-

sentially, for a specific MR model, the instrumental inequalities are a set of inequalities based on combinations of proportions of particular values of the proposed instruments, exposure, and outcome being less than one. If these inequalities do not hold (meaning the sum of proportions is greater than one), the instrumental conditions are incompatible with the observed data distribution. However, if the instrumental inequalities do hold, we do not have evidence for or against the instrumental conditions. Thus, the instrumental inequalities can be used to falsify (but not verify) an MR model. Importantly, the instrumental inequalities cannot show why the instrumental conditions are violated, only that they are. For MR studies, such violations can result from a number of different bias structures, including pleiotropy, selection bias, population stratification, and assortative mating (Diemer et al., 2020; VanderWeele et al., 2014). One key limitation of the instrumental inequalities is that they cannot be applied to continuous exposures (Bonet, 2001). While it is relatively easy to resolve this by discretising the continuous exposures of interest, the instrumental conditions can then be violated by the measurement error induced by this discretization, which is why we chose deciles over a smaller number of quantiles. When multiple SNPs are proposed as instruments in MR, we can also apply the instrumental inequalities to sets of SNPs jointly (see Chapter 4). For each exposure-outcome combination, we applied the instrumental inequalities to models proposing each SNP as an instrument individually, and (where multiple SNPs had been proposed as instruments), to a model proposing unweighted allele score deciles as an instrument, using R code developed in Chapter 4. As a sensitivity analysis to understand how results are impacted by residual population stratification, we also calculated the instrumental inequalities using inverse probability weights to adjust for 10 principal components (see Appendix for details). All analyses were conducted in R version 3.2.6 (Team, 2020).

## 5.4   Results

The study population in our final analytic subpopulations consisted of participants with a mean ages between 56 (range 38-78 years) and 57 years (range 38-73 years) who were primarily of white, British ethnicity. The proportion of women varied between 53.5%-54.2% across analytic subpopulations (Table 5.1).

The instrumental inequalities held for all 132 SNPs proposed as instruments when considering each SNP individually (Table 5.2). The instrumental in-

equalities also held when allele scores were proposed as instruments for alcohol consumption, HDL-, and LDL-cholesterol. However, the instrumental inequalities were violated when proposing allele scores as instruments for vitamin D and triglyceride levels, indicating the instrumental conditions were violated for those models. Results were generally consistent when inverse probability weighted for 10 principal components (see Appendix for details).

## 5.5   Discussion

In our investigation of 6 exposures and an accompanying 132 SNPs used in prior MR studies in UK Biobank, we detected no violations of the instrumental conditions when considering each SNP individually as a proposed instrument. Violations of the instrumental conditions were detected for some allele scores proposed as instruments.

These findings suggest that the instrumental inequalities may be helpful in detecting violations of the instrumental conditions for sets of SNPs proposed as instruments, but, per prior conventional wisdom, may not detect which specific SNP or SNPs are not instruments. Moreover, these findings do not explain why the allele scores are not instruments: we do not know if the conditions are violated due to a selection bias, pleiotropy, or another structural violation; or whether it is due to one or several SNPs marginally violating the instrumental conditions; or whether the violation is study-specific and not indicating a structural violation with that allele score being an instrument for that exposure-outcome in another study setting (Chapter 4). It is worth reiterating that detecting no violations when considering each SNP individually should only be interpreted as a failure to falsify, and not as support for the validity. It is possible that the instrumental conditions are violated for the SNP, exposure and outcome combinations in this study, but the instrumental inequalities did not detect these violations. Our results also need to be interpreted in light of the imposed categories of the exposures (see Chapter 4).

MR studies rely on strong, unverifiable assumptions, but investigators have an arsenal of tools for falsifying these assumptions and attempting to mitigate violations, along with robust methods that leverage alternative assumptions as a means to relax some others (Bowden et al., 2015; Bowden et al., 2016). The instrumental inequalities are an easily implementable technique, which, if integrated into this MR toolbox, could help to identify violations of the instrumental conditions in common MR settings with multiple proposed instruments,

including biases that may be difficult to identify through other means.

Table 5.1: Summary of Analytic Study Populations

| | Vitamin D Concentration | Alcohol Consumption | C-reactive Protein | Triglycerides | HDL-cholesterol | LDL-cholesterol |
|---|---|---|---|---|---|---|
| N | 443334 | 486195 | 463294 | 463934 | 424978 | 463435 |
| Sex: female | 237,302 (53.5%) | 263,656 (54.2%) | 251,169 (54.2%) | 251,467 (54.2%) | 228,607 (53.8%) | 228,607 (53.8%) |
| Recruitment age in years | | | | | | |
| Mean [SD] | 56 [8] | 57 [8] | 57 [8] | 57 [8] | 57 [8] | 57 [8] |
| Range | 38 - 78 | 38 - 73 | 38 - 73 | 38 - 73 | 38 - 73 | 38 - 73 |
| Self-reported ethnicity | | | | | | |
| British | 392,145 (88.5%) | 429,905 (88.4%) | 409,326 (88.4%) | 409,882 (88.3%) | 375,466 (88.3%) | 409,451 (88.4%) |
| Irish | 11,705 (2.6%) | 12,701 (2.6%) | 12,113 (2.6%) | 12,126 (2.6%) | 11,090 (2.6%) | 12,110 (2.6%) |
| Any other white background | 14,318 (3.2%) | 15,745 (3.2%) | 14,952 (3.2%) | 14,974 (3.2%) | 13,690 (3.2%) | 14,957 (3.2%) |
| Other | 25,166 (5.7%) | 27,844 (5.8%) | 26,903 (5.8%) | 26,952 (5.9%) | 24,372 (5.9%) | 26,917 (5.8%) |
| Coronary artery disease | 21,496 (4.8%) | 22,964 (4.7%) | 21,817 (4.7%) | 21,859 (4.7%) | 20,062 (4.7%) | 21,838 (4.7%) |
| Vitamin D, c-reactive protein, triglycerides, HDL-cholesterol, LDL-cholesterol | | | | | | |
| Mean [SD] | 48.64 nmol/L [21.1] | | 2.60 mg/L [4.4] | 1.75 mmol/L [1.03] . | 1.45 mmol/L [90.38] | 3.56 mmol/L [0.87] |
| Alcohol consumption | | | | | | |
| 'Never' | | 39,163 (8.1%) | | | | |
| 'Special occasions only' | | 55,958 (11.5%) | | | | |
| '1 to 3 times a month' | | 54,141 (11.1%) | | | | |
| 'Once or twice a week' | | 125,500 (25.8%) | | | | |
| '3 to 4 times a week' | | 112,390 (23.1%) | | | | |
| 'Daily or almost daily' | | 99,043 (20.4%) | | | | |

Table 5.2: Summary of results of the instrumental inequalities applied to MR models for the effect of vitamin D concentration, alcohol consumption, C-reactive protein, triglycerides, HDL-cholesterol, LDL-cholesterol concentrations on coronary artery disease.

| Exposure | Number of Proposed Instruments | Number of Violations of IV inequalities marginally | Maximum value of the inequalities for all SNPs marginally | Violation of the IV inequalities for allele score | Maximum value of instrumental inequalities for allele score |
|---|---|---|---|---|---|
| Vitamin D | 11 | 0/11 | 0.36 | Yes | 1.01 |
| Alcohol consumption | 2 | 0/2 | 0.27 | No | 0.26 |
| C-reactive protein | 1 | 0/1 | 0.16 | | |
| Triglycerides | 8 | 0/8 | 0.18 | Yes | 1.02 |
| HDL-cholesterol | 43 | 0/43 | 0.17 | No | 0.12 |
| LDL-cholesterol | 67 | 0/67 | 0.18 | No | 0.12 |

# Appendix

## Details of systematic review to identify commonly proposed genetic instruments

To identify SNPs commonly proposed as instruments for each exposure, we searched the UK Biobank archive (https://www.ukbiobank.ac.uk) and PubMed and the using the following search terms:

| | |
|---|---|
| **Vitamin D** | ("Mendelian Randomization Analysis"[Mesh] OR mendelian-random*[tiab]) AND (UK Biobank[tiab] OR UK-biobank[tiab]) AND ("Vitamin D"[Mesh] OR "25-Hydroxyvitamin D 2"[Mesh]) |
| **Alcohol** | ("Mendelian Randomization Analysis"[Mesh] OR mendelian-random*[tiab]) AND (UK Biobank[tiab] OR UK-biobank[tiab]) AND ("Drinking Behavior"[Mesh]) |
| **CRP** | ("Mendelian Randomization Analysis"[Mesh] OR mendelian-random*[tiab]) AND (UK Biobank[tiab] OR UK-biobank[tiab]) AND ("C-Reactive Protein"[Mesh]) |
| **Lipoprotein (LDL-cholesterol, HDL-cholesterol, triglycerides)** | ("Mendelian Randomization Analysis"[Mesh] OR mendelian-random*[tiab]) AND (UK Biobank[tiab] OR UK-biobank[tiab]) AND ('Triglycerides"[Mesh] OR "Cholesterol, HDL"[Mesh] OR "Cholesterol, LDL" [Mesh]) |

Databases were searched from their start date to March 2020. Initial searches resulted in 30 potentially eligible studies. Studies were eligible for inclusion in the review if they (1) explicitly reported using an MR approach, (2) studied either vitamin D, alcohol use, triglyceride levels, CRP, LDL-, or HDL-cholesterol as an exposure, and (3) conducted the analysis within a UK Biobank sample. Importantly, studies were not required to use coronary artery disease as an outcome. Of the 30 articles, 13 were rejected based on full text review. After review, 8 articles on vitamin D concentration, 3 articles on alcohol use, 2 articles on CRP, and 4 articles on lipoproteins (LDL-cholesterol, HDL-cholesterol,

or triglycerides) met criteria and were included in the review. A list of included articles and SNPs proposed in each is available below. A list of included articles and SNPs proposed in each will be available in the supplement to the published version of this article, and is omitted from this dissertation for brevity.

# ICD-10 and OPCS-4 coding and descriptions

Table 5.3: ICD-10 Codes

| International Classification of Disease, tenth revision (ICD-10) |
| --- |
| I20X Angina pectoris |
| I21X Acute myocardial infarction |
| I22X Subsequent myocardial infarction |
| I23X Current complications following acute myocardial infarction |
| I24X Other acute ischemic heart diseases |
| I25.1 Atherosclerotic heart disease |
| 25.2 Old myocardial infarction |
| I25.5 Ischaemic cardiomyopathy |
| I25.6 Silent myocardial ischemia |
| I25.8 Other forms of chronic ischaemic heart disease |
| I25.9 Chronic ischaemic heart disease, unspecified) |
| I23X Current complications following acute myocardial infarction |
| I24X Other acute ischemic heart diseases |
| I25X I251 Atherosclerotic heart disease |
| I25.2 Old myocardial infarction |
| I25.5 Ischaemic cardiomyopath |

## Table 5.4: ICD-10 Codes Continued

| International Classification of Disease, tenth revision (ICD-10) |
| --- |
| I25.6 Silent myocardial ischemia |
| I25.8 Other forms of chronic ischaemic heart disease |
| I25.9 Chronic ischaemic heart disease, unspecified) |
| K40 Saphenous vein graft replacement of coronary artery |
| K41 Other autograft replacement of coronary artery |
| K43 Prosthetic replacement of coronary artery |
| K46 Other bypass of coronary artery |
| K49 Transluminal balloon angioplasty of coronary artery |
| K75 Percutaneous transluminal balloon angioplasty and stenting of coronary artery |
| K45 Connection of thoracic artery to coronary artery |
| K50.1 Percutaneous transluminal laser coronary angioplasty |
| K50.2 Percutaneoucors transluminal coronary thrombolysis using streptokinase |
| K50.3 Transluminal atherectomy of coronary artery |

Table 5.5:  OPCS-4 Codes

| Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures, fourth revision (OPCS-4) |
| --- |
| K40 Saphenous vein graft replacement of coronary artery |
| K41 Other autograft replacement of coronary artery |
| K43 Prosthetic replacement of coronary artery |
| K46 Other bypass of coronary artery |
| K49 Transluminal balloon angioplasty of coronary artery |
| K75 Percutaneous transluminal balloon angioplasty and stenting of coronary artery |
| K45 Connection of thoracic artery to coronary artery |
| K50.1 Percutaneous transluminal laser coronary angioplasty |
| K50.2 Percutaneoucors transluminal coronary thrombolysis using streptokinase |
| K50.3 Transluminal atherectomy of coronary artery |

## Data Fields in UK Biobank

| Exposure | Data Field |
| --- | --- |
| Vitamin D | 30890 |
| Alcohol Intake Frequency | 1558 |
| C-reactive Protein | 30710 |
| Triglycerides | 30870 |
| HDL-cholesterol | 30760 |
| LDL-cholesterol | 30780 |

| Outcome | Data Field |
| --- | --- |
| Non-cancer illness code | 20002 |
| Operation Code | 20004 |
| Vascular/heart problems diagnosed by a doctor | 6150 |
| Diagnoses ICD10 | 41270 |
| Operative Procedures OPC4 | 41272 |

# Flowcharts of participant inclusion in analytic study populations
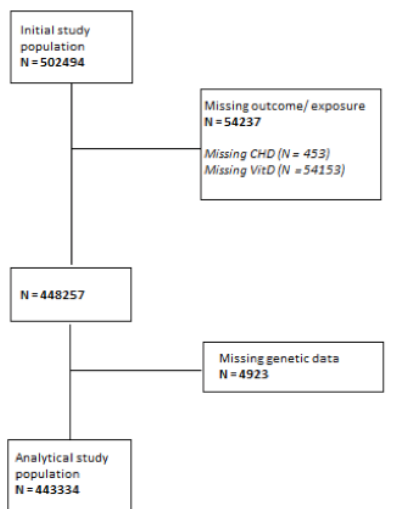


Figure 5.1: Vitamin D
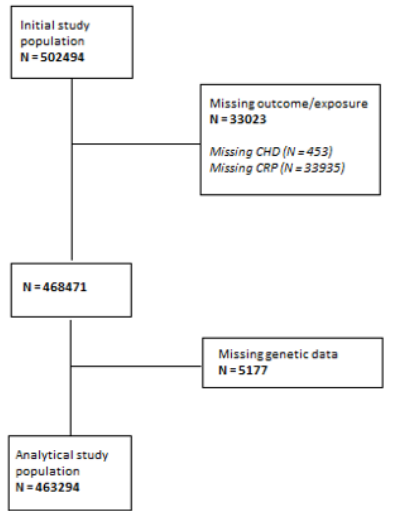
Figure 5.2: C-reactive Protein
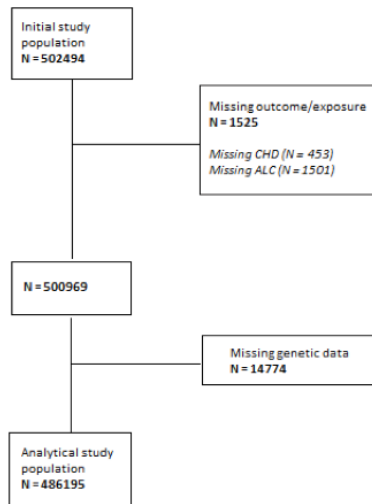


Figure 5.3: Alcohol Intake

194
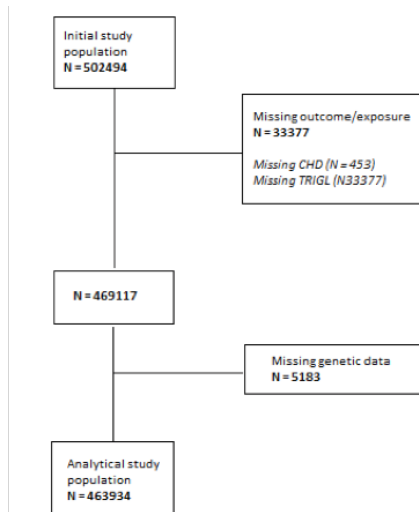
Figure 5.4: Triglycerides
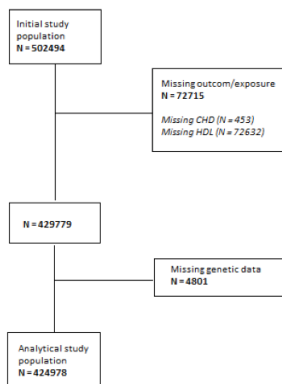


Figure 5.5: HDL-cholesterol
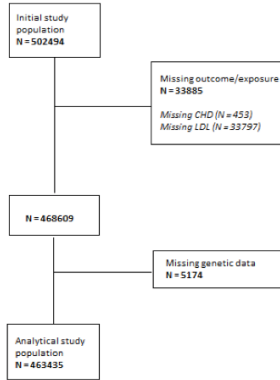
Figure 5.6: LDL-cholesterol

# Details of the inverse probability weighting procedure

For each proposed instrument $Z$, unstabilized inverse probability probability weights $(W^a)$ were estimated as follows,

$$W^A = 1/P(Z_l|PC_1, PC_2, PC_3, PC_4, PC_5, PC_6, PC_7, PC_8, PC_9, PC_{10})$$

To estimate $W^A$, we fitted multinomial logistic regression models predicting $Z_l$ assuming the principal components contributed additively and linearly on the logit scale. Values were subsequently back-transformed to probabilities, and we calculated
$1/P(Z_l|PC_1, PC_2, PC_3, PC_4, PC_5, PC_6, PC_7, PC_8, PC_9, PC_{10})$
for each individual using these back-transformed probabilities.

# Results of the instrumental inequalities in pseudo-populations inverse probability weighted for 10 principal components

Table 5.6: Summary of results of the instrumental inequalities applied to MR models for the effect of vitamin D concentration, alcohol consumption, C-reactive protein, triglycerides, HDL–cholesterol, LDL–cholesterol concentrations on coronary artery disease in pseudo-populations IP weighted for 10 principal components.

| Exposure | Number of Proposed Instruments | Number of Violations of IV inequalities marginally | Maximum value of the inequalities for all SNPs marginally | Violation of the IV inequalities for allele score | Maximum value of instrumental inequalities for allele score |
|---|---|---|---|---|---|
| Vitamin D | 11 | 0/11 | 0.36 | Yes | 1.01 |
| Alcohol consumption | 2 | 0/2 | 0.27 | No | 0.26 |
| C-reactive protein | 1 | 0/1 | 0.16 | | |
| Triglycerides | 8 | 0/8 | 0.18 | Yes | 1.02 |
| HDL–cholesterol | 43 | 0/43 | 0.17 | No | 0.12 |
| LDL–cholesterol | 67 | 0/67 | 0.18 | No | 0.12 |

# References

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176.

Beasley, M., Freidin, M. B., Basu, N., Williams, F. M., & Macfarlane, G. J. (2019). What is the effect of alcohol consumption on the risk of chronic widespread pain? a mendelian randomisation study using uk biobank. *Pain*, *160*(2), 501–507.

Bonet, B. (2001). Instrumentality tests revisited. *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 48–55.

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International journal of epidemiology*, *44*(2), 512–525.

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

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliot, L., Sharp, K., Mother, A., Vukcevic, D., Delanueau, O., O'Connell, J., & Cortes, A. (2018). The uk biobank resource with deep phenotypic and genomic data. *Nature*, *562*(7726), 203–209.

Diemer, E. W., Labrecque, J., Tiemeier, H., & Swanson, S. A. (2020). Application of the instrumental inequalities to a mendelian randomization study with multiple proposed instruments. *Epidemiology*, *31*(1), 65–74.

Fry, D., Almond, R., Moffat, S., Gordon, M., & Singh, P. (2019). Uk biobank biomarker project: Companion document to accompany serum biomarker data.

Funck-Brentano, T., Nethander, M., Movérare-Skrtic, S., Richette, P., & Ohlsson, C. (2019). Causal factors for knee, hip, and hand osteoarthritis: A mendelian randomization study in the uk biobank. *Arthritis & rheumatology*, *71*(10), 1634–1641.

Havdahl, A., Mitchell, R., Paternoster, L., & Smith, G. D. (2019). Investigating causality in the association between vitamin d status and self-reported tiredness. *Scientific reports*, *9*(1), 1–8.

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 360–372.

Johansson, Å., Marroni, F., Hayward, C., Franklin, C. S., Kirichenko, A. V., Jonasson, I., Hicks, A. A., Vitart, V., Isaacs, A., Axenovich, T., et al. (2010). Linkage and genome-wide association analysis of obesity-related phenotypes: Association of weight with the mgat1 gene. *Obesity*, *18*(4), 803–808.

Lor, G. C., Risch, H. A., Fung, W., Yeung, S. A., Wong, I. O., Zheng, W., & Pang, H. (2019). Reporting and guidelines for mendelian randomization analysis: A systematic review of oncological studies. *Cancer epidemiology*, *62*, 101577.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 435–443.

Richardson, T. S., & Robins, J. M. (2010). Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 415–444.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med*, *12*(3), e1001779.

Swanson, S. A. (2019). A practical guide to selection bias in instrumental variable analyses. *Epidemiology*, *30*(3), 345–349.

Team, R. C. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., & Kraft, P. (2014). Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.)*, *25*(3), 427.

Chapter 6

# Bounding the average causal effect in Mendelian randomization studies with multiple proposed instruments: An application to prenatal alcohol exposure and attention deficit hyperactivity disorder

Elizabeth W. Diemer, Alexandra Havdahl, Ole A. Andreassen, Marcus R. Munafo, Pal R. Njolstad, Henning Tiemeier, Luisa Zuccolo, Sonja A. Swanson

## 6.1 Abstract

Point estimation in Mendelian randomization (MR), a type of instrumental variable model, requires strong homogeneity assumptions beyond the core instrumental conditions. Bounding approaches, which do not require homogeneity assumptions, are infrequently applied in MR. Using data from the Norwegian Mother, Father, and Child Cohort Study and Avon Longitudinal Study of Parents and Children (n=4,457, 6,216) to study the average causal effect of prenatal alcohol exposure on offspring attention deficit hyperactivity disorder symptoms, we proposed 11 maternal SNPs as instruments. We then computed bounds assuming a subset of SNPs were jointly valid instruments, for all combinations of SNPs where the MR model was not falsified. The MR assumptions were violated for all sets with more than 4 SNPs in one cohort and for all sets with more than 2 SNPs in the other. Bounds assuming one SNP was an individually valid instrument barely improved on assumption-free bounds. Bounds tightened as more SNPs were assumed to be jointly valid instruments, and occasionally identified the direction of effect, though bounds from different sets varied. Our results suggest that, when proposing multiple SNPs as instruments, MR bounds can contextualize plausible magnitudes and directions of an effect. Computing bounds over multiple assumption sets underscores the need for evaluation of the unverifiable assumptions of each MR model.

## 6.2   Introduction

When estimating causal effects using methods based on confounder adjustment, studies are vulnerable to bias from unmeasured confounding. This is especially problematic for exposure-outcome relationships where confounders are complex or difficult to measure. Mendelian randomization (MR), an instrumental variable model proposing single nucleotide polymorphisms (SNPs) as instruments, is an increasingly popular alternative. Under certain conditions, MR allows for estimation of causal effects even in the presence of unmeasured confounding. Specifically, when proposing a single SNP as an instrument, MR requires that the SNP is associated with the exposure, does not affect the outcome except through the exposure, and individuals at different levels of the SNP are exchangeable with regards to counterfactual outcome (Hernan and Robins, 2018). To obtain a point estimate for the average causal effect in the population, investigators must additionally make one of a set of possible homogeneity assumptions, described in detail elsewhere (Balke and Pearl, 1997; Manski, 1990; Robins, 1989; Tchetgen et al., 2017). Unfortunately, these point estimating conditions are often biologically implausible in MR (Diemer et al., 2020; Hernán and Robins, 2006).

In contrast, bounding of the average causal effect can be conducted under the 3 primary MR conditions alone. Historically, bounding approaches have been unpopular, possibly because bounds based on a single binary proposed instrument are often wide (Swanson and Hernán, 2013). However, when multiple SNPs are proposed as instruments, there are underrecognized opportunities. First, we might tighten bounds by proposing joint sets of SNPs as instruments (Richardson and Robins, 2014; Swanson, 2017). Second, by comparing bounds computed under different assumptions, we might learn more about our reliance on assumptions in informing plausible effect sizes (Cole et al., 2019; Robins and Greenland, 1996; Swanson et al., 2018).

This approach may be especially helpful for MR studies of the effect of pregnancy alcohol consumption on offspring outcomes. While several non-MR studies have found positive associations between maternal pregnancy alcohol use and offspring attention deficit hyperactivity disorder (ADHD) (Han et al., 2015; Linnet et al., 2003; Pagnin et al., 2019), these estimated effects may be confounded by other maternal health behaviors. However, because offspring alcohol exposure depends both on the amount of alcohol consumed by the mother and the speed of the mother's alcohol metabolism, most versions of homogeneity assumptions required for point estimation using MR are violated when propos-

ing alcohol dehydrogenase-related SNPs as instruments: the effect of alcohol exposure would likely be heterogeneous across offspring of mothers with different genetic variants (Hernán and Robins, 2006; Swanson and Hernán, 2013). Additionally, because the effect of alcohol exposure is likely heterogeneous for other reasons, homogeneity assumptions are also suspect when proposing non-alcohol dehydrogenase SNPs as instruments (Hernán and Robins, 2006). Here, we demonstrate the use of bounds derived from multiple proposed instruments in an MR study where effect heterogeneity is expected, and provide adaptable software for the implementation of the bounds across combinations of proposed instruments.

## 6.3   Methods

**Data**
The Avon Longitudinal Study of Parents and Children (ALSPAC) is a longitudinal birth cohort, which aimed to recruit all pregnant women in former Avon county with a due date between April 1st, 1991 and December 31st, 1992 and continues to follow the offspring. 75.3% of contacted women agreed to participate, resulting in a total of 14,541 pregnancies enrolled during this period. When the oldest children were approximately 7 years old, the study recruited additional eligible children who had not previously participated. The study now includes data on the offspring of 15,454 pregnancies. Further detail is available elsewhere (Boyd et al., 2013; Fraser et al., 2013; Northstone et al., 2019). The study website contains details on available data through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. Ethical approval was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. We restricted analyses to singleton pairs of self-reported white European origin with complete data on maternal genotype, maternal pregnancy drinking behavior, and offspring outcomes, resulting in a total analytic sample of 4,457 mother-child pairs (Appendix Figure 6.5).

The Norwegian Mother, Father and Child Cohort Study (MoBa) is a population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health. Participants were recruited from all over Norway from 1999-2008. The women consented to participation in 41% of the pregnancies.

The cohort now includes 114,500 children, 95,200 mothers, and 75,200 fathers. Detailed information is available elsewhere (Magnus et al., 2016; Paltiel et al., 2014). The current study is based on version 12 of the quality-assured data files released for research in January 2019. The establishment of MoBa and initial data collection was based on a license from the Norwegian Data Protection Agency and approval from The Regional Committee for Medical and Health Research Ethics. MoBa is now based on regulations related to the Norwegian Health Registry Act. The current study was approved by The Norwegian Regional Committee for Medical and Health Research Ethics (2016/1702). For this study, we restricted our sample to singleton pairs with complete data on maternal genetics, maternal pregnancy drinking behavior, and offspring outcomes, resulting in a final analytic sample of 6,216 mother-child pairs (Appendix Figure 6.6).

**Measures**
*Genetic variants*
We selected SNPs based on a recent genome-wide association study of alcohol use in UK Biobank (Clarke et al., 2017). Of the 14 SNPs identified at genome-wide significance in that analysis, we excluded 3 SNPs previously found to be associated with traits that could violate MR assumptions via pleiotropy, or were within genes that were associated with such traits (Brazel et al., 2019; Chambers et al., 2008; Kanai et al., 2018; Linnér et al., 2019; Strawbridge et al., 2018; Zhong et al., 2019). The 11 independent SNPs we thus proposed as instruments were rs145452708, rs193099203, rs11940694, rs29001570, rs3114045, rs140280172, rs9841829, rs35081954, rs9991733, rs149127347, chr18:72124965. ALSPAC mothers were genotyped using the Illumina human660W-quad, and imputed to the 1000 Genome Project. MoBa mothers were genotyped using either Illumina HumanCoreExome or Illumina Global Screening Array, and genotypes were imputed to Haplotype Reference Consortium (HRC) version 1.1. Details of ALSPAC and MoBa genotyping procedures are available in the Appendix.

In contrast to GWA studies, measurement error of SNPs proposed as instruments will not bias average causal effect estimates of the exposure of interest on the outcome, as long as measurement error of the SNPs is at most differentially associated with the exposure, and not with the outcome (Hernán and Robins, 2006). For this reason, we did not exclude proposed instruments with minor allele frequencies under 5% or imputation quality below 0.8. However, assortative mating can violate MR assumptions (Hartwig et al., 2018). While Hardy Weinberg equilibrium tests for all SNPs proposed as instruments were conducted as part of the quality control pipeline in both cohorts, these

tests may be underpowered to detect small deviations (Salanti et al., 2005). However, such deviations could cause large biases in MR. We estimated the correlation between maternal and paternal genotype for each SNP proposed as an instrument in one cohort to identify SNPs which may be particularly vulnerable to this bias (Appendix).

Because there is incomplete overlap of loci between 1000Genomes and HRC, not all 11 SNPs were available in MoBa. Proxies for unavailable SNPs were selected using LDProxy, based on maximum $r^2$ (Machiela and Chanock, 2015). Within MoBa, rs145441283 was used as a proxy for rs193099203 and rs1154447 was used as a proxy for rs35081954. Because chr18:72124965 was unavailable in either cohort, rs201288331 was used as proxy in ALSPAC, and rs12955142 was used as a proxy in MoBa.

*Alcohol Use*
Alcohol use in the second and third trimester was assessed via postal questionnaire around gestational weeks 18 and 32 in ALSPAC, in which mothers reported their average volume and frequency of alcohol consumption in the last few weeks. In MoBa, mothers reported average volume and frequency of alcohol consumption between gestational weeks 13-24 and after week 25 via a postal questionnaire at week 30. Although drinking in pregnancy is not truly a binary process, and mild drinking likely incurs different effects than heavy drinking, the bounding approach used here (described below) requires a binary exposure. For that reason, mothers were categorized as ever drinkers if they reported drinking any amount of alcohol during the second or third trimester, and never drinkers if they did not report any drinking during either trimester. Because heavy and moderate drinking were included in the same category, this approach may be vulnerable to bias from poorly defined interventions. To evaluate whether this caused violations of the instrumental inequalities, we applied the instrumental inequalities when grouping alcohol consumption into 3 categories (never drinking, <2 drinks per week, $\geq$ 2 drinks per week), 4 categories (never drinking, < 1 drink per week, 1-2 drinks per week, > 2 drinks per week), and 7 categories (never drinking, <1 drink per week, 1-2 drinks per week, 3-4 drinks per week, 5-6 drinks per week, 7-13 drinks per week, > 13 drinks per week). In secondary analyses, we restricted the study population to compare never drinking and moderate drinking, defined as drinking less than or equal to 32 grams of alcohol per week (approximately 2 drinks per week). Restricting the analytic population in this way can generate selection bias (Swanson et al., 2015), which is why this is not the primary approach.

*ADHD*

Table 6.1: Prevalence of Maternal Alcohol Use and Offspring Attention Deficit-Hyperactivity (ADHD) Symptoms in the Avon Longitudinal Study of Parents and Children (ALSPAC) and the Norwegian Mother, Father, and Child Study (MoBa).

|  | ALSPAC | MoBa |
|---|---|---|
| n | 4457 | 6216 |
| Alcohol use during 2nd and 3rd trimester of pregnancy | | |
|   0 g/week | 66.9 (1522) | 90.6 (5606) |
|   = 32 g/week | 9.5 (216) | 9.0 (555) |
|   > 32 g/week | 23.6 (536) | 0.4 (25) |
| Offspring ADHD symptoms | 2.0 (90) | 2.6 (163) |

In ALSPAC, mother-reported ADHD symptoms at age 7 were assessed using the Development and Well-being Assessment (Goodman et al., 2000). In MoBa, mother-reported ADHD symptoms at age 5 were assessed using the Child Behavior Checklist attention deficit hyperactivity subscale (Achenbach and Rescorla, 2000). Children with subscale T scores at or above the 98th percentile within the full MoBa cohort (raw score 8, equivalent to the 84th percentile in published norm data) were considered to have ADHD symptoms (Achenbach and Rescorla, 2000). Table 6.1 shows the prevalence of maternal alcohol use and offspring ADHD symptoms.

**Statistical Analysis**

When multiple SNPs are believed to be individually valid instruments, several MR models using different subsets of SNPs, and thus slightly different assumptions, are possible. We could conduct MR models separately for each SNP proposed as an instrument. If we were willing to assume several SNPs were individually and jointly valid instruments, we could also conduct MR analyses proposing the set of SNPs as joint instruments. Our analysis plan included computing bounds under combinations of assumptions related to the 10 SNPs being proposed as instruments, as described below. Prior to computing any of these bounds, we applied the instrumental inequalities to attempt to falsify each assumption set (Bonet, 2001; Pearl, 1995). Specifically, in each cohort, we applied the Balke-Pearl instrumental inequalities across all possible combinations of the SNPs proposed as instruments and to a categorical, unweighted allele score, using code developed previously (Diemer et al., 2020). We also

applied the Bonet instrumental inequalities to each SNP individually. All sets that violated the instrumental inequalities (e.g., resulted in values greater than 1 for the Balke-Pearl inequalities, or greater than 2 for the Bonet inequalities) were eliminated from further analysis. When multiple SNPs are proposed as joint instruments, the MR model can also be falsified if the bounds calculated using the sets flip, meaning the lower bound is higher than the upper bound. Sets that produced flipped bounds were also removed from the results.

As increasingly large numbers of SNPs are proposed as joint instruments, it is increasingly likely that the MR conditions, and thus the instrumental inequalities, will be violated by chance, rather than by a structural bias in the super-population of interest. These random violations are similar to the concept of "random confounding" in randomized control trials (Greenland and Mansournia, 2015). Such violations are of particular concern to the analyses in ALSPAC, due to the relatively small sample size and the number of combinations of SNPs proposed as instruments. However, as with random confounding in randomized control trials, if random violations of the MR conditions are present within a sample, an MR analysis in that sample is expected to produce biased effect estimates (Diemer et al., 2020). By eliminating all sets that violated the instrumental inequalities, we could eliminate all sets for which the MR conditions were clearly falsified. However, because it is unclear which violations of the inequalities represent structural violations of the MR conditions, as opposed to random violations, the extent to which results of the instrumental inequalities in this study can be generalized to other datasets is unclear.

In the setting of a binary exposure and outcome, bounds on the average causal effect can be calculated using exposure and outcome data alone, without any assumptions (Manski, 1990; Robins, 1989). These assumption-free bounds will always have width 1 and always include the null, meaning they cannot identify the direction of effect. Under the MR assumptions, narrower bounds on the average causal effect are possible. When a set of SNPs are assumed to be jointly valid instruments, the set can be combined into a single variable, with levels representing every unique combination of alleles from the included SNPs. This combined variable can then be used to generate bounds using the expression described by Richardson and Robins, 2014. To evaluate differences in the bounds across different joint instruments in each cohort, we calculated Richardson-Robins bounds for all combinations of the 11 SNPs that did not violate the instrumental inequalities (Appendix).

If at least some number $k$ SNPs, but not all 11, were jointly valid instruments, then the average causal effect would lie within the union of the Richardson-

Robins bounds computed proposing combinations of $k$ SNPs as joint instruments (Swanson, 2017). To explore this, we computed bounds in each cohort assuming only a subset of the 11 SNPs were jointly valid instruments, for all subset sizes where at least some combinations did not violate the instrumental inequalities.

In the context of alcohol-dehydrogenase related SNPs and prenatal alcohol, the additional homogeneity assumption required for point estimation of the average causal effect in MR is likely invalid. However, in order to explore how conclusions from point estimation and bounding in MR differ, we computed point estimates using two stage least squares (Appendix).

Although MoBa and ALSPAC are relatively ethnically homogenous, residual population stratification may bias our results. We therefore also calculated the instrumental inequalities and bounds for each possible combination of the proposed instruments using inverse probability weighting to adjust for 10 principal components (Appendix) (Hernan and Robins, 2018).

All analyses were conducted in R version 3.6.1 (Team, 2020). Adaptable R code for application of the instrumental variable bounds, filtered by the instrumental inequalities, will be published in the supplement to the published version of this article, and are omitted from this dissertation for brevity.

## 6.4   Results

When comparing any alcohol consumption to no alcohol consumption, in ALSPAC, the instrumental inequalities held for all SNPs individually, 28 combinations of 2 SNPs, 16 combinations of 3 SNPs, two combinations of 4 SNPs, and no combinations of 5 or more SNPs (Appendix Figure 6.7). In MoBa, the instrumental inequalities held for 9 combinations of 2 SNPs, and did not hold for any combination of 3 or more SNPs (Appendix Figure 6.8). In addition, the instrumental inequalities failed to hold for 3 SNPs individually in MoBa. A similar amount and pattern of instrumental inequality violations were observed when comparing moderate alcohol consumption to no alcohol consumption (Appendix Figures 6.9-6.10). Results of the instrumental inequalities were also broadly similar when categorizing alcohol consumption into 3,4, or 7 categories (Appendix Figures 6.11-6.16), and when samples were IP weighted for 10 principal components (Appendix Figures 6.17-6.20).

In ALSPAC, bounds assuming at least one SNP was an individually valid in-
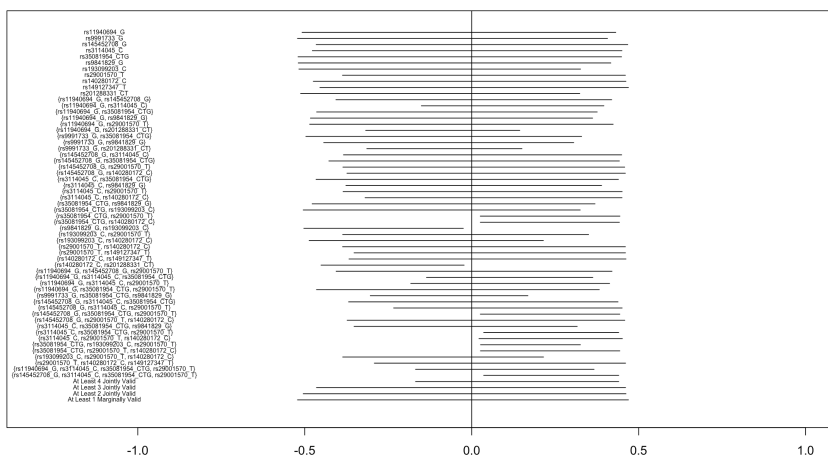
Figure 6.1: Bounds on the average causal effect of any vs no alcohol consumption during the second and third trimester on offspring attention deficit hyperactivity disorder symptoms in the Avon Longitudinal Study of Parents and Children, proposing different sets of SNPs as instruments.

strument were very wide (-0.52, 0.47), and barely improved on the assumption-free bounds (-0.53,0.47). Bounds calculated using each instrument individually were similarly wide (Figure 6.1). As the number of SNPs assumed to be jointly valid instruments increased, the bounds narrowed substantially, and sometimes fell completely on one side of the null, identifying the direction of effect. However, bounds from different sets of proposed instruments varied substantially, even identifying opposite directions of effect. With few exceptions, point estimates generally fell within the bounds (Appendix Table 6.2).

In MoBa, bounds were consistent across different assumptions (Figure 6.2). In all cases, the bounds covered the null. In most cases, the bounds did not differ substantially from the assumption-free bounds (-0.12, 0.88), with the narrowest bounds computed being based on the assumption that two SNPs (rs29001570 and rs9841829) were jointly valid (-0.07, 0.73). In 5 of 16 sets of proposed instruments, point estimates fell outside of the bounds (Appendix Table 6.3).

Bounds computed to estimate the effect of moderate alcohol consumption, rather than any alcohol consumption, followed a similar pattern in both cohorts (Figures 6.3-6.4, Appendix Tables 6.4-6.5). In ALSPAC, bounds proposing combinations of 3 or more SNPs narrowed more substantially than the
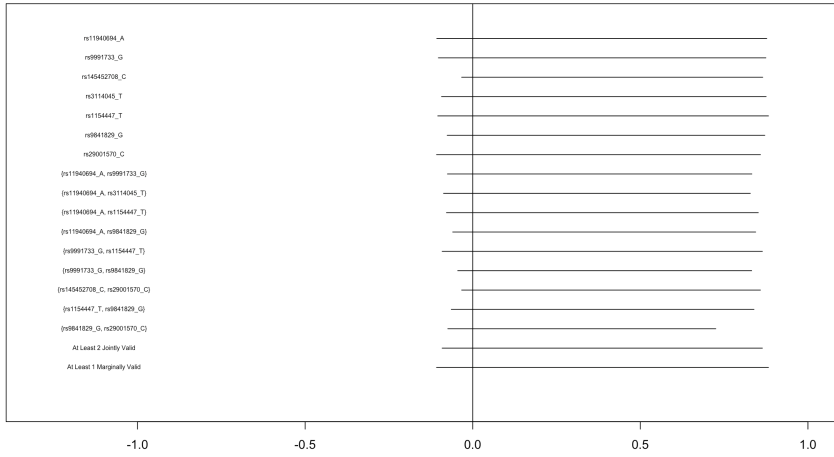
Figure 6.2: Bounds on the average causal effect of any alcohol vs. no alcohol consumption during the second and third trimesters of pregnancy on offspring attention deficit hyperactivity disorder symptoms in the Norwegian Mother, Father, and Child Study, proposing varying combinations of SNPs as instruments.

any alcohol models, though bounds still varied substantially and several sets resulted in flipped bounds. Results in each cohort were generally consistent when IP weighted for 10 principal components (Appendix Figures 6.21-6.24, Appendix Tables 6.6-6.9). Correlation between maternal and paternal genotypes was generally very small (Appendix Table 6.10).
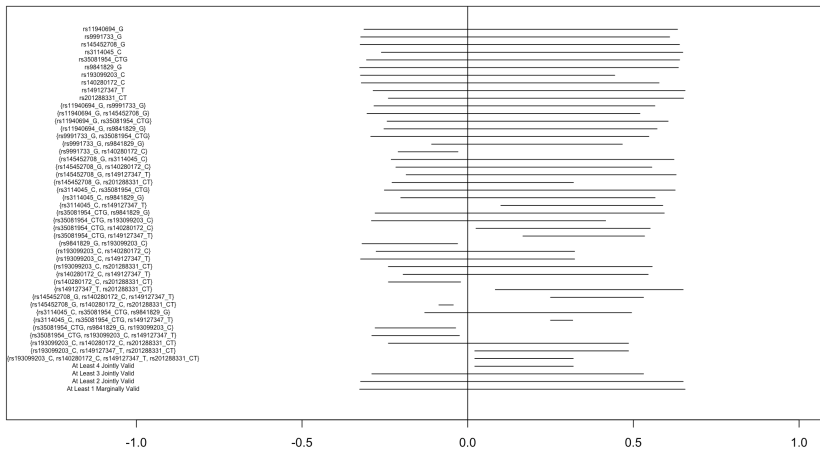
Figure 6.3: Bounds on the average causal effect of moderate (< 2 drinks/week) vs no alcohol consumption during the second and third trimester on offspring attention deficit hyperactivity disorder symptoms in the Avon Longitudinal Study of Parents and Children, proposing varying combinations of SNPs as instruments.

## 6.5 Discussion

When single SNPs were proposed as instruments, bounds on the average causal effect of both any and moderate prenatal alcohol consumption on offspring ADHD were wide, and were consistent with negative, null, and positive effects. However, in ALSPAC, as increasing number of SNPs were assumed to be joint instruments, bounds narrowed and sometimes identified the direction of effect, though bounds varied substantially across different proposed instruments. In MoBa, the instrumental inequalities held for far fewer sets of proposed instruments compared to ALSPAC. Bounds on the average causal effect of moderate and any alcohol consumption on offspring ADHD remained wide and fairly constant across several different sets of assumptions in MoBa.

Although bounds proposing a single SNP as an instrument barely improved on the assumption-free bounds, the width of the bounds did decrease as we incorporated stronger assumptions. Our ability to evaluate how incorporating stronger assumptions might narrow the bounds was limited by the fact that the strongest assumption sets we considered a priori (that all 11 SNPs were jointly valid instruments) were found to be violated. Nonetheless, bounds in
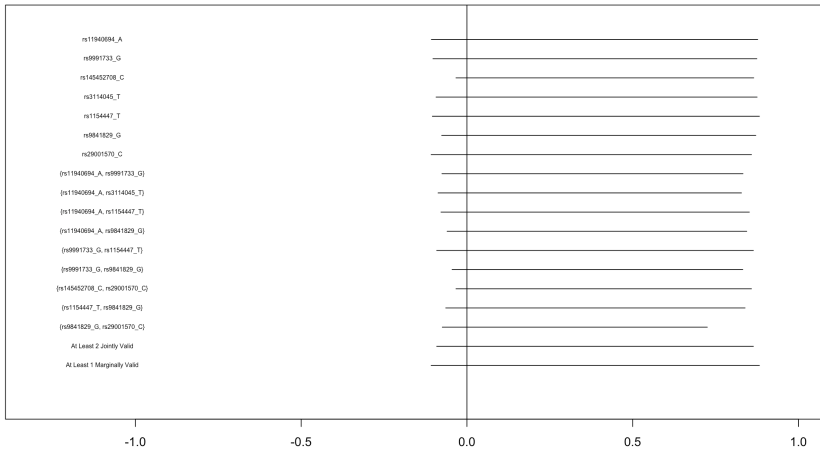
Figure 6.4: Bounds on the average causal effect of moderate alcohol consumption moderate (<2 drinks/week) vs no alcohol consumption during the second and third trimester on offspring attention deficit hyperactivity disorder symptoms in the Norwegian Mother, Father, and Child Study, proposing varying combinations of SNPs as instruments.

our analysis did narrow as larger numbers of SNPs were proposed as joint instruments, and sometimes identified the direction of effect. This suggests that, when multiple SNPs are proposed as jointly valid instruments, bounds may be able to inform decision-making without additional point estimating assumptions. This may be especially helpful for contexts, like MR studies of prenatal alcohol exposure, where homogeneity assumptions are implausible.

An advantage of computing bounds over many different assumptions is that such approaches can clarify how different assumptions can change study conclusions (Swanson et al., 2015). In our application, we were only able to identify a direction of effect under the strong assumption that multiple SNPs were jointly valid instruments. Moreover, in ALSPAC, proposing different sets of SNPs as joint instruments resulted in bounds that identified opposite directions of effects. This variation would have been difficult to identify in many MR point estimation approaches, but is clearly apparent when bounds are evaluated over several possible assumptions. In highlighting these variations, computing bounds over many different assumptions about the SNPs proposed as instruments could shift the focus of MR studies towards the question of what assumptions are most plausible, and thus which range of effects we should be

213

most confident in.

This property may be enhanced by combining bounding with applications of the instrumental inequalities, which could allow for the elimination of analyses based on clearly invalid assumptions (Diemer et al., 2020). Our results showed that at least 7 of the SNPs in our analysis could not be valid instruments in ALSPAC, and at least 9 of the 11 could not be valid instruments in MoBa. This is surprising, as the full set of proposed instruments contained 4 SNPs in alcohol dehydrogenase regions, whose relationship to alcohol consumption is relatively well understood. This detected bias could have resulted from several different causes (some of which are detailed in the Appendix, see Chapter 3 for further detail), but indicates that MR studies of prenatal alcohol exposure may be more vulnerable to bias than was previously understood, and should be viewed with caution. Further investigation is needed to clarify how maternal alcohol-related SNPs impact offspring behavioral health.

The variation in the bounds across assumption sets also illustrates how strongly point estimation in MR relies on the homogeneity assumptions. Even under the strongest unfalsified assumption sets, bounds often covered a moderately large range of effect sizes, meaning point estimation under those sets would still depend heavily on the homogeneity assumptions. Under weaker sets of assumptions, like proposing a single SNP as an instrument, the conclusions of MR studies using point estimation would be informed almost entirely by those additional homogeneity assumptions. This suggests that greater attention should be paid to evaluating the validity of point-estimating assumptions in MR. In our application, point estimates sometimes fell outside the bounds, indicating a violation of the point-identifying assumptions. These sets included SNPs inside and outside of alcohol dehydrogenase regions. While violations of homogeneity were expected in our context, this suggests the resulting bias was severe, and future MR studies might benefit from closer evaluation of the plausibility of the point estimating assumptions.

Even in settings where both the primary MR assumptions and the additional point estimating assumptions are plausible, presentation of the bounds alongside point estimates could help readers and investigators to understand how strongly MR studies depend on assumptions. This is true even, and perhaps especially, when the bounds are wide. Several studies have called for presentation of bounds in observational studies, particularly for instrumental variable models like MR (Cole et al., 2019; Robins and Greenland, 1996; Swanson et al., 2018; Swanson et al., 2015). Robins and Greenland noted that "wide bounds make clear the degree to which public health decisions are dependent

on merging the data with strong prior beliefs" (12) (Robins and Greenland, 1996). Incorporating bounds into MR practice would clarify the amount of information present in the data alone, and the need for critical evaluation of assumptions within each study's unique context.

Further research is needed to extend bounding approaches for instrumental variables and MR in several ways, including but not limited to extensions for: estimation procedures incorporating sampling variability (Swanson et al., 2018; Tamer, 2010); time-varying interventions (Labrecque and Swanson, 2019; Robins, 1994, 2014); conditional instrumental variables incorporating measured covariates (Hernán and Robins, 2006); non-binary exposures (Burgess and Labrecque, 2018; VanderWeele et al., 2014); and two-sample approaches (Lawlor, 2016). Though this list is not exhaustive, we believe it represents priorities for maximizing the usefulness and applicability of bounding in MR.

## 6.6 Conclusion

Our results show that, when multiple SNPs are proposed as instruments, it is possible to narrow bounds on the average causal effect. The extent of this narrowing will likely depend on the study question and population, but sometimes may allow for identification of the direction of effect. Further, the variation of the bounds observed across different proposed instruments provides a clear example of how bounding can be used to evaluate how heavily an MR analysis depends on assumptions regarding a particular SNP.

MR studies frequently propose large numbers of SNPs as joint instruments, and thus make equivalently large numbers of assumptions about the joint validity of those proposed instruments. Adding to the growing arsenal of sensitivity analyses, bounding may allow researchers to leverage these assumptions to make meaningful conclusions about effects without additional homogeneity assumptions. Even when homogeneity assumptions are biologically plausible, estimating bounds across different combinations of proposed instruments may allow investigators to better evaluate the dependence of their conclusions on those assumptions.

# Appendix

## Description of genotyping in ALSPAC

ALSPAC children were genotyped using the Illumina HumanHap550 quad chip genotyping platforms. The resulting raw genome-wide data were subjected to standard quality control methods. Individuals were excluded on the basis of gender mismatches; minimal or excessive heterozygosity; disproportionate levels of individual missingness ($> 3\%$) and insufficient sample replication (IBD < 0.8). Population stratification was assessed by multidimensional scaling analysis and compared with Hapmap II (release 22) European descent (CEU), Han Chinese, Japanese and Yoruba reference populations; all individuals with non-European ancestry were removed. SNPs with a call rate of $< 95\%$ or evidence for violations of Hardy-Weinberg equilibrium ($P < 5 * 10^{-7}$) were removed. Cryptic relatedness was measured as proportion of identity by descent (IBD > 0.1). Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation. 9,115 subjects and 500,527 SNPs passed these quality control filters.

ALSPAC mothers were genotyped using the Illumina human660W-quad array at Centre National de Génotypage (CNG) and genotypes were called with Illumina GenomeStudio. PLINK (v1.07) was used to carry out quality control measures on an initial set of 10,015 subjects and 557,124 directly genotyped SNPs. SNPs were removed if they displayed more than 5% missingness or a Hardy-Weinberg equilibrium P value of less than 1 x $10^{-06}$. Samples were excluded if they displayed more than 5% missingness, had indeterminate X chromosome heterozygosity or extreme autosomal heterozygosity. Samples showing evidence of population stratification were identified by multidimensional scaling of genome-wide identity by state pairwise distances using the four HapMap populations as a reference, and then excluded. Cryptic relatedness was assessed using a IBD estimate of more than 0.125 which is expected to correspond to roughly 12.5% alleles shared IBD or a relatedness at the first cousin level. Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation. 9,048 subjects and 526,688 SNPs passed these quality control filters.

After combining genotype data in the mothers and the children, SNPs with genotype missingness above 1% were removed due to poor quality (11,396 SNPs removed) and a further 321 subjects were removed due to potential ID mismatches. This resulted in a dataset of 17,842 subjects. Imputation of the

target data was performed using Impute V2.2.2 against the 1000 genomes reference panel (Phase 1, Version 3) (all polymorphic SNPs excluding singletons), using all 2186 reference haplotypes (including non-Europeans).

This gave 8,196 eligible mothers with available genotype data after exclusion of related subjects using cryptic relatedness measures described previously.

## Description of genotyping in MoBa

Genotyping of MoBa participants is currently ongoing, and this analysis was conducted using the first available maternal genetic data. Approximately 17,000 trios from MoBa were genotyped in 3 batches. Samples were selected randomly, and excluded from genotyping if the trio met any of the following exclusion criteria: 1) offspring stillborn, 2) offspring deceased, 3) twin offspring, 4) non-existent Medical Birth Registry data, 5) missing anthropometric measures at birth in Medical Birth Registry, 6) pregnancies where the mother did not answer the first questionnaire (as a proxy for higher fallout rate), 7) missing parental DNA samples. The first batch, comprising 20,664 individuals (including parents and children), was genotyped at the Genomics Core Facility (Iceland) using the Illumina HumanCoreExome (Illumina, San Diego, USA) genotyping array, version 12 1.1. The second batch, comprising 12,874 individuals, was genotyped at the Genomics Core Facility (Iceland) using the Illumina HumanCoreExome (Illumina, San Diego, USA) genotyping array, version 24 1.0. The third batch, comprising 17,949 individuals, was genotyped at Erasmus MC (the Netherlands) using the Illumina Global Screening Array (Illumina, San Diego, USA) version 24 1. Genotypes were called using GenomeStudio (Illumina, San Diego, USA) and converted to PLINK format files.

PLINK version 1.90 beta 3.36 (http://pngu.mgh.harvard.edu/purcell/plink/) was used to conduct the quality control, which has been previous described by Helgeland et al (Helgeland et al., 2019). Known problematic SNPs previously reported by the Cohorts for Heart and Aging Research in Genomic Epidemiology consortium and Psychiatric Genomics Consortium were excluded from each batch. Duplicate samples were removed, and each batch was split into parents and offspring. Quality control was conducted separately for parents and offspring.

Individuals were excluded if they had a genotyping call rate below 95% or autosomal zygosity greater than four standard deviations from the sample mean.

SNPs were excluded if they were ambiguous, had a genotyping call rate below 98%, or Hardy-Weinberg equilibrium p-value less than $1 \times 10^{-6}$. Population stratification was assessed using the HapMap phase 3 release 3 as a reference, by principal component analysis using EIGENSTRAT version 6.1.4. Visual inspection identified a homogenous population of European ethnicity, and individuals of non-european ethnicity were removed. A sex check was done by assessing the sex declared in the pedigree with the genetic sex, which was imputed based on the heterozygosity of chromosome X. When sex discrepancies were identified, the individual was flagged. Relatedness was assessed by flagging one individual from each pairwise comparison of identity-by-descent with a pi-hat greater than 0.1.

Parent and offspring datasets were then merged into one dataset per genotyping batch. All individuals passing the genotyping call rate and normal heterozygosity measures were included in the merged datasets, meaning individuals who had previously been flagged or excluded for being a duplicate, having a sex discrepancy, being an ethnic outlier, or having a high level of relatedness, were included. Concordance checks were then conducted on validated duplicates. Duplicate and tri-allelic SNPs, as well as SNPs that were discordant between validated duplicates were excluded. Indivudals with a genotyping call rate below 98% in the merged datasets were removed. Insertions and deletions were excluded.

Phasing was conducted using Shapeit 2 release 837 and the duoHMM approach was used to account for pedigree structure. Imputation was conducted using the Haplotype Reference Consortium release 1.1 as the reference panel. The Sanger Imputation Server was used to perform the imputations with the Positional Burrows-Wheeler Transform. Phasing and imputation were conducted separately for each genotyping batch.

Imputation quality control was performed by initially converting dosages to best-guess genotypes. Individuals were removed if they had a genotyping call rate less than 99% or were of non-European ethnicity. SNPs with a genotyping call rate less than 98%, or a Hardy-Weinberg equilibrium p-value less than $1 \times 10^{-6}$ were removed. Relatedness was assessed intergenerationally and across batches by flagging one individual from each pairwise comparison of identity-by-descent with a pi-hat greater than 0.15 (excepting known parent-offspring relationships). Individuals were flagged for removal only if the other member of the pair would otherwise be included in the same analysis. One individual from each pair was flagged at random, except when retaining one individual would keep more duo/trio data available, in which case the other member was

dropped. After quality control, a core homogenous sample of European ethnicity (based on PCA of markers overlapping with available HapMap markers) individuals across all batches and array were available for analysis, resulting in a total n of 14,804 mothers prior to analysis-specific exclusions.

## Expression for Richardson-Robins bounds for all possible combinations of instruments

Richardson and Robins, 2014 considered a model in which $X$ and $Y$ are binary, taking states $\{0, 1\}$, and $Z$ takes states $\{1, 2, .., k\}$ under 4 different assumptions:

(i) $Z \perp\!\!\!\perp Y^{x_1}, Y^{x_0}, X^{z_1}, ..., X^{z_k}$

(ii) $Z \perp\!\!\!\perp Y^{x_0}, Y^{x_1}$

(iii) for $i \in \{1, ..., k\}, j \in \{0, 1\}, Z \perp\!\!\!\perp X^{z_i}, Y^{x_j}$

(iv) there exists a $U$ such that $U \perp\!\!\!\perp Z$ and for $j \in \{0, 1\}, Y^{x_j} \perp\!\!\!\perp X, Z | U$

Each of these assumptions is a slightly different version of the IV conditions used in the literature (Swanson et al., 2018). Under assumption (i), (ii), (iii), or (iv), for all $i, j \in 0, 1$, $P(Y^{x_i} = j) \leq g(i, j)$ where

$g(i, j) = min\{min_z[P(X = i, Y = j | Z = z) + P(X = 1 - i, | Z = z)],$
$min_{z, \tilde{z}: z \neq \tilde{z}}[P(X = i, Y = j | Z = z) + P(X = 1 - i, Y = 0 | Z = z) + P(X = i, Y = j | Z = \tilde{z}) + P(X = 1 - i, Y = 1 | Z = \tilde{z})]\}$

Because $P(Y^{x_0})$ and $P(Y^{x_1})$ are variation independent, the average causal effect of X on Y, denoted $ACE(X \rightarrow Y)$, is bounded by

$1 - g(1, 0) - g(0, 1) \leq ACE(X \rightarrow Y) \leq g(0, 0) + g(1, 1) - 1$

Returning to our setting with multiple proposed instruments, we can consider the set of proposed instruments $B = \{b_1, b_2, .., b_n\}$. We note that any combination of the proposed instruments in $B$ that are themselves categorical variables can be combined into a single joint instrument $Z$ which takes states $\{1, 2, .., k\}$, where each state is a unique possible combination of values of the proposed joint instruments in the subset. Thus the Richardson-Robins bounds can be applied to any joint instrument $Z$, assuming (i), (ii), (iii), or (iv) hold both individually and jointly for each proposed instrument included in $Z$. In our application, we considered this for all possible subsets of our set of proposed instruments.

## Point estimation procedures

For each proposed joint instrument $Z_l$, point estimates for the average causal effect were estimated using two-stage least squares, using linear regression models for both steps. 95% confidence intervals were estimated using basic bootstrap. Two stage squares were estimated using the ivreg() function from the AER package (Kleiber et al., 2020), and bootstrapping was conducting using the boot.ci() function from the boot package (Ripley, 2010).

In the context of categorical exposures and outcomes, two stage least squares using linear regression is vulnerable to measurement error, which can result in predicted values of the exposure outside of the 0-1 range, and may violate the assumption of bivariate normally distributed errors (Rassen et al., 2009). However, some research has suggested this issue may be primarily theoretical, and has limited impact on practical applications (Angrist, 2001; Johnston et al., 2008; Rassen et al., 2009). Some MR researchers attempt to avoid this issue by using models based on logistic regression. However, these approaches will produce estimates of the causal odds ratio, rather than the average causal effect on the risk difference scale. In order to produce point estimates of the average causal effect on the same risk difference scale as the bounds, we therefore chose to use two stage least squares based on linear regression.

## Expression of inverse probability weights for each proposed joint instrument

For each proposed joint instrument $Z_l$ , unstabilized inverse probability weights (Robins, 1997) to account for 10 principal components were estimated as follows:

$$W^A = 1/P(Z_l|PC_1, PC_2, PC_3, PC_4, PC_5, PC_6, PC_7, PC_8, \ PC_9, PC_{10})$$

To estimate $W^A$, we fitted multinomial logistic regression models predicting $Z_l$ assuming the principal components contributed additively and linearly on the logit scale. Values were subsequently back-transformed to probabilities, and we calculated

$$1/P(Z_l|PC_1, PC_2, PC_3, PC_4, PC_5, PC_6, PC_7, PC_8, PC_9, PC_{10})$$

for each individual using these back-transformed probabilities.

## Possible violations of the MR assumptions in this analysis

The MR assumptions are strong, and any or all the SNPs proposed as instruments in our analysis may have been affected by a number of different biases. While the complete case approach used in this analysis aligns with common practice in MR, both analytic samples were much smaller than the original recruited cohort, meaning the results may have been affected by selection bias due to loss to followup and missing data. Further, because both cohorts were recruited based on the presence of a pregnancy, and offspring ADHD status can only be evaluated in women who become pregnant and carry to term, the MR conditions would have been violated if a woman's alcohol consumption impacted her probability of becoming pregnant (Diemer et al., 2020). Results were largely consistent when inverse probability weighted for 10 principal components, suggesting that the results were not affected by residual population stratification. The correlation between maternal and paternal genotype was small, implying the study was not strongly biased by assortative mating. While the MR assumptions can be violated if offspring outcomes are affected by offspring genotype, this path was unlikely to have impacted our analyses, because alcohol dehydrogenase genes are not expressed in fetuses or young children, who process alcohol through a different mechanism (van Faassen and Niemelä, 2011). The MR conditions could also be violated if maternal genetic variants impacted offspring ADHD through mechanisms other than alcohol consumption, such a consumption of other substances, if maternal alcohol consumption after birth also impacted offspring ADHD, or if the relationship between maternal genotype and alcohol consumption changed over the course of pregnancy.

## Supplementary figures 1-20
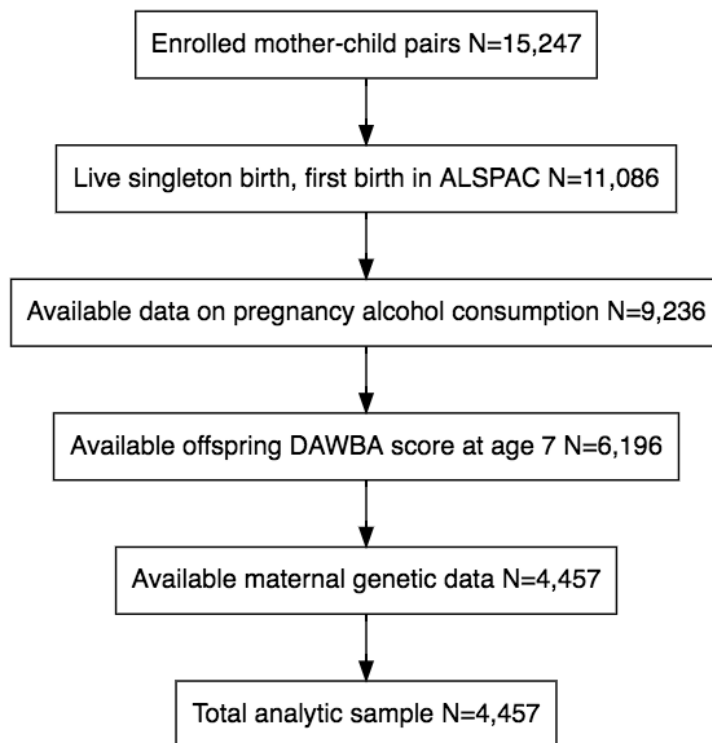


Figure 6.5: Flowchart of Avon Longitudinal Study of Parents and Children included in analytic sample
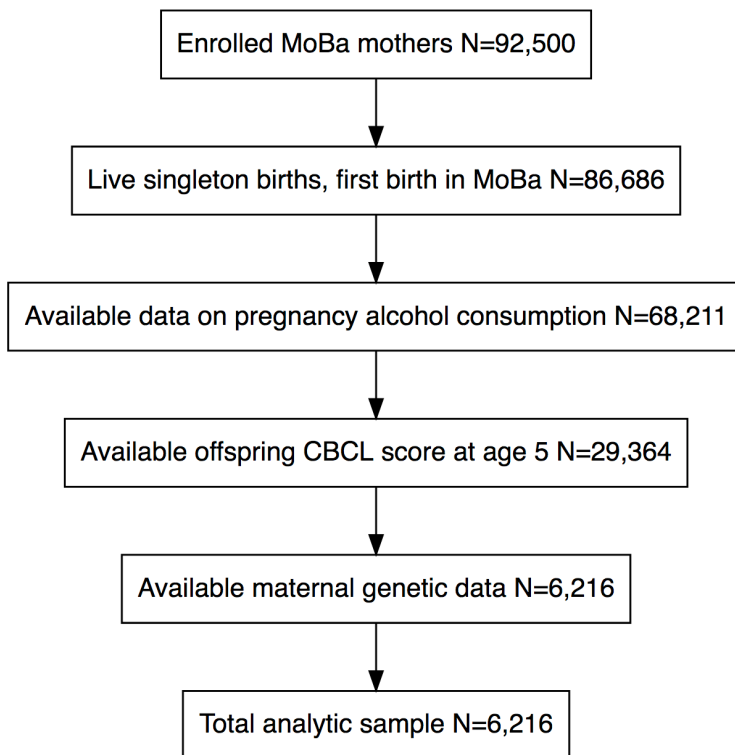
Figure 6.6: Flowchart of Norwegian Mother, Father, and Child Study included in analytic sample
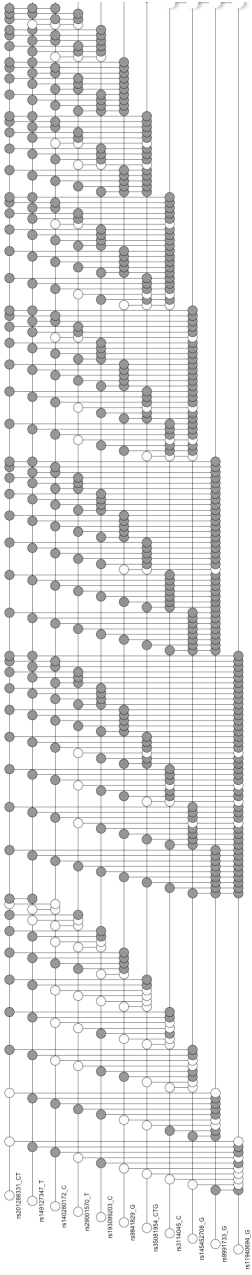
Figure 6.7: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of any vs. no alcohol consumption during pregnancy of offspring ADHD in the Avon Longitudinal Study of Parents and Children. This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.
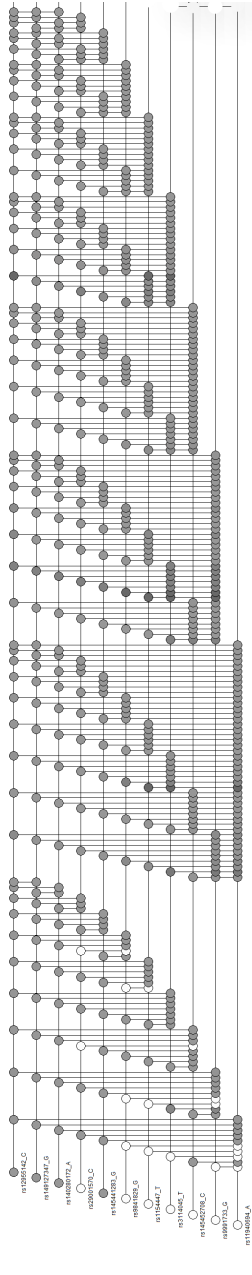


Figure 6.8: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of any vs. no alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother, Father, and Child Study. This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.
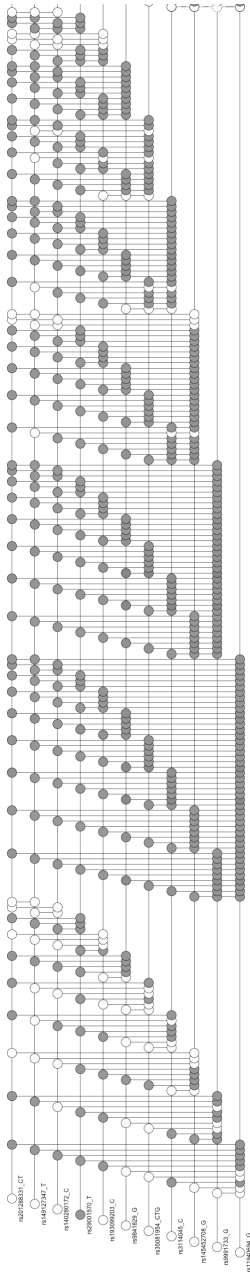
Figure 6.9: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of moderate vs. no alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children.This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.
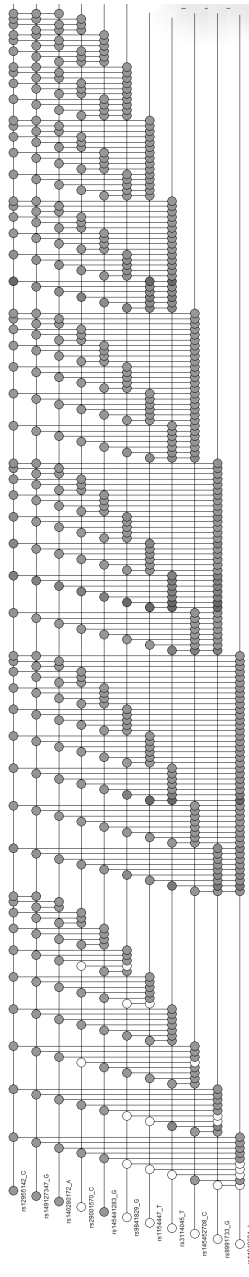


Figure 6.10: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of moderate vs. no alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother, Father, and Child Study. This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.

226

Figure 6.11: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children, grouping alcohol consumption into 3 categories. This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.



Figure 6.12: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children, grouping alcohol consumption into 4 categories.This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.

227

Figure 6.13: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children, grouping alcohol consumption into 7 categories.This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.
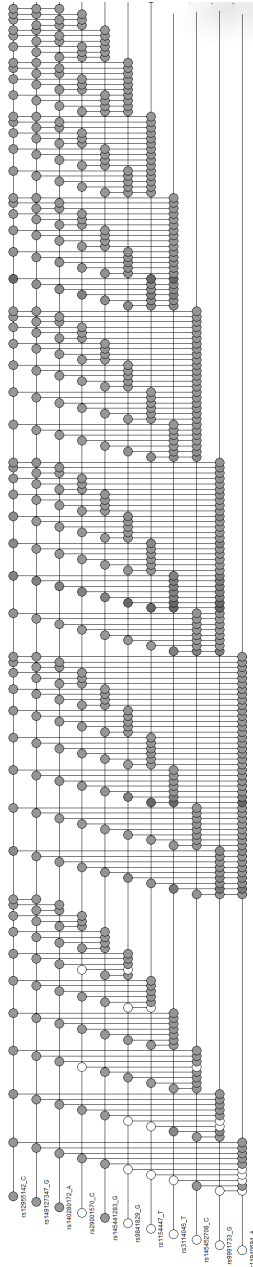


Figure 6.14: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother, Father, and Child Study, grouping alcohol consumption into 3 categories.This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.
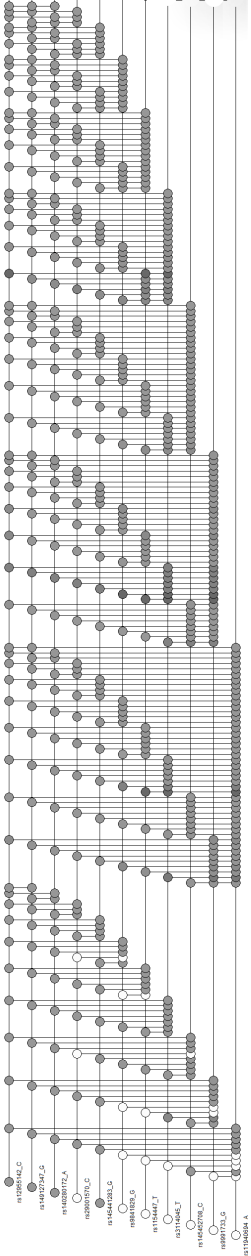
Figure 6.15: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother, Father, and Child Study, grouping alcohol consumption into 4 categories.This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.
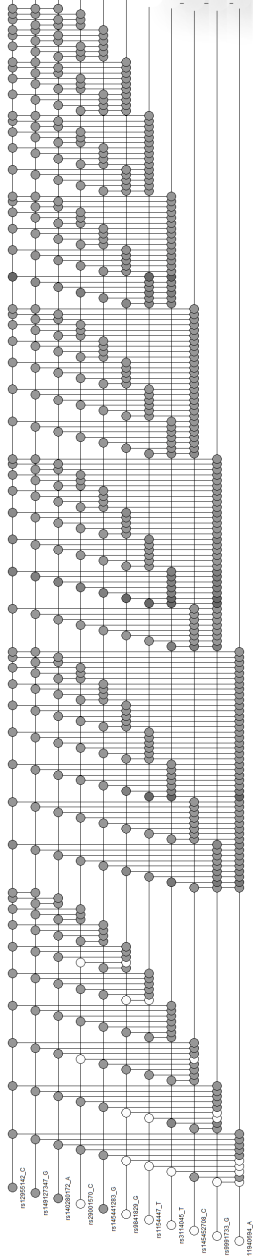


Figure 6.16: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother, Father, and Child Study, grouping alcohol consumption into 7 categories. This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.

229
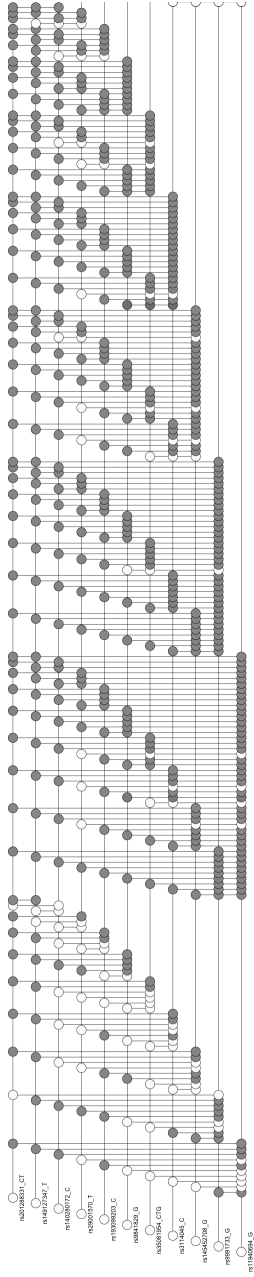
Figure 6.17: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of any vs. no alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children, inverse probability weighted for 10 principal components. This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.
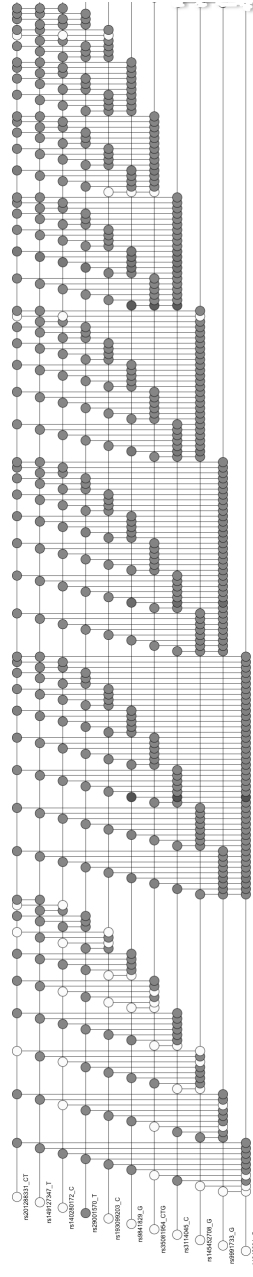


Figure 6.18: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of moderate vs. no alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children, inverse probability weighted for 10 principal components. This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.
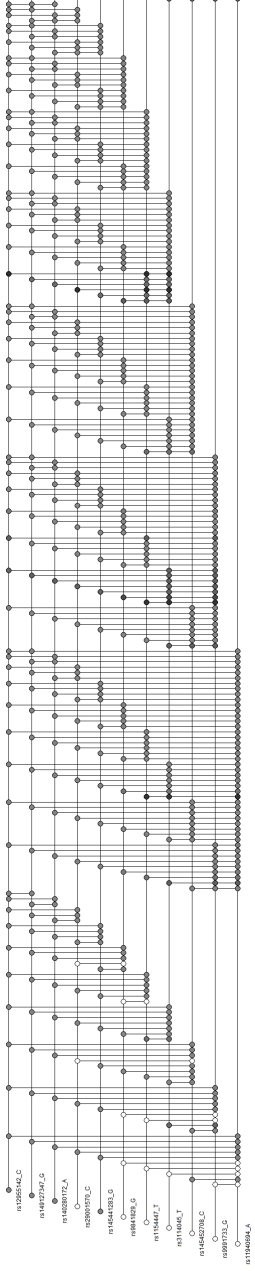
Figure 6.19: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of any vs. no alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother, Father, and Child Study, inverse probability weighted for 10 principal components. This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.
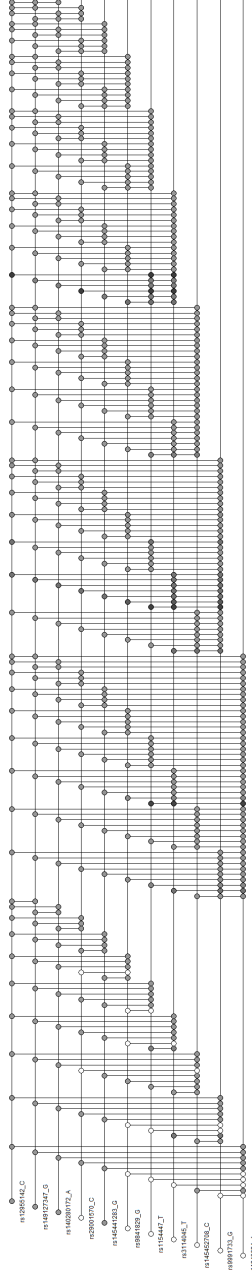


Figure 6.20: Cropped visualization of the application of the instrumental inequalities to models for the average causal effect of moderate vs. no alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother, Father, and Child Study, inverse probability weighted for 10 principal components. This visualization is cropped such that sets of proposed instruments not shown violated the instrumental inequalities.

231
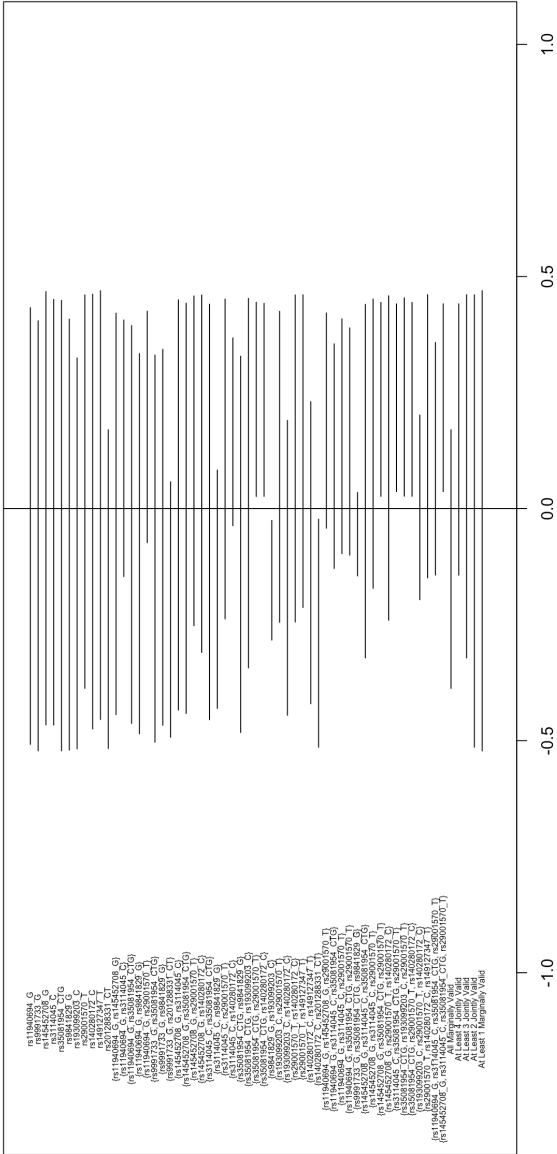
Figure 6.21: Bounds on the average causal effect of any vs. no alcohol consumption on offspring ADHD in the Avon Longitudinal Study of Parents and Children, IP weighted for 10 principal components.
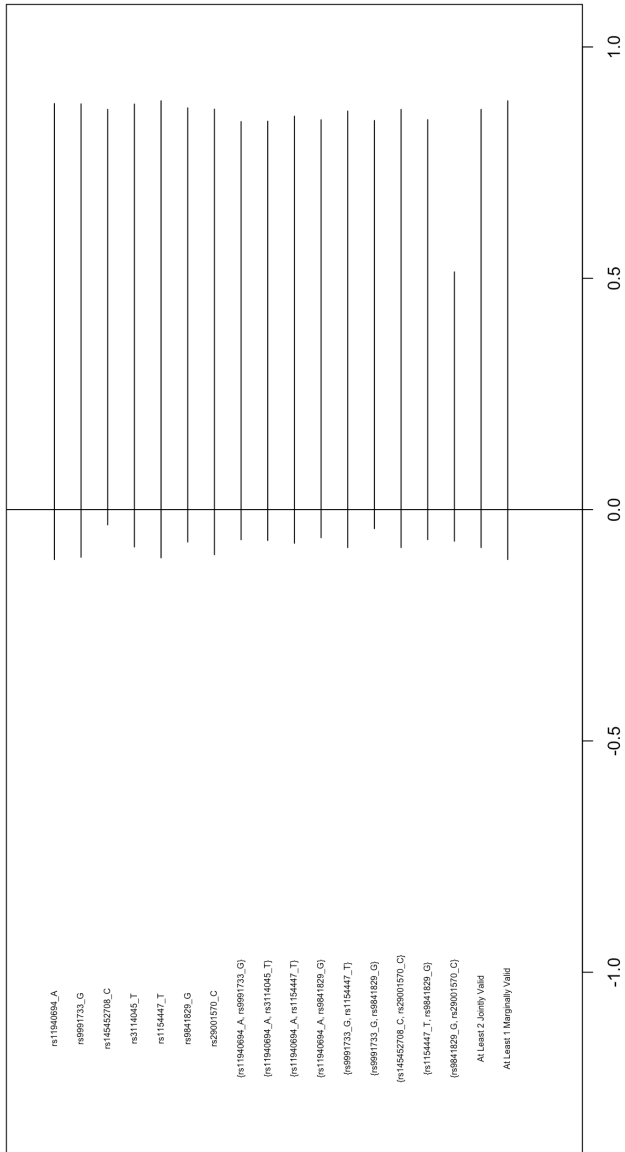
Figure 6.22: Bounds on the average causal effect of any vs. no alcohol consumption on offspring ADHD in the Norwegian Mother, Father, and Child Study, IP weighted for 10 principal components.
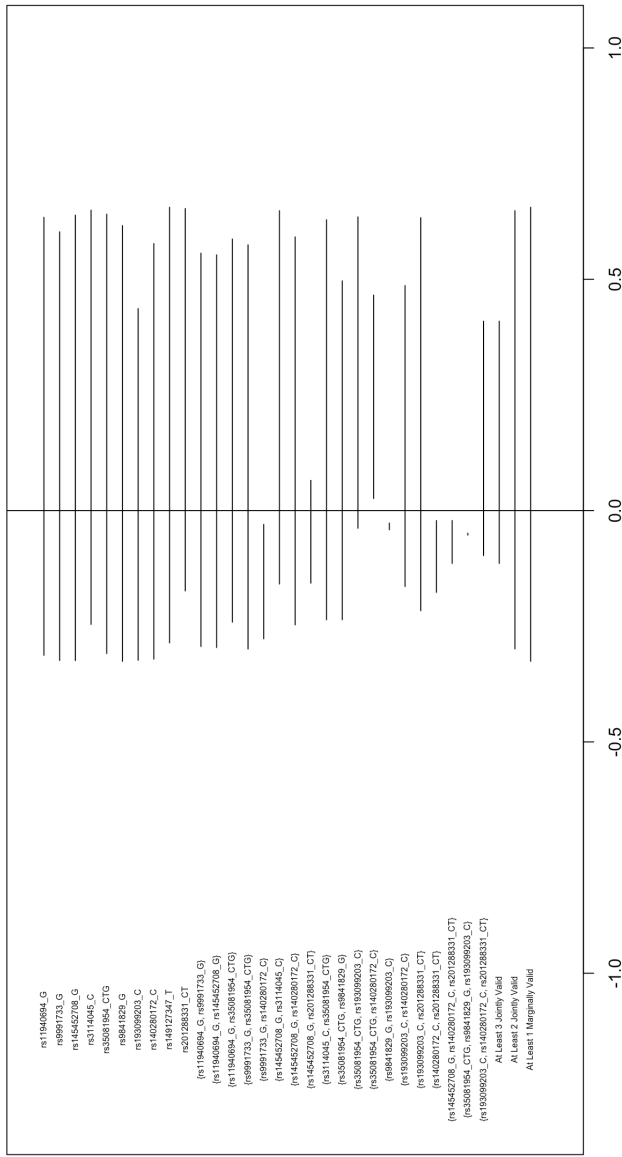
Figure 6.23: Bounds on the average causal effect of moderate vs. no alcohol consumption on offspring ADHD in the Avon Longitudinal Study of Parents and Children, IP weighted for 10 principal components.

234

Figure 6.24: Bounds on the average causal effect of moderate vs. no alcohol consumption on offspring ADHD in the Norwegian Mother, Father, and Child Study, IP weighted for 10 principal components.

# Supplementary tables

Table 6.2: Bounds and point estimates for the average causal effect of any vs. no alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| rs11940694 | -0.51 | 0.43 | 0.06 (-0.19, 0.31) |
| rs9991733 | -0.52 | 0.41 | 0.02 (-0.59, 0.53) |
| rs145452708 | -0.47 | 0.47 | 0.11 (-2.22, 2.06) |
| rs3114045 | -0.48 | 0.45 | 0.02 (-0.37, 0.39) |
| rs35081954 | -0.52 | 0.45 | -0.31 (-1.46, 0.51) |
| rs9841829 | -0.52 | 0.42 | -0.03 (-0.75, 0.61) |
| rs193099203 | -0.52 | 0.33 | -0.16 (-2.17, 1.19) |
| rs29001570 | -0.39 | 0.46 | 0.14 (-1.11, 1.18) |
| rs140280172 | -0.47 | 0.46 | 0.14 (-1.11, 1.18) |
| rs149127347 | -0.45 | 0.47 | -1.7 (-9.31, 0.34) |
| rs201288331 | -0.51 | 0.32 | 0 (-0.6, 0.58) |
| {rs11940694, rs145452708} | -0.41 | 0.42 | 0.09 (-0.05, 0.25) |
| {rs11940694, rs3114045} | -0.15 | 0.40 | 0.07 (-0.05, 0.2) |
| {rs11940694, rs35081954} | -0.46 | 0.38 | 0.06 (-0.06, 0.23) |
| {rs11940694, rs9841829} | -0.48 | 0.36 | 0.04 (-0.1, 0.23) |
| {rs11940694, rs29001570} | -0.49 | 0.42 | 0.08 (-0.13, 0.29) |
| {rs11940694, rs201288331} | -0.32 | 0.14 | 0.02 (-0.11, 0.17) |
| {rs9991733, rs35081954} | -0.50 | 0.33 | -0.05 (-0.31, 0.11) |
| {rs9991733, rs9841829} | -0.44 | 0.31 | 0.15 (-0.02, 0.44) |
| {rs9991733, rs201288331} | -0.31 | 0.15 | -0.07 (-0.22, 0.05) |
| {rs145452708, rs3114045} | -0.38 | 0.45 | 0.02 (-0.19, 0.25) |

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
| --- | --- | --- | --- |
| {rs145452708, rs35081954} | -0.43 | 0.44 | -0.03 (-0.45, 0.31) |
| {rs145452708, rs29001570} | -0.39 | 0.46 | 0.09 (-0.33, 0.57) |
| {rs145452708, rs140280172} | -0.37 | 0.46 | 0.04 (-0.22, 0.45) |
| {rs3114045, rs35081954} | -0.47 | 0.44 | 0.12 (-0.19, 0.53) |
| {rs3114045, rs9841829} | -0.38 | 0.39 | 0.02 (-0.14, 0.19) |
| {rs3114045, rs29001570} | -0.39 | 0.45 | 0.09 (-0.14, 0.46) |
| {rs3114045, rs140280172} | -0.32 | 0.45 | 0.03 (-0.13, 0.28) |
| {rs35081954, rs9841829} | -0.48 | 0.37 | -0.2 (-0.65, -0.09) |
| {rs35081954, rs193099203} | -0.50 | 0.32 | -0.27 (-0.84, -0.12) |
| {rs35081954, rs29001570} | 0.03 | 0.44 | -0.07 (-0.48, 0.28) |
| {rs35081954, rs140280172} | 0.03 | 0.44 | -0.05 (-0.35, 0.22) |
| {rs9841829, rs193099203} | -0.50 | -0.03 | -0.06 (-0.29, 0.11) |
| {rs193099203, rs29001570} | -0.39 | 0.35 | -0.06 (-0.37, 0.17) |
| {rs193099203, rs140280172} | -0.49 | 0.21 | -0.06 (-0.24, 0.03) |
| {rs29001570, rs140280172} | -0.39 | 0.46 | 0.26 (-0.01, 0.91) |
| {rs29001570, rs149127347} | -0.35 | 0.46 | -0.13 (-1.02, 0.75) |
| {rs140280172, rs149127347} | -0.37 | 0.46 | 0.36 (-0.16, 1.27) |
| {rs140280172, rs201288331} | -0.45 | -0.02 | 0.07 (-0.18, 0.34) |
| {rs11940694, rs145452708, rs29001570} | -0.41 | 0.42 | 0.11 (0, 0.26) |
| {rs11940694, rs3114045, rs35081954} | -0.14 | 0.36 | 0.07 (-0.01, 0.21) |
| {rs11940694, rs3114045, rs29001570} | -0.18 | 0.41 | 0.09 (-0.01, 0.26) |
| {rs11940694, rs35081954, rs29001570} | -0.46 | 0.38 | 0.09 (-0.04, 0.26) |
| {rs9991733, rs35081954, rs9841829} | -0.30 | 0.17 | -0.01 (-0.12, 0.09) |
| {rs145452708, rs3114045, rs35081954} | -0.37 | 0.44 | 0.16 (0.02, 0.5) |
| {rs145452708, rs3114045, rs29001570} | -0.23 | 0.45 | 0.05 (-0.06, 0.23) |
| {rs145452708, rs35081954, rs29001570} | 0.03 | 0.44 | 0.02 (-0.2, 0.28) |
| {rs145452708, rs29001570, rs140280172} | -0.37 | 0.46 | 0.04 (-0.13, 0.31) |

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
| --- | --- | --- | --- |
| {rs3114045, rs35081954, rs941829} | -0.35 | 0.32 | -0.01 (-0.17, 0.15) |
| {rs3114045, rs35081954, rs29001570} | 0.04 | 0.44 | 0.18 (0, 0.61) |
| {rs3114045, rs29001570, rs140280172} | 0.02 | 0.45 | 0.07 (-0.01, 0.28) |
| {rs35081954, rs193099203, rs29001570} | 0.03 | 0.33 | -0.1 (-0.41, 0.08) |
| {rs35081954, rs29001570, rs140280172} | 0.03 | 0.44 | 0 (-0.19, 0.21) |
| {rs193099203, rs29001570, rs140280172} | -0.39 | 0.22 | -0.01 (-0.11, 0.07) |
| {rs29001570, rs140280172, rs149127347} | -0.29 | 0.46 | 0.07 (-0.37, 0.46) |
| {rs11940694, rs3114045, rs35081954, rs29001570} | -0.17 | 0.37 | 0.09 (0.01, 0.24) |
| {rs145452708, rs3114045, rs35081954, rs29001570} | 0.04 | 0.44 | 0.16 (0.03, 0.43) |
| At Least 4 Jointly Valid | -0.17 | 0.44 | |
| At Least 3 Jointly Valid | -0.46 | 0.46 | |
| At Least 2 Jointly Valid | -0.50 | 0.46 | |
| At Least 1 Marginally Valid | -0.52 | 0.47 | |

Table 6.3: Bounds and point estimates for the average causal effect of any vs. no alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother and Child Study

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| rs11940694 | -0.11 | 0.88 | -0.71 (-2.74, 0.81) |
| rs9991733 | -0.10 | 0.87 | -0.32 (-2.68, 1.43) |
| rs145452708 | -0.03 | 0.87 | 0.49 (-1.9, 3.28) |
| rs3114045 | -0.09 | 0.88 | 0.24 (-2.95, 3.38) |
| rs35081954 | -0.10 | 0.88 | 0.13 (-1.06, 1.5) |
| rs9841829 | -0.08 | 0.87 | 0.3 (-0.37, 0.79) |
| rs29001570 | -0.11 | 0.86 | -0.94 (-10.64, 5.47) |
| {rs11940694, rs9991733} | -0.08 | 0.83 | -0.47 (-1.1, -0.2) |
| {rs11940694, rs3114045} | -0.09 | 0.83 | -0.57 (-1.63, -0.43) |
| {rs11940694, rs35081954} | -0.08 | 0.85 | 0.05 (-0.35, 0.47) |
| {rs11940694, rs9841829} | -0.06 | 0.84 | -0.04 (-0.33, 0.27) |
| {rs9991733, rs35081954} | -0.09 | 0.86 | 0.04 (-0.38, 0.54) |
| {rs9991733, rs9841829} | -0.04 | 0.83 | 0.21 (-0.05, 0.58) |
| {rs145452708, rs29001570} | -0.03 | 0.86 | 0.02 (-1.26, 0.97) |
| {rs35081954, rs9841829} | -0.06 | 0.84 | 0.22 (-0.03, 0.66) |
| {rs9841829, rs29001570} | -0.07 | 0.73 | 0.17 (-0.15, 0.47) |
| At Least 2 Jointly Valid | -0.09 | 0.86 | |
| At Least 1 Marginally Valid | -0.11 | 0.88 | |

Table 6.4: Bounds and point estimates for the average causal effect of moderate vs. no alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| rs11940694 | -0.31 | 0.63 | 0.06 (-0.33, 0.39) |
| rs9991733 | -0.32 | 0.61 | 0.12 (-0.95, 1.55) |
| rs145452708 | -0.32 | 0.64 | -0.09 (-3.96, 2.39) |
| rs3114045 | -0.26 | 0.65 | 0.21 (-0.29, 0.88) |
| rs35081954 | -0.31 | 0.64 | -0.17 (-0.64, 0.24) |
| rs9841829 | -0.33 | 0.64 | 0.03 (-0.75, 1.15) |
| rs193099203 | -0.32 | 0.44 | -0.09 (-0.97, 0.7) |
| rs140280172 | -0.32 | 0.58 | -0.24 (-3.18, 1.99) |
| rs149127347 | -0.29 | 0.66 | 0.61 (-3.08, 6.3) |
| rs201288331 | -0.24 | 0.65 | 0.12 (-0.36, 0.75) |
| {rs11940694, rs9991733} | -0.28 | 0.56 | -0.01 (-0.25, 0.18) |
| {rs11940694, rs145452708} | -0.30 | 0.52 | 0 (-0.21, 0.17) |
| {rs11940694, rs35081954} | -0.24 | 0.60 | 0 (-0.16, 0.15) |
| {rs11940694, rs9841829} | -0.25 | 0.57 | 0.06 (-0.11, 0.29) |
| {rs9991733, rs35081954} | -0.29 | 0.55 | -0.09 (-0.36, 0.09) |
| {rs9991733, rs9841829} | -0.11 | 0.47 | 0.2 (0.05, 0.59) |
| {rs9991733, rs140280172} | -0.21 | -0.03 | -0.01 (-0.17, 0.14) |
| {rs145452708, rs3114045} | -0.23 | 0.62 | 0.13 (-0.07, 0.43) |
| {rs145452708, rs140280172} | -0.22 | 0.56 | -0.18 (-0.67, -0.02) |
| {rs145452708, rs149127347} | -0.19 | 0.63 | 0.3 (-0.42, 1.02) |
| {rs145452708, rs201288331} | -0.23 | 0.32 | 0 (-0.26, 0.21) |
| {rs3114045, rs35081954} | -0.25 | 0.63 | -0.13 (-0.56, 0.14) |
| {rs3114045, rs9841829} | -0.20 | 0.57 | 0.02 (-0.15, 0.25) |
| {rs3114045, rs149127347} | 0.10 | 0.59 | 0.25 (-0.08, 0.59) |

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| {rs35081954, rs9841829} | -0.28 | 0.59 | -0.2 (-0.53, -0.09) |
| {rs35081954, rs193099203} | -0.29 | 0.42 | -0.13 (-0.36, 0.04) |
| {rs35081954, rs140280172} | 0.03 | 0.55 | -0.16 (-0.42, 0) |
| {rs35081954, rs149127347} | 0.17 | 0.53 | -0.11 (-0.7, 0.08) |
| {rs9841829, rs193099203} | -0.32 | -0.03 | -0.07 (-0.29, 0.13) |
| {rs193099203, rs140280172} | -0.28 | 0.32 | -0.06 (-0.19, 0.01) |
| {rs193099203, rs149127347} | -0.32 | 0.32 | -0.09 (-0.26, -0.02) |
| {rs193099203, rs201288331} | -0.24 | 0.56 | 0.01 (-0.24, 0.29) |
| {rs140280172, rs149127347} | -0.19 | 0.54 | 0.08 (-0.55, 0.41) |
| {rs140280172, rs201288331} | -0.24 | -0.02 | -0.03 (-0.29, 0.13) |
| {rs149127347, rs201288331} | 0.08 | 0.65 | 0.09 (-0.59, 0.5) |
| {rs145452708, rs3114045, rs149127347} | 0.11 | -0.11 | 0.1 (-0.22, 0.3) |
| {rs145452708, rs140280172, rs149127347} | 0.25 | 0.53 | 0.18 (-0.37, 0.47) |
| {rs145452708, rs140280172, rs201288331} | -0.09 | -0.04 | -0.08 (-0.27, 0.06) |
| {rs145452708, rs149127347, rs201288331} | 0.20 | -0.20 | 0.08 (-0.33, 0.26) |
| {rs3114045, rs35081954, rs9841829} | -0.13 | 0.49 | -0.18 (-0.43, -0.1) |
| {rs3114045, rs35081954, rs149127347} | 0.25 | 0.32 | -0.1 (-0.5, 0.01) |
| {rs35081954, rs9841829, rs193099203} | -0.28 | -0.04 | -0.13 (-0.33, -0.05) |
| {rs35081954, rs193099203, rs149127347} | -0.29 | -0.03 | -0.11 (-0.27, -0.01) |
| {rs35081954, rs140280172, rs149127347} | 0.33 | 0.32 | 0.05 (-0.32, 0.22) |
| {rs193099203, rs140280172, rs149127347} | 0.02 | -0.02 | -0.03 (-0.11, -0.01) |
| {rs193099203, rs140280172, rs201288331} | -0.24 | 0.49 | 0.01 (-0.13, 0.18) |
| {rs193099203, rs149127347, rs201288331} | 0.02 | 0.49 | 0.01 (-0.15, 0.18) |
| {rs140280172, rs149127347, rs201288331} | 0.12 | -0.12 | 0.02 (-0.41, 0.2) |
| {rs145452708, rs140280172, rs149127347, rs201288331} | 0.50 | -0.50 | 0.09 (-0.3, 0.22) |
| {rs193099203, rs140280172, rs149127347, rs201288331} | 0.02 | 0.32 | 0.01 (-0.13, 0.12) |
| At Least 4 Jointly Valid | 0.02 | 0.32 | |

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| At Least 3 Jointly Valid | -0.29 | 0.53 | |
| At Least 2 Jointly Valid | -0.32 | 0.65 | |
| At Least 1 Marginally Valid | -0.33 | 0.66 | |

Table 6.5: Bounds and point estimates for the average causal effect of moderate vs. no alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother, Father, and Child Study
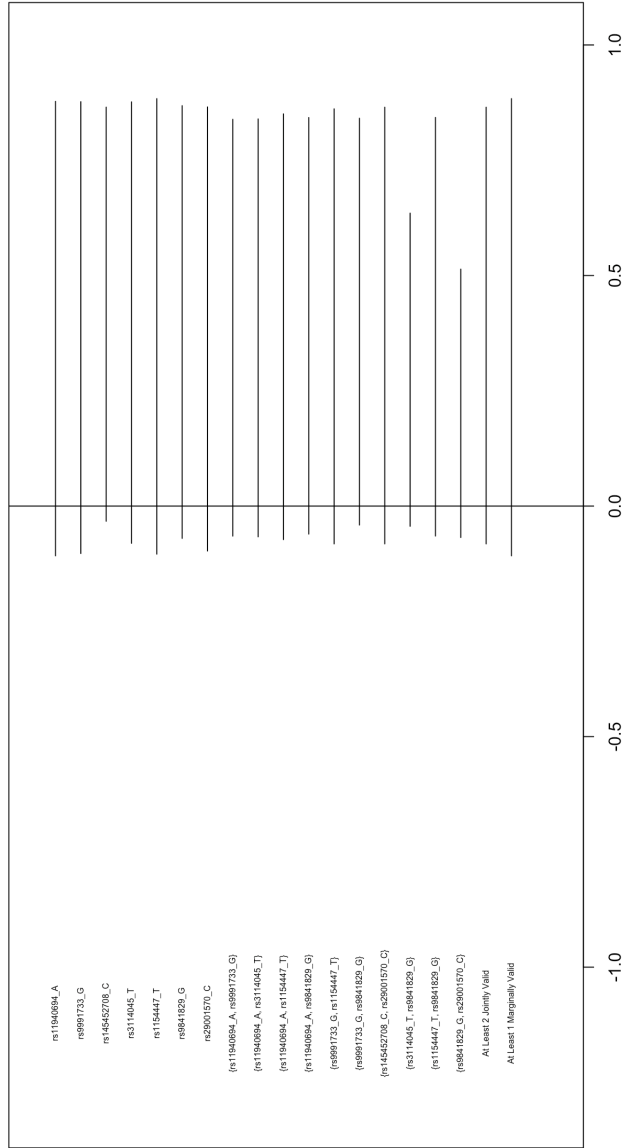
| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| rs11940694 | -0.11 | 0.88 | -0.71 (-2.85, 0.63) |
| rs9991733 | -0.10 | 0.87 | -0.32 (-2.63, 1.66) |
| rs145452708 | -0.03 | 0.87 | 0.49 (-2.14, 3.48) |
| rs3114045 | -0.09 | 0.88 | 0.24 (-2.56, 3.33) |
| rs35081954 | -0.10 | 0.88 | 0.13 (-1.19, 1.7) |
| rs9841829 | -0.08 | 0.87 | 0.3 (-0.36, 0.78) |
| rs29001570 | -0.11 | 0.86 | -0.94 (-8.66, 5.33) |
| {rs11940694, rs9991733} | -0.08 | 0.83 | -0.47 (-1.14, -0.17) |
| {rs11940694, rs3114045} | -0.09 | 0.83 | -0.57 (-1.63, -0.46) |
| {rs11940694, rs35081954} | -0.08 | 0.85 | 0.05 (-0.3, 0.43) |
| {rs11940694, rs9841829} | -0.06 | 0.84 | -0.04 (-0.31, 0.25) |
| {rs9991733, rs35081954} | -0.09 | 0.86 | 0.04 (-0.38, 0.47) |
| {rs9991733, rs9841829} | -0.04 | 0.83 | 0.21 (-0.03, 0.58) |
| {rs145452708, rs29001570} | -0.03 | 0.86 | 0.02 (-0.99, 0.87) |
| {rs35081954, rs9841829} | -0.06 | 0.84 | 0.22 (-0.05, 0.67) |
| {rs9841829, rs29001570} | -0.07 | 0.73 | 0.17 (-0.12, 0.49) |
| At Least 2 Jointly Valid | -0.09 | 0.86 | |
| At Least 1 Marginally Valid | -0.11 | 0.88 | |

Table 6.6: Bounds and point estimates for the average causal effect of any vs. no alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children, inverse probability weighted for 10 principal components

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| rs11940694 | -0.51 | 0.43 | 0.05 (-0.21, 0.3) |
| rs9991733 | -0.52 | 0.40 | 0.06 (-0.51, 0.56) |
| rs145452708 | -0.47 | 0.47 | 0.11 (-1.78, 2.2) |
| rs3114045 | -0.47 | 0.45 | 0.29 (-0.02, 1.1) |
| rs35081954 | -0.52 | 0.45 | -0.33 (-1.88, 0.47) |
| rs9841829 | -0.52 | 0.41 | 0.08 (-0.58, 0.95) |
| rs193099203 | -0.52 | 0.32 | -0.16 (-1.55, 1.15) |
| rs29001570 | -0.39 | 0.46 | 0.14 (-1.55, 1.15) |
| rs140280172 | -0.47 | 0.46 | 0.42 (-3.81, 5.03) |
| rs149127347 | -0.45 | 0.47 | -1.62 (-8.6, 2) |
| rs201288331 | -0.52 | 0.17 | -0.06 (-0.33, 0.11) |
| {rs11940694, rs145452708} | -0.44 | 0.42 | 0.2 (0.21, 0.47) |
| {rs11940694, rs3114045} | -0.15 | 0.41 | 0.07 (0.03, 0.13) |
| {rs11940694, rs35081954} | -0.46 | 0.39 | 0 (-0.21, 0.13) |
| {rs11940694, rs9841829} | -0.49 | 0.33 | -0.03 (-0.35, 0.2) |
| {rs11940694, rs29001570} | -0.07 | 0.43 | 0.04 (-0.03, 0.16) |
| {rs9991733, rs35081954} | -0.50 | 0.33 | -0.03 (-0.22, 0.11) |
| {rs9991733, rs9841829} | -0.47 | 0.34 | 0.29 (0.19, 0.87) |
| {rs9991733, rs201288331} | -0.49 | 0.06 | -0.06 (-0.1, -0.02) |
| {rs145452708, rs3114045} | -0.43 | 0.45 | 0.21 (0.21, 0.55) |
| {rs145452708, rs35081954} | -0.44 | 0.44 | 0.2 (0.1, 0.67) |
| {rs145452708, rs29001570} | -0.25 | 0.46 | 0.07 (-0.11, 0.23) |
| {rs145452708, rs140280172} | -0.31 | 0.46 | 0.05 (-0.06, 0.26) |

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
| --- | --- | --- | --- |
| {rs3114045, rs35081954} | -0.45 | 0.44 | 0.27 (0.12, 0.84) |
| {rs3114045, rs9841829} | -0.43 | 0.08 | -0.03 (-0.21, 0.03) |
| {rs3114045, rs29001570} | -0.24 | 0.45 | 0.07 (-0.09, 0.21) |
| {rs3114045, rs140280172} | -0.04 | 0.37 | 0.03 (-0.05, 0.1) |
| {rs35081954, rs9841829} | -0.48 | 0.33 | -0.11 (-0.45, 0.06) |
| {rs35081954, rs193099203} | -0.34 | 0.45 | 0.12 (0.1, 0.34) |
| {rs35081954, rs29001570} | 0.03 | 0.44 | 0.04 (-0.01, 0.18) |
| {rs35081954, rs140280172} | 0.03 | 0.44 | 0.07 (-0.03, 0.26) |
| {rs9841829, rs193099203} | -0.28 | -0.03 | 0.02 (-0.08, 0.14) |
| {rs193099203, rs29001570} | -0.24 | 0.42 | 0.04 (-0.13, 0.21) |
| {rs193099203, rs140280172} | -0.45 | 0.19 | -0.01 (-0.09, 0.06) |
| {rs29001570, rs140280172} | -0.24 | 0.46 | 0.06 (-0.18, 0.18) |
| {rs29001570, rs149127347} | -0.21 | 0.46 | 0.05 (-0.48, 0.64) |
| {rs140280172, rs149127347} | -0.42 | 0.23 | -0.03 (-0.53, 0.25) |
| {rs140280172, rs201288331} | -0.51 | -0.02 | -0.04 (-0.15, 0.05) |
| {rs11940694, rs145452708, rs29001570} | -0.04 | 0.42 | 0.03 (-0.05, 0.09) |
| {rs11940694, rs3114045, rs35081954} | -0.13 | 0.35 | 0.06 (0.03, 0.13) |
| {rs11940694, rs3114045, rs29001570} | -0.10 | 0.41 | 0.05 (0.02, 0.08) |
| {rs11940694, rs35081954, rs29001570} | -0.10 | 0.39 | 0.04 (-0.02, 0.15) |
| {rs9991733, rs35081954, rs9841829} | -0.15 | 0.03 | -0.03 (-0.1, 0.04) |
| {rs145452708, rs3114045, rs35081954} | -0.32 | 0.44 | 0.12 (0.07, 0.26) |
| {rs145452708, rs3114045, rs29001570} | -0.17 | 0.45 | 0.05 (0, 0.13) |
| {rs145452708, rs35081954, rs29001570} | 0.03 | 0.44 | 0.04 (-0.03, 0.11) |
| {rs145452708, rs29001570, rs140280172} | -0.24 | 0.46 | 0.06 (0.02, 0.16) |
| {rs3114045, rs35081954, rs29001570} | 0.04 | 0.44 | 0.04 (-0.07, 0.16) |
| {rs35081954, rs193099203, rs29001570} | 0.03 | 0.45 | 0.04 (0.03, 0.1) |
| {rs35081954, rs29001570, rs140280172} | 0.03 | 0.44 | 0.03 (-0.01, 0.07) |

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
| --- | --- | --- | --- |
| {rs193099203, rs29001570, rs140280172} | -0.20 | 0.20 | 0.01 (-0.02, 0.04) |
| {rs29001570, rs140280172, rs149127347} | -0.15 | 0.46 | -0.04 (-0.29, 0.11) |
| {rs11940694, rs3114045, rs35081954, rs29001570} | -0.14 | 0.36 | 0.06 (0.04, 0.11) |
| {rs145452708, rs3114045, rs35081954, rs29001570} | 0.04 | 0.44 | 0.05 (0.01, 0.1) |
| At Least 4 Jointly Valid | -0.14 | 0.44 | |
| At Least 3 Jointly Valid | -0.32 | 0.46 | |
| At Least 2 Jointly Valid | -0.51 | 0.46 | |
| At Least 1 Marginally Valid | -0.52 | 0.47 | |

Table 6.7: Bounds and point estimates for the average causal effect of moderate vs. no alcohol consumption during pregnancy on offspring ADHD in the Avon Longitudinal Study of Parents and Children, inverse probability weighted for 10 principal components

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| rs11940694 | -0.31 | 0.63 | 0.02 (-0.37, 0.4) |
| rs9991733 | -0.32 | 0.60 | 0.24 (-0.83, 1.62) |
| rs145452708 | -0.32 | 0.64 | -0.08 (-3.8, 2.58) |
| rs3114045 | -0.25 | 0.65 | 0.26 (-0.23, 1.38) |
| rs35081954 | -0.31 | 0.64 | -0.23 (-0.87, 0.42) |
| rs9841829 | -0.33 | 0.62 | 0.24 (-0.69, 1.48) |
| rs193099203 | -0.32 | 0.44 | -0.09 (-0.8, 0.71) |
| rs140280172 | -0.32 | 0.58 | -0.24 (-3.7, 2.68) |
| rs149127347 | -0.29 | 0.66 | 0.61 (-3.52, 5.29) |
| rs201288331 | -0.17 | 0.65 | 0.14 (-0.15, 0.61) |
| {rs11940694, rs9991733} | -0.29 | 0.56 | 0.09 (-0.1, 0.4) |
| {rs11940694, rs145452708} | -0.30 | 0.55 | -0.11 (-0.38, -0.04) |
| {rs11940694, rs35081954} | -0.24 | 0.59 | -0.07 (-0.31, 0.01) |
| {rs9991733, rs35081954} | -0.30 | 0.57 | -0.09 (-0.36, 0.02) |
| {rs9991733, rs140280172} | -0.28 | -0.03 | -0.03 (-0.12, 0.05) |
| {rs145452708, rs3114045} | -0.16 | 0.65 | 0.12 (0.03, 0.32) |
| {rs145452708, rs140280172} | -0.25 | 0.59 | -0.1 (-0.39, 0.02) |
| {rs145452708, rs201288331} | -0.16 | 0.07 | -0.02 (-0.17, 0.01) |
| {rs3114045, rs35081954} | -0.24 | 0.63 | 0.17 (0, 0.68) |
| {rs35081954, rs9841829} | -0.24 | 0.50 | 0.03 (-0.17, 0.3) |
| {rs35081954, rs193099203} | -0.04 | 0.63 | 0.07 (0.04, 0.19) |
| {rs35081954, rs140280172} | 0.03 | 0.47 | 0 (-0.1, 0.14) |
| {rs9841829, rs193099203} | -0.04 | -0.03 | 0.04 (-0.06, 0.18) |

247

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| {rs193099203, rs140280172} | -0.16 | 0.49 | -0.01 (-0.11, 0.07) |
| {rs193099203, rs201288331} | -0.22 | 0.63 | 0.03 (-0.15, 0.27) |
| {rs140280172, rs201288331} | -0.18 | -0.02 | -0.04 (-0.16, 0.07) |
| {rs145452708, rs140280172, rs201288331} | -0.11 | -0.02 | -0.02 (-0.07, 0.01) |
| {rs35081954, rs9841829, rs193099203} | -0.05 | -0.05 | 0.03 (-0.06, 0.12) |
| {rs193099203, rs140280172, rs201288331} | -0.10 | 0.41 | 0 (-0.08, 0.04) |
| At Least 3 Jointly Valid | -0.11 | 0.41 | |
| At Least 2 Jointly Valid | -0.30 | 0.65 | |
| At Least 1 Marginally Valid | -0.33 | 0.66 | |

Table 6.8: Bounds and point estimates for the average causal effect of any vs. no alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother and Child Study, inverse probability weighted for 10 principal components

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
|---|---|---|---|
| rs11940694 | -0.11 | 0.88 | -0.83 (-2.89, 0.44) |
| rs9991733 | -0.10 | 0.88 | -0.31 (-2.83, 1.75) |
| rs145452708 | -0.03 | 0.87 | 0.49 (-2.61, 5.07) |
| rs3114045 | -0.08 | 0.88 | -0.04 (-1.92, 1.99) |
| rs35081954 | -0.10 | 0.88 | 0.2 (-1.46, 2.07) |
| rs9841829 | -0.07 | 0.87 | 0.41 (-0.42, 0.8) |
| rs29001570 | -0.10 | 0.87 | -2.28 (-17.79, 6.09) |
| {rs11940694, rs9991733} | -0.07 | 0.84 | -0.39 (-1.07, 0.47) |
| {rs11940694, rs3114045} | -0.07 | 0.84 | -0.23 (-0.86, 0.25) |
| {rs11940694, rs35081954} | -0.07 | 0.85 | 0.18 (-0.18, 0.69) |
| {rs11940694, rs9841829} | -0.06 | 0.84 | 0.22 (-0.01, 0.67) |
| {rs9991733, rs35081954} | -0.08 | 0.86 | -0.02 (-0.74, 0.52) |
| {rs9991733, rs9841829} | -0.04 | 0.84 | 0.27 (0, 0.72) |
| {rs145452708, rs29001570} | -0.08 | 0.87 | -0.4 (-2.2, 0.25) |
| {rs35081954, rs9841829} | -0.06 | 0.84 | 0.33 (0.03, 0.91) |
| {rs9841829, rs29001570} | -0.07 | 0.51 | -0.06 (-0.17, 0.03) |
| At Least 2 Jointly Valid | -0.08 | 0.87 | |
| At Least 1 Marginally Valid | -0.11 | 0.88 | |

249

Table 6.9: Bounds and point estimates for the average causal effect of moderate vs. no alcohol consumption during pregnancy on offspring ADHD in the Norwegian Mother and Child Study, inverse probability weighted for 10 principal components

| Proposed Instruments | Lower Bound | Upper Bound | Point Estimate (95% CI) |
| --- | --- | --- | --- |
| rs11940694 | -0.11 | 0.88 | -0.83 (-3.12, 0.35) |
| rs9991733 | -0.10 | 0.88 | -0.31 (-2.65, 1.26) |
| rs145452708 | -0.03 | 0.87 | 0.49 (-2.03, 2.9) |
| rs3114045 | -0.08 | 0.88 | -0.04 (-1.89, 2.22) |
| rs35081954 | -0.10 | 0.88 | 0.2 (-1.24, 1.78) |
| rs9841829 | -0.07 | 0.87 | 0.41 (-0.34, 0.8) |
| rs29001570 | -0.10 | 0.87 | -2.28 (-15.35, 3.77) |
| {rs11940694, rs9991733} | -0.07 | 0.84 | -0.39 (-1.08, 0.49) |
| {rs11940694, rs3114045} | -0.07 | 0.84 | -0.23 (-0.83, 0.41) |
| {rs11940694, rs35081954} | -0.07 | 0.85 | 0.18 (-0.12, 0.7) |
| {rs11940694, rs9841829} | -0.06 | 0.84 | 0.22 (0.01, 0.67) |
| {rs9991733, rs35081954} | -0.08 | 0.86 | -0.02 (-0.78, 0.54) |
| {rs9991733, rs9841829} | -0.04 | 0.84 | 0.27 (-0.02, 0.71) |
| {rs145452708, rs29001570} | -0.08 | 0.87 | -0.4 (-2.48, 0.24) |
| {rs3114045, rs9841829} | -0.04 | 0.64 | -0.06 (-0.53, 0.12) |
| {rs35081954, rs9841829} | -0.06 | 0.84 | 0.33 (0.04, 0.96) |
| {rs9841829, rs29001570} | -0.07 | 0.51 | -0.06 (-0.18, 0.01) |
| At Least 2 Jointly Valid | -0.08 | 0.87 | |
| At Least 1 Marginally Valid | -0.11 | 0.88 | |

Table 6.10: Correlation between maternal and paternal genotypes in the Norwegian Mother, Father, and Child Study.

| SNP | Chromosome | MAF | HWE p-value | Batch specific imputation score (m12) | Batch specific imputation score (m24) | Batch specific imputation score (njl) | MoBa maternal-paternal correlation |
|---|---|---|---|---|---|---|---|
| rs1154447 | 4 | 0.452 | 0.226 | 0.986 | 0.985 | 0.793 | -0.001 |
| rs11940694 | 4 | 0.381 | 0.265 | 1.000 | 1.000 | 1.000 | 0.006 |
| rs12955142 | 18 | 0.046 | 0.401 | 0.810 | 0.807 | 0.976 | 0.007 |
| rs140280172 | 4 | 0.001 | 1.000 | 0.716 | 0.739 | 0.652 | -0.002 |
| rs145441283 | 4 | 0.002 | 0.051 | 0.499 | 0.548 | 0.531 | -0.004 |
| rs145452708 | 4 | 0.003 | 0.090 | 0.752 | 0.748 | 0.787 | -0.006 |
| rs149127347 | 4 | 0.001 | 0.013 | 0.537 | 0.663 | 0.658 | -0.002 |
| rs29001570 | 4 | 0.004 | 1.000 | 0.710 | 0.753 | 0.840 | 0.010 |
| rs3114045 | 4 | 0.121 | 0.334 | 0.965 | 0.960 | 0.970 | 0.008 |
| rs9841829 | 3 | 0.238 | 0.011 | 0.999 | 0.999 | 0.996 | 0.003 |
| rs9991733 | 4 | 0.271 | 0.575 | 0.999 | 0.999 | 0.958 | 0.004 |

# References

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the aseba preschool forms and profiles*. Burlington, VT: University of Vermont, Research center for children, youth, families.

Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of business & economic statistics*, *19*(1), 2–28.

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176.

Bonet, B. (2001). Instrumentality tests revisited. *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 48–55.

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., & Davey Smith, G. (2013). Cohort profile: The 'children of the 90s'—the index offspring of the avon longitudinal study of parents and children. *International journal of epidemiology*, *42*(1), 111–127.

Brazel, D. M., Jiang, Y., Hughey, J. M., Turcot, V., Zhan, X., Gong, J., Batini, C., Weissenkampen, J. D., Liu, M., & Surendran, P. (2019). Exome chip meta-analysis fine maps causal variants and elucidates the genetic architecture of rare coding variants in smoking and alcohol use. *Biological psychiatry*, *85*(11), 946–955.

Burgess, S., & Labrecque, J. A. (2018). Mendelian randomization with a binary exposure variable: Interpretation and presentation of causal estimates. *European journal of epidemiology*, *33*(10), 947–952.

Chambers, J. C., Elliott, P., Zabaneh, D., Zhang, W., Li, Y., Froguel, P., Balding, D., Scott, J., & Kooner, J. S. (2008). Common genetic variation near mc4r is associated with waist circumference and insulin resistance. *Nature genetics*, *40*(6), 716.

Clarke, T.-K., Adams, M. J., Davies, G., Howard, D. M., Hall, L. S., Padmanabhan, S., Murray, A. D., Smith, B. H., Campbell, A., & Hayward, C. (2017). Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in uk biobank (n= 112 117). *Molecular psychiatry*, *22*(10), 1376.

Cole, S. R., Hudgens, M. G., Edwards, J. K., Brookhart, M. A., Richardson, D. B., Westreich, D., & Adimora, A. A. (2019). Nonparametric bounds for the risk function. *American journal of epidemiology*, *188*(4), 632–636.

Diemer, E. W., Labrecque, J., Tiemeier, H., & Swanson, S. A. (2020). Application of the instrumental inequalities to a mendelian randomization study with multiple proposed instruments. *Epidemiology*, *31*(1), 65–74.

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., & Ness, A. (2013). Cohort profile: The avon longitudinal study of parents and children: Alspac mothers cohort. *International journal of epidemiology*, *42*(1), 97–110.

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The development and well-being assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of child psychology and psychiatry*, *41*(5), 645–655.

Greenland, S., & Mansournia, M. A. (2015). Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *European journal of epidemiology*, *30*(10), 1101–1110.

Han, J.-Y., Kwon, H.-J., Ha, M., Paik, K.-C., Lim, M.-H., Lee, S. G., Yoo, S.-J., & Kim, E.-J. (2015). The effects of prenatal exposure to alcohol and environmental tobacco smoke on risk for adhd: A large population-based study. *Psychiatry research*, *225*(1-2), 164–168.

Hartwig, F. P., Davies, N. M., & Davey Smith, G. (2018). Bias in mendelian randomization due to assortative mating. *Genetic epidemiology*, *42*(7), 608–620.

Helgeland, Ø., Vaudel, M., Juliusson, P. B., Holmen, O. L., Juodakis, J., Bacelis, J., Jacobsson, B., Lindekleiv, H., Hveem, K., Lie, R. T., et al. (2019). Genome-wide association study reveals dynamic role of genetic variation in infant and early childhood growth. *Nature communications*, *10*(1), 1–10.

Hernan, M. A., & Robins, J. M. (2018). *Causal inference*. Chapman; Hall/CRC.

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 360–372.

Johnston, K., Gustafson, P., Levy, A., & Grootendorst, P. (2008). Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in medicine*, *27*(9), 1539–1556.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., & Matsuda, K. (2018). Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nature genetics*, *50*(3), 390–400.

253

Kleiber, C., Zeileis, A., & Zeileis, M. A. (2020). Package 'aer'. *R package version 1.2, 4*.

Labrecque, J. A., & Swanson, S. A. (2019). Interpretation and potential biases of mendelian randomization estimates with time-varying exposures. *American journal of epidemiology*, *188*(1), 231–238.

Lawlor, D. A. (2016). Commentary: Two-sample mendelian randomization: Opportunities and challenges. *International journal of epidemiology*, *45*(3), 908.

Linnér, R. K., Biroli, P., Kong, E., Meddens, S. F. W., Wedow, R., Fontana, M. A., Lebreton, M., Tino, S. P., Abdellaoui, A., & Hammerschlag, A. R. (2019). Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature genetics*, *51*(2), 245–257.

Linnet, K. M., Dalsgaard, S., Obel, C., Wisborg, K., Henriksen, T. B., Rodriguez, A., Kotimaa, A., Moilanen, I., Thomsen, P. H., & Olsen, J. (2003). Maternal lifestyle factors in pregnancy risk of attention deficit hyperactivity disorder and associated behaviors: Review of the current evidence. *American Journal of Psychiatry*, *160*(6), 1028–1040.

Machiela, M. J., & Chanock, S. J. (2015). Ldlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, *31*(21), 3555–3557.

Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Daltveit, A. K., Handal, M., Haugen, M., Høiseth, G., & Knudsen, G. P. (2016). Cohort profile update: The norwegian mother and child cohort study (moba). *International journal of epidemiology*, *45*(2), 382–388.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, *80*(2), 319–323.

Northstone, K., Lewcock, M., Groom, A., Boyd, A., Macleod, J., Timpson, N., & Wells, N. (2019). The avon longitudinal study of parents and children (alspac): An update on the enrolled sample of index children in 2019. *Wellcome open research*, *4*.

Pagnin, D., Grecco, M. L. Z., & Furtado, E. F. (2019). Prenatal alcohol use as a risk for attention-deficit/hyperactivity disorder. *European archives of psychiatry and clinical neuroscience*, *269*(6), 681–687.

Paltiel, L., Anita, H., Skjerden, T., Harbak, K., Bækken, S., Kristin, S. N., Knudsen, G. P., & Magnus, P. (2014). The biobank of the norwegian mother and child cohort study–present status. *Norsk epidemiologi*, *24*(1-2).

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 435–443.

Rassen, J. A., Schneeweiss, S., Glynn, R. J., Mittleman, M. A., & Brookhart, M. A. (2009). Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *American journal of epidemiology*, *169*(3), 273–284.

Richardson, T. S., & Robins, J. M. (2014). Ace bounds; sems with equilibrium conditions. *Statistical Science*, *29*(3), 363–366.

Ripley, M. B. (2010). Package 'boot'. *beaver*, *9*.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, *23*(8), 2379–2412.

Robins, J. M. (1997). Marginal structural models.

Robins, J. M. (2014). Structural nested failure time models. *Wiley StatsRef: statistics reference online*.

Robins, J. M., & Greenland, S. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, *91*(434), 456–458.

Salanti, G., Amountza, G., Ntzani, E. E., & Ioannidis, J. P. A. (2005). Hardy–weinberg equilibrium in genetic association studies: An empirical evaluation of reporting, deviations, and power. *European journal of human genetics*, *13*(7), 840–848.

Strawbridge, R. J., Ward, J., Cullen, B., Tunbridge, E. M., Hartz, S., Bierut, L., Horton, A., Bailey, M. E. S., Graham, N., & Ferguson, A. (2018). Genome-wide analysis of self-reported risk-taking behaviour and cross-disorder genetic correlations in the uk biobank cohort. *Translational psychiatry*, *8*(1), 1–11.

Swanson, S. A. (2017). Commentary: Can we see the forest for the ivs? mendelian randomization studies with multiple genetic variants. *Epidemiology*, *28*(1), 43–46.

Swanson, S. A., & Hernán, M. A. (2013). Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, *24*(3), 370–374.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., & Richardson, T. S. (2018). Partial identification of the average treatment effect using in-

strumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, *113*(522), 933–947.

Swanson, S. A., Holme, Ø., Løberg, M., Kalager, M., Bretthauer, M., Hoff, G., Aas, E., & Hernán, M. A. (2015). Bounding the per-protocol effect in randomized trials: An application to colorectal cancer screening. *Trials*, *16*(1), 541.

Tamer, E. (2010). Partial identification in econometrics. *Annu. Rev. Econ.*, *2*(1), 167–195.

Tchetgen, E. J. T., Sun, B., & Walter, S. (2017). The genius approach to robust mendelian randomization inference. *arXiv preprint arXiv:1709.07779*.

Team, R. C. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., & Kraft, P. (2014). Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.)*, *25*(3), 427.

van Faassen, E., & Niemelä, O. (2011). Biochemistry of prenatal alcohol exposure.

Zhong, V. W., Kuang, A., Danning, R. D., Kraft, P., Van Dam, R. M., Chasman, D. I., & Cornelis, M. C. (2019). A genome-wide association study of bitter and sweet beverage consumption. *Human molecular genetics*, *28*(14), 2449–2457.

## Chapter 7

# Partial identification of the average causal effect using bounds computed in multiple study populations: the challenge of combining Mendelian randomization studies

Elizabeth W. Diemer, Luisa Zuccolo, Sonja A. Swanson

# 7.1 Abstract

**Background**: Researchers often use random-effects or fixed-effects meta-analysis to combine findings from multiple study populations. However, the causal interpretation of these models is not always clear, and they do not easily translate to settings where bounds, rather than point estimates, are computed.

**Methods**: If bounds on an average causal effect of interest in a well-defined population are computed in multiple study populations under specified identifiability assumptions, then under those assumptions that average causal effect would lie within all study-specific bounds and thus the intersection of the study-specific bounds. We demonstrate this by pooling bounds on the average causal effect of prenatal alcohol exposure on attention deficit-hyperactivity disorder symptoms computed under several sets of assumptions in Mendelian randomization (MR) analyses conducted in two European cohorts.

**Results**: For all assumption sets considered, pooled bounds were wide and did not identify the direction of effect. The narrowest pooled bound computed was [-.041, .306].

**Conclusions**: All pooled bounds computed in our application covered the null, illustrating how strongly point estimates from prior MR studies of this effect rely on within-study homogeneity assumptions. We discuss how the interpretation of both pooled bounds and point estimation in MR is complicated by possible heterogeneity of effects and heterogeneity of exposure definitions across populations. We argue this highlights a broader need for consideration of effect modification and consistency when basing clinical or policy recommendations on MR studies.

## 7.2 Introduction

When data from multiple study populations are available, combining evidence across populations can improve our understanding of causal effects. For example, researchers commonly attempt to synthesize information from multiple studies using meta-analysis, combining study-specific point estimates using either random-effects or fixed-effects models to obtain pooled effect estimates (DerSimonian and Laird, 1986; Higgins et al., 2009; Laird and Mosteller, 1990). However, the causal interpretation of estimates derived from these meta-analyses is not always clear, especially when random-effects models are used (Dahabreh et al., 2020; Manski, 2020). Moreover, traditional meta-analytic approaches do not readily translate to pooling information from studies in which bounds rather than point estimates are computed (Swanson et al., 2018; Tamer, 2010).

Here, we describe and apply an alternative approach to standard meta-analysis, which pools information from study-specific bounds as opposed to study-specific point estimates. In brief, we demonstrate how, if each study is viewed as a random sample from the same well-defined superpopulation, logical combinations of the data and underlying assumptions allow for partial identification of causal effects by the intersection or union of the bounds computed in each study (Manski, 2020). While bounds can be computed in a variety of study designs, our application focuses on pooling two sets of Mendelian randomization (MR) analyses, an application of instrumental variable methods proposing genetic variants as instruments, in order to bound the average causal effect of alcohol consumption during pregnancy on offspring attention deficit hyperactivity disorder (ADHD) symptoms (Chapter 6). The individual studies computed bounds under many different sets of assumptions, as they had proposed multiple genetic variants as instrumental variables (Chapter 6; Swanson, 2017), thereby giving an opportunity to explore how different sets of assumptions may come together in this pooled approach. We begin by describing the general theory.

## 7.3 Pooling two bounds computed across studies under the same set of assumptions

Suppose we are interested in the average causal effect of an exposure $A$ on an outcome $Y$, $\mathrm{E}(Y^{a=a} - Y^{a=a'})$, in some well-defined population. We conduct

$k$ studies, which we will index with $S = \{1, 2, \ldots k\}$, and within each study have computed bounds on this population average causal effect for each of the $k$ studies under some arbitrary set of identifiability assumptions. Then, assuming all sets of identifiability assumptions hold, the average causal effect $E(Y^{a=a} - Y^{a=a'})$ is bounded by the intersection of all these bounds, that is, $[max_s(LB_s), min_s(UB_s)]$. A simple proof of this is given in the Appendix. (We note that we are not claiming that these bounds are sharp; see Appendix.) Notably, if the intersection of the bounds computed in each study is an empty set, that is evidence that at least one of the identifiability assumptions in at least one study is violated.

Before continuing, we wish to flag that the logic of the above statements, and the proof in the Appendix, rely on several subtle points that merit scrutiny in practice. For one, the computation of the bounds in each study as bounding the population average causal effect will rely on principles that have been described in the context of transportability (Dahabreh et al., 2020). Namely, we must have a well-defined population in mind, and specify why each of these studies are targeting an effect in that population. Most often, this will require some form of homogeneity assumption (Steele et al., 2020), as implicitly is required for interpretability of traditional fixed-effect meta analyses. We return to this and other likely challenges that would arise in common data settings in the discussion.

One could also consider pooling bounds under a relaxed set of assumptions that results in using set unions rather than set intersections. Suppose, for example, we wished to compute bounds under the assumption that at least one of the k studies' identifiability assumptions held, but we do not have evidence of which study. In that case, the average causal effect would lie in the union of the bounds from each study population, that is, $[min_s(LB_s), max_s(UB_s)]$ (Swanson, 2017). However, in many settings, it is difficult to imagine a bias that would invalidate at least one study without invalidating all included studies, particularly if the same identifiability assumptions are evoked for computing all study-specific bounds.

## 7.4 Pooling multiple bounds computed across studies under multiple sets of assumptions

A single study may present multiple opportunities to bound the same average causal effect under slightly different identifiability assumptions. For example,

in MR, researchers often propose multiple genetic variants as instruments to estimate the same exposure-outcome relationship. In this case, researchers could consider generating bounds separately for each genetic variant, under the assumption that each genetic variant was an individually valid instrument. Alternatively, under the assumption that multiple genetic variants were individually and jointly valid instruments, bounds could be calculated by proposing a set of genetic variants as a joint instrument (Chapter 6). This approach could be applied not only to the complete set of genetic variants proposed as instruments, but also to every possible subset of those genetic variants.

In this case, bounds can be pooled across study populations separately for each assumption set used to generate the bounds. In an MR study proposing multiple genetic variants as instruments, investigators can generate pooled bounds on $E(Y^{a=a} - Y^{a=a'})$ separately for each subset of genetic variants proposed as instruments. These pooled bounds can then be compared to one another to "triangulate" results and thus indirectly and directly evaluate the dependence of the results on the validity of the MR conditions for each genetic variant proposed as an instrument.

In addition, investigators can consider pooling bounds across different sets of assumptions. If two studies computed bounds on the same average causal effect using methods that relied on two different assumption sets, then we would expect the average causal effect to be within the intersection of those bounds under the combined (but study-specific) assumptions.

## 7.5   Application

*Data*

We computed pooled bounds on an average causal effect of maternal alcohol consumption during pregnancy on offspring ADHD, using results of our previous MR analysis conducted in the Avon Longitudinal Study of Parents and Children (ALSPAC) and the Norwegian Mother, Father, and Child Study (MoBa) (Chapter 6; Boyd et al., 2013; Fraser et al., 2013; Magnus et al., 2016). For an individual single nucleotide polymorphism (SNP), partial identification of the average causal effect is achieved if the SNP $Z$ is associated with the exposure $A$, the SNP $Z$ has no effect on the outcome $Y$ except through the exposure $A$ and individuals at different levels of the SNP $Z$ are exchangeable with regards to counterfactual outcome (Hernán and Robins, 2006). When a set of SNPs are proposed as joint instruments, these conditions must hold for the set

of SNPs individually and jointly. Importantly, conditions 2 and 3 are unverifiable. The previous study laid out several reasons why the MR conditions may not hold in this context, including selection on pregnancy, various forms of pleiotropy, assortative mating, and time-varying SNP-exposure relationships (Chapter 6).

We previously computed bounds under the MR model in each cohort separately, proposing 11 maternal SNPs (rs145452708, rs193099203, rs11940694, rs29001570, rs3114045, rs140280172, rs9841829, rs35081954, rs9991733, rs149127347, chr18:72124965) as instruments for the effect of any alcohol consumption during pregnancy on offspring ADHD. Within MoBa, rs145441283 was used as a proxy for rs193099203 and rs1154447 was used as a proxy for rs35081954. Because chr18:72124965 was unavailable in either cohort, rs201288331 was used as proxy in ALSPAC, and rs12955142 was used as a proxy in MoBa. The outcome was mother-reported ADHD symptoms in the clinical range in the offspring, measured using either the Development and Wellbeing Assessment or the Child Behavior Checklist Attention Deficit Hyperactivity subscale (Achenbach and Rescorla, 2000; Goodman et al., 2000). In the first model, the exposure $A=1$ if mothers reported any alcohol consumption during the second and third trimester of pregnancy, and $A=0$ if they did not report any alcohol consumption. In the second model, mothers who consumed more than 32 grams of alcohol per week (equivalent to approximately 2 cans of beer or glasses of wine) were removed from the analytic dataset. While this second question focuses more explicitly on the effects of light alcohol consumption on offspring ADHD, conditioning on the exposure in this way can result in selection bias (Swanson, Robins, et al., 2015).

*Statistical analyses conducted in the prior study*

Analyses in both cohorts were restricted to mother-child pairs without missing data on the exposure, outcome, or any of the proposed genetic instruments. Within ALSPAC, analyses were restricted to participants of self-reported white British ancestry. Because MoBa does not collect data on self-reported ancestry, we did not restrict the MoBa sample based on ancestry. However, only 5.6% of all MoBa participants report a first language other than Norwegian, suggesting the study population is primarily of Scandinavian ancestry (Magnus et al., 2006). These restrictions resulted in analytic samples of 4,457 mother-child pairs in ALSPAC and 6,216 mother-child pairs in MoBa. Prevalence of alcohol consumption and ADHD symptoms in both cohorts are shown in Table 6.1.

Prior to calculating bounds, we had eliminated combinations of SNPs proposed

as instruments for which the MR conditions were falsified (Bonet, 2001; Diemer et al., 2020; Pearl, 1995). For each set that was not falsified, the Richardson-Robins bounds were then calculated (Richardson and Robins, 2014).

*Statistical analysis for the pooled results*

To pool results, we assume the bounds computed within ALSPAC and MoBa identify the average causal effect in the population of western European mother-child pairs. Here, we assumed that any assumption set that was falsified in either cohort represented a structural violation of the MR conditions in the population of interest, and removed the set from further analysis.

Otherwise, for each subset of the SNPs proposed as instruments, we pooled bounds by taking the intersection of bounds calculated in ALSPAC and MoBa. Because we do not have an *a priori* reason to believe that a source of bias might exist that is completely unique to only one cohort, we do not present union bounds.

To evaluate the sensitivity of the results to potential residual population stratification, we also apply this method to models incorporating inverse-probability weights for 10 principal components. All analyses were conducted in R version 3.6.1 (Team, 2020).

## 7.6   Results

We first consider pooling bounds on the effect of any alcohol consumption compared to no alcohol consumption during pregnancy among European women, proposing single SNPs as instruments. As shown in Table 7.1, under the assumptions that the SNP in question is a valid instrument in both study populations, and that there is no effect modification by study population, the estimated bounds on the true average causal effect in the study population will be the intersection of the bounds calculated in each cohort. E.g., when rs11940694 is proposed as an instrument, bounds on the average causal effect were [.508, .431] in ALSPAC and [-.108, .878] in MoBa. The pooled bounds are therefore [-.108, .431]. Notably, because the instrumental inequalities failed to hold for 4 individual SNPs in MoBa, we have evidence that those SNPs are not valid instruments in at least one cohort (MoBa), and therefore do not meet the assumptions necessary for pooling. For every SNP proposed as an instrument individually, the pooled bounds were consistent with maternal alcohol

consumption slightly decreasing risk of offspring ADHD, having no effect, or increasing risk of offspring ADHD to a small or moderate degree.

Table 7.1: Pooled Bounds assuming bounds in all cohorts are valid, assuming single SNPs are individually valid instruments.

| Proposed Instruments | ALSPAC | | MoBa | | Pooled | | Assumptions |
|---|---|---|---|---|---|---|---|
| | Lower Bound | Upper Bound | Lower Bound | Upper Bound | Lower Bound | Upper Bound | |
| rs11940694 | -0.508 | 0.431 | -0.108 | 0.878 | -0.108 | 0.431 | 1. rs11940694 is a valid instrument in both studies 2. No effect modification by study population 3. Consistency |
| rs140280172 | -0.474 | 0.462 | NA | NA | NA | NA | |
| rs145452708 | -0.466 | 0.467 | -0.033 | 0.865 | -0.033 | 0.467 | 1. rs145452708 is a valid instrument in both studies 2. No effect modification by study population 3. Consistency |
| rs149127347 | -0.455 | 0.469 | NA | NA | NA | NA | |
| rs193099203 | -0.518 | 0.326 | NA | NA | NA | NA | |
| rs201288331 | -0.512 | 0.324 | NA | NA | NA | NA | |
| rs29001570 | -0.387 | 0.460 | -0.097 | 0.866 | -0.097 | 0.46 | 1. rs29001570 is a valid instrument in both studies 2. No effect modification by study population 3. Consistency |
| rs3114045 | -0.477 | 0.450 | -0.081 | 0.877 | -0.081 | 0.45 | 1. rs3114045 is a valid instrument in both studies 2. No effect modification by study population |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | 3. Consistency |
| rs35081954 | -0.520 | 0.449 | -0.104 | 0.884 | -0.104 | 0.449 | 1. rs35081954 is a valid instrument<br>2. No effect modification by study population<br>3. Consistency |
| rs9841829 | -0.520 | 0.417 | -0.07 | 0.868 | -0.07 | 0.417 | 1. rs9841829 is a valid instrument in both studies<br>2. No effect modification by study population<br>3. Consistency |
| rs9991733 | -0.522 | 0.407 | -0.103 | 0.877 | -0.103 | 0.407 | 1. rs9991733 is a valid instrument in both studies<br>2. No effect modification by study population<br>3. Consistency |

Table 7.2: Pooled bounds assuming bounds in all cohorts are valid, assuming multiple SNPs are individually and jointly valid instruments.

| Proposed Instruments | ALSPAC | | MoBa | | Pooled | | Assumptions |
|---|---|---|---|---|---|---|---|
| | Lower Bound | Upper Bound | Lower Bound | Upper Bound | Lower Bound | Upper Bound | |
| {rs9991733, rs9841829} | -0.443 | 0.306 | -0.041 | 0.841 | -0.041 | 0.306 | 1. rs9991733 & rs9841829 are valid instruments in both study populations<br>2. No effect modification by study population<br>3. Consistency |
| {rs9991733, rs350819544} | -0.497 | 0.329 | -0.082 | 0.861 | -0.082 | 0.329 | 1. rs9991733 & rs35081954 are valid instruments in both study populations<br>2. No effect modification by study population<br>3. Consistency |
| {rs35081954, rs9841829} | -0.478 | 0.369 | -0.065 | 0.843 | -0.065 | 0.369 | 1. rs35081954 & rs9841829 are valid instruments in both study populations<br>2. No effect modification by study population<br>3. Consistency |
| {rs145452708, rs29001570} | -0.386 | 0.458 | -0.082 | 0.865 | -0.082 | 0.458 | 1. rs145452708 & rs29001570 are valid instruments in both study populations<br>2. No effect modification by study population<br>3. Consistency |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| {rs11940694, rs9841829} | -0.483 | 0.362 | -0.061 | 0.843 | -0.061 | 0.362 | 1. rs11940694 & rs9841829 are valid instruments in both study populations<br>2. No effect modification by study population<br>3. Consistency |
| {rs11940694, rs35081954} | -0.464 | 0.376 | -0.073 | 0.851 | -0.073 | 0.376 | 1. rs11940694 & rs35081954 are valid instruments in both study populations<br>2. No effect modification by study population<br>3. Consistency |
| {rs11940694, rs3114045} | -0.151 | 0.396 | -0.067 | 0.840 | -0.067 | 0.396 | 1. rs11940694 & rs3114045 are valid instruments in both study populations<br>2. No effect modification by study population<br>3. Consistency |

When pooling the bounds computed in each cohort assuming multiple SNPs were valid instruments, the pooled bounds are slightly narrower than those generated proposing individual SNPs as instruments (Table 7.2). Overall, similar to bounds generated proposing single SNPs as instruments, the pooled bounds were consistent with maternal alcohol consumption slightly reducing risk of offspring ADHD, having no effect, or increasing risk of offspring ADHD to a small or moderate degree. The pooled bounds were generally similar when computing bounds for the effect of light alcohol consumption, and when weighting for 10 principal components (Figure 7.1).
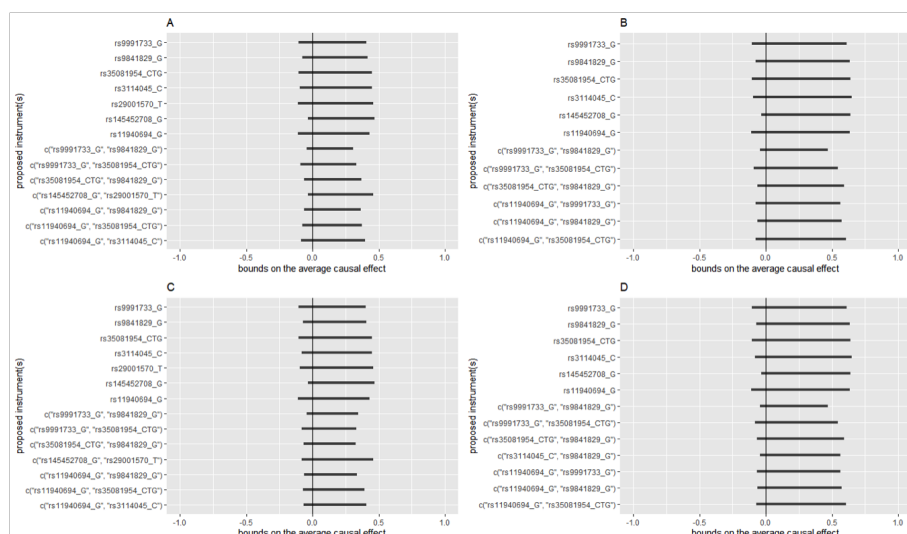


Figure 7.1: Pooled bounds on the average causal effect of alcohol consumption during pregnancy on offspring ADHD symptoms under different exposure definitions, and with and without inverse probability weighting to account for residual confounding. A and C show bounds on the average causal effect of any vs no alcohol consumption during pregnancy, in the unweighted study population (A) and a pseudo-population IP weighted for 10 principal components (C). B and D show bounds on the average causal effect of light alcohol consumption ($\leq 32$ grams/week) in the unweighted study population (B), and in a pseudo-population IP weighted for 10 principal components (D).

## 7.7    Discussion

Methods for combining bounds generated using MR in different studies have not been clearly established. Here, we demonstrate a straightforward approach for pooling MR bounds calculated in different cohorts with available individual level data, and clarify the assumptions necessary to perform such an analysis. Not only does this pooling procedure provide a method for synthesizing results from MR bounds analyses in multiple cohorts, it also will necessarily produce bounds that are equal in length or narrower than bounds computed in each cohort separately. In fact, because the narrowness of a set-intersection bound depends both on the size of the bounds being pooled and their position relative to one another, pooling theoretically can yield substantially narrower bounds even when the bounds from each study population are fairly wide (Manski, 2020).

As with any causal inference, it is critical to clearly define the population of interest. The ambiguity that results from an ill-defined population compound when we consider pooling data across studies (Dahabreh et al., 2020; Hernán and VanderWeele, 2011). It is common to imagine study populations as being drawn from an infinite super-population of individuals meeting particular eligibility criteria, and to aim to extend inferences to that infinite super-population. For the current application, one could argue that we are interested in the effect of alcohol consumption during pregnancy on offspring ADHD among all western European women who have become pregnant since the beginning of study recruitment, or will ever become pregnant in the future, a population that is effectively infinite. However, the idea that study populations were randomly sampled from such an infinite super-population is a fiction (Hernan and Robins, 2018; Robins, 1988). Within our application, each study population was restricted to a particular country, and, like all studies, were restricted to particular time periods. Beyond this, previous research has found that participants in cohort studies differ from non-participants in meaningful ways (Goldberg et al., 2001; Macera et al., 1990; Nilsen et al., 2009; Nohr et al., 2006).

How then, are we able to justify using these study-specific data and results to bound a population average causal effect? One answer is that these pooling methods implicitly require an assumption of no effect modification of the exposure-outcome relationship by study $S$ on the relevant scale (here, additive). In practice, even if each study population was not a true random sample of the super-population, this assumption would hold if $S$ was not related to either the outcome or any effect measure modifiers of the exposure-outcome relation-

ship. However, if the distribution of modifiers differed in different cohorts, this assumption would be violated. This is because the average causal effect in the super-population would be a weighted average of the effect within strata of the modifiers, with weights based on the distribution of modifiers in the super-population. Meanwhile, the pooled bounds we computed are based on the distribution of modifiers present within each cohort. If study populations differed in the distribution of an effect modifier from the super-population, then the true average causal effect would not necessarily lie within the intersection or union or study-specific bounds.

Unfortunately, this homogeneity assumption is implausible in many settings, including ours. Where individual level data on the proposed instrument, exposure, outcome, and potential effect modifiers are available in both the populations used to generate the bounds, and the distribution of potential effect modifiers is known for the target population of interest, it is possible that existing methods for transporting point estimates of average causal effects could be adapted to bounds of average causal effects in order to ameliorate this issue, e.g., by reweighting study populations to reflect the distribution of effect modifiers in a target population (Dahabreh et al., 2020). In the specific context of MR, this is substantially more complicated in practice compared to alternative analyses, as many plausible effect modifiers may be downstream of the SNP proposed as an instrument. In the case of our application, for example, the effect of alcohol consumption in pregnancy on offspring ADHD may be modified by the speed at which a woman metabolizes alcohol, or by offspring genotype. Moreover, since the existing transportability methods require an assumption of homogeneity conditional on covariates (Dahabreh et al., 2020; Steele et al., 2020), this may be further difficult to justify in the context of bounds computed under MR or other IV-based assumptions. An intrinsic motivation for the use of partial identification over point estimation in IV approaches is the desire to avoid strong, potentially implausible homogeneity assumptions (Swanson et al., 2018). It is therefore somewhat troublesome that, in order to pool bounds across study populations, we must make another homogeneity assumption.

Though this issue presents a specific complication to the use of pooled bounds, it also highlights a broader issue with the conduct and interpretation of MR studies, which are now frequently being used as evidence for policy interventions (Dixon et al., 2020; Harrison et al., 2020; Scholder et al., 2014). This includes the application here, in which previous MR studies on alcohol consumption during pregnancy have been cited in support of policy recommendations (Kenny and Hedges, 2018; to the UK Chief Medical Officers, 2016). Yet, as established, the study populations these effects are estimated in are not

necessarily selected randomly from the population in which the guidelines or policies are being given. This is all the more true when MR study populations are restricted to white European ancestry groups to avoid bias from population stratification (Davies et al., 2018). Extending inferences from these MR studies to a defined population then also requires homogeneity assumptions. Regardless of whether such a homogeneity assumption might truly hold, they are rarely, if ever, discussed.

There is also an issue of consistency in exposure definition that one needs to consider: pooling study-specific bounds on a population average causal effect is further complicated in practice by whether consistency can be reasonably assumed. Formally, these pooling methods require that if $A_i = a$ then $Y_i^a = Y_i$ for every individual $i$ in the target population and the included study populations. This implies that the exposure of interest must be the same across studies, and that study participation does not impact the outcome. Within observational studies, this may become especially problematic when a single binary exposure encompasses several versions of treatment, but the distribution of those treatment versions differs between study populations. In our primary example, we have grouped into two categories, never versus ever drinking during pregnancy. However, beyond possible issues of measurement error, it is likely that the amount of drinking, and not just the presence, during pregnancy affects ADHD symptom risk. If so, and individuals in each study population who consume alcohol differ in the amount of alcohol they consume, then there is relevant treatment variation (Hernán and VanderWeele, 2011), and the causal interpretation of bounds pooled across these study populations would be unclear. In practice, this may be an especially important consideration for MR studies that suggest they are targeting the "lifetime effect" of an exposure, rather than the 9 month gestation period our application focused on. In such MR designs, a single definition of the exposure could encompass many different exposure trajectories over the life-course, the distribution of which may differ between study populations.

We have focused primarily on identification, without discussing issues of statistical imprecision. However, bounds are impacted by the uncertainty created by sampling variation. While there is a growing literature on confidence interval estimation and statistical inference for bounds (Swanson et al., 2018; Tamer, 2010), currently there is no consensus on the best approach to accounting for this uncertainty in bounding approaches within studies, including the set intersection methods we describe. This is doubly important for MR studies that use the instrumental inequalities for falsification, as the instrumental inequalities are themselves partially identified parameters (Richardson and Robins, 2010).

Estimation would also be further complicated if the population of interest was in fact finite (Chan, 2017).

In our applied example, while bounds on the effects of prenatal alcohol exposure on ADHD did narrow, they did not identify a direction of effect. Readers might therefore question whether the many complications of pooling in practice are worthwhile, or how such pooled bounds could actually be integrated into decision-making. Importantly, bounds do not necessarily replace point identification strategies, but instead can be presented alongside point estimates. Indeed, to make recommendations about drinking behaviors and offspring ADHD risk based on these MR applications, we would need either to add further point-identifying assumptions or to use another causal inference approach. Yet bounds still have a vital place in such discourse. As has been extensively argued previously, bounds, especially wide bounds, can help show how strongly a particular analysis relies on assumptions (Cole et al., 2019; Robins and Greenland, 1996; Swanson et al., 2018; Swanson, Holme, et al., 2015) . Within individual MR studies with multiple SNPs proposed as instruments, computing bounds using different subsets of SNPs allows investigators to evaluate how results are affected by assumptions about both homogeneity and the validity of specific SNPs proposed as instruments. By quantifying the degree to which an analysis depends on such assumptions, bounding approaches can identify cases where potential violations of these assumptions should be more closely evaluated. Although pooled bounds are not directly comparable to fixed- or random- effects meta-analysis, incorporating pooled bounds into meta-analyses could similarly show how the conclusions of such an analysis might be impacted by heterogeneity of effects within studies. The use of such bounds also highlights the implicit assumption of homogeneity of effects and consistency across populations made whenever MR estimates are generalized to broader populations. By making these assumptions explicit, pooling approaches could help researchers and readers to identify areas in need of further investigation (e.g. evaluation of the extent to which effects of interest vary across populations of interest).

## 7.8   Conclusion

The use of these or related pooled-bounding methods in practice is complicated by issues of effect homogeneity, definitions of populations of interest, and consistency. While these issues pose a challenge to the use of pooling or meta-analytic methods, they also make clear the implicit assumptions made each time MR estimates are used to inform policy recommendations for larger

populations, or more generally, when any estimates from MR studies or other observational studies are being "triangulated". The presentation of both study-specific and pooled bounds across different assumption sets can help clarify the extent to which the conclusions of an analysis depend on the investigator's assumptions, rather than the data alone. Calculation of pooled bounds may also help investigators and readers to clarify the exact causal questions under study, and identify assumptions that should be further investigated.

# References

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the aseba preschool forms and profiles*. Burlington, VT: University of Vermont, Research center for children, youth, families.

Bonet, B. (2001). Instrumentality tests revisited. *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 48–55.

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., & Davey Smith, G. (2013). Cohort profile: The 'children of the 90s'—the index offspring of the avon longitudinal study of parents and children. *International journal of epidemiology*, *42*(1), 111–127.

Chan, W. (2017). Partially identified treatment effects for generalizability. *Journal of Research on Educational Effectiveness*, *10*(3), 646–669.

Cole, S. R., Hudgens, M. G., Edwards, J. K., Brookhart, M. A., Richardson, D. B., Westreich, D., & Adimora, A. A. (2019). Nonparametric bounds for the risk function. *American journal of epidemiology*, *188*(4), 632–636.

Dahabreh, I. J., Petito, L. C., Robertson, S. E., Hernán, M. A., & Steingrimsson, J. A. (2020). Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology*, *31*(3), 334–344.

Davies, N. M., Holmes, M. V., & Smith, G. D. (2018). Reading mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *Bmj*, *362*, k601.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177–188.

Diemer, E. W., Labrecque, J., Tiemeier, H., & Swanson, S. A. (2020). Application of the instrumental inequalities to a mendelian randomization study with multiple proposed instruments. *Epidemiology*, *31*(1), 65–74.

Dixon, P., Hollingworth, W., Harrison, S., Davies, N. M., & Smith, G. D. (2020). Mendelian randomization analysis of the causal effect of adiposity on hospital costs. *Journal of Health Economics*, *70*, 102300.

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., & Ness, A. (2013). Cohort profile: The avon longitudinal study of parents and children: Alspac mothers cohort. *International journal of epidemiology*, *42*(1), 97–110.

Goldberg, M., Chastang, J. F., Leclerc, A., Zins, M., Bonenfant, S., Bugel, I., Kaniewski, N., Schmaus, A., Niedhammer, I., & Piciotti, M. (2001). Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: A prospective study of the french gazel cohort and its target population. *American journal of epidemiology*, *154*(4), 373–384.

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The development and well-being assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of child psychology and psychiatry*, *41*(5), 645–655.

Harrison, S., Dixon, P., Jones, H. E., Davies, A. R., Howe, L. D., & Davies, N. M. (2020). Robust causal inference for long-term policy decisions: Cost effectiveness of interventions for obesity using mendelian randomization. *medRxiv*.

Hernan, M. A., & Robins, J. M. (2018). *Causal inference*. Chapman; Hall/CRC.

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 360–372.

Hernán, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology (Cambridge, Mass.)*, *22*(3), 368.

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 137–159.

Kenny, C., & Hedges, S. (2018). *Parental alcohol misuse and children* (tech. rep.). UK Parliament, Parliamentary Office of Science and Technology.

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International journal of technology assessment in health care*, *6*(1), 5–30.

Macera, C. A., Jackson, K. L., Davis, D. R., Kronenfeld, J. J., & Blair, S. N. (1990). Patterns of non-response to a mail survey. *Journal of clinical epidemiology*, *43*(12), 1427–1430.

Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Daltveit, A. K., Handal, M., Haugen, M., Høiseth, G., & Knudsen, G. P. (2016). Cohort profile update: The norwegian mother and child cohort study (moba). *International journal of epidemiology*, *45*(2), 382–388.

Magnus, P., Irgens, L. M., Haug, K., Nystad, W., Skjærven, R., & Stoltenberg, C. (2006). Cohort profile: The norwegian mother and child cohort study (moba). *International journal of epidemiology*, *35*(5), 1146–1150.

Manski, C. F. (2020). Toward credible patient-centered meta-analysis. *Epidemiology*, *31*(3), 345–352.

Nilsen, R. M., Vollset, S. E., Gjessing, H. K., Skjaerven, R., Melve, K. K., Schreuder, P., Alsaker, E. R., Haug, K., Daltveit, A. K., & Magnus, P. (2009). Self-selection and bias in a large prospective pregnancy cohort in norway. *Paediatric and perinatal epidemiology*, *23*(6), 597–608.

Nohr, E. A., Frydenberg, M., Henriksen, T. B., & Olsen, J. (2006). Does low participation in cohort studies induce bias? *Epidemiology*, 413–418.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 435–443.

Richardson, T. S., & Robins, J. M. (2010). Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 415–444.

Richardson, T. S., & Robins, J. M. (2014). Ace bounds; sems with equilibrium conditions. *Statistical Science*, *29*(3), 363–366.

Robins, J. M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine*, *7*(7), 773–785.

Robins, J. M., & Greenland, S. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, *91*(434), 456–458.

Scholder, S. V. K., Wehby, G. L., Lewis, S., & Zuccolo, L. (2014). Alcohol exposure in utero and child academic achievement. *Econ. J.*, *124*(576), 634–667.

Steele, R. J., Schnitzer, M. E., & Shrier, I. (2020). Importance of homogeneous effect modification for causal interpretation of meta-analyses. *Epidemiology*, *31*(3), 353–355.

Swanson, S. A. (2017). Commentary: Can we see the forest for the ivs? mendelian randomization studies with multiple genetic variants. *Epidemiology*, *28*(1), 43–46.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., & Richardson, T. S. (2018). Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, *113*(522), 933–947.

Swanson, S. A., Holme, Ø., Løberg, M., Kalager, M., Bretthauer, M., Hoff, G., Aas, E., & Hernán, M. A. (2015). Bounding the per-protocol effect in randomized trials: An application to colorectal cancer screening. *Trials*, *16*(1), 541.

Swanson, S. A., Robins, J. M., Miller, M., & Hernán, M. A. (2015). Selecting on treatment: A pervasive form of bias in instrumental variable analyses. *American journal of epidemiology*, *181*(3), 191–197.

Tamer, E. (2010). Partial identification in econometrics. *Annu. Rev. Econ.*, *2*(1), 167–195.

Team, R. C. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria.

to the UK Chief Medical Officers, G. D. G. (2016). Health risks from alcohol: New guidelines - list of supporting evidence.

# Chapter 8

# Discussion

The main aim of this dissertation was to explore strategies to improve the study of causal effects of maternal exposures on offspring outcomes in observational data. To achieve this, we investigated potential physiologic mechanisms by which exposures during pregnancy might affect offspring outcomes, and explored the use of MR to study effects of prenatal exposures and offspring outcomes. In this discussion, I review the main findings of this project, highlight key implications of these results, and propose future directions for further research in this area.

## 8.1   Principal Findings

In Chapter 2, we aimed to explore the effect of maternal mid-pregnancy vitamin D sufficiency on offspring DNA methylation in cord blood. In an analysis of data from 3,738 mother-child pairs across 7 European cohorts, we did not find evidence of associations between maternal mid-pregnancy vitamin D sufficiency and offspring DNA methylation at any measured CpG site. Under the assumptions of exchangeability, positivity, consistency, and no model misspecification, this suggests that either that maternal vitamin D sufficiency has a null effect on offspring cord blood methylation in these populations, or that the effects are too small to be detected with available sample sizes. The assumptions necessary for estimating causal effects using this study are somewhat

more reasonable in this study than in many previous EWAS, as maternal vitamin D status during pregnancy is unlikely to be affected by offspring DNA methylation at birth, meaning the study is less vulnerable to reverse causation. However, as discussed in detail with Chapter 2, it remains difficult to justify an assumption of no unmeasured confounding in EWAS. This is because the determinants of DNA methylation, and thus potential confounders, are not fully known, and may be affected by an individual's own tendency toward health promoting behaviors and perceived stress. In the specific context of null findings in EWAS, as in Chapter 2, considerations of confounding would also require grappling with faithfulness regarding a possibly perfect cancellation of bias and effect size.

How could possible confounding bias in EWAS be addressed? Some researchers have proposed incorporating MR into EWAS in order to avoid issues related to unmeasured confounding (Felix and Cecil, 2019; Relton and Davey Smith, 2012). The appeal of MR in this setting is understandable. Under the conditions laid out in the introduction, MR allows for unbiased estimation of average causal effects even in the presence of unmeasured confounding. For the study of prenatal exposures, where exposure-outcome confounders are often complex and difficult to measure, the MR conditions may seem like a more reasonable choice. Moreover, prenatal MR studies may be less vulnerable to bias due to time-varying SNP-exposure relationships than other MR designs, as offspring are only directly exposed to proposed maternal genetic instruments for 9 months, rather than their entire life course. However, prenatal MR is a unique context, and may be subject to unique biases not found in other MR designs.

To explore these biases, in Chapter 3, we conducted a systematic review of prenatal MR studies. There, we aimed to explore the nature and reporting of key strengths and weaknesses of the MR design in the context of pregnancy exposures and offspring outcomes. We found that researchers rarely discussed issues specific to the prenatal MR context, including selection on pregnancy, pleiotropy via offspring genotype, or pleiotropy via pre- or post-conceptional maternal exposure status. In addition, although the majority of prenatal MR studies in this review presented point estimates, the additional assumptions necessary for point estimation were rarely discussed. This is especially concerning in the context of prenatal MR, where certain estimands are defined differently, and there is biologic evidence that the point-identifying assumptions cannot hold for certain exposures of interest (see Chapter 1 and Chapter 6). In this review, falsification strategies were rarely applied, and no studies applied the instrumental inequalities to attempt to falsify their MR model.

As discussed in Chapter 4, although the primary MR assumptions cannot be verified, they can be falsified (Balke and Pearl, 1997; Pearl, 1995). Pearl showed that the instrumental conditions, and thus the MR conditions, imply a set of inequalities. If these inequalities do not hold, the MR model cannot hold in the dataset. Bonet subsequently proved that the instrumental variable model, and thus the MR model, actually imply additional inequalities beyond those described by Pearl (Bonet, 2001). However, neither the inequalities described by Balke & Pearl nor Bonet, known as the instrumental inequalities, had ever been applied to an MR study with multiple proposed instruments. In Chapter 4, we applied the instrumental inequalities to an MR study of the effect of maternal vitamin D sufficiency on offspring psychiatric symptoms, proposing 4 maternal SNPs as instruments. We found that, within our dataset, the MR assumptions were violated for at least half of the 4 SNPs proposed as instruments. This provides a clear example of how, in the context of multiple proposed instruments, the instrumental inequalities can be used to detect violations of the instrumental conditions. Further, simulations conducted as part of this study suggest that the inequalities will be increasingly violated as the magnitude of violations grow, and are more sensitive as larger numbers of variables are proposed as joint instruments.

The results of the instrumental inequalities are always specific to the research question and study population to which they were applied. After finding that the instrumental inequalities identified violations of the MR model for at least half of the 4 SNPs proposed as instruments in Chapter 4, we were interested in exploring the utility of the instrumental inequalities in detecting violations of the MR conditions in different research contexts. In Chapter 5, we therefore applied the instrumental inequalities to MR models proposing SNPs as instruments for the effects of 6 common exposures on cardiovascular disease in the UK Biobank. In that study, we detected no violations of the MR conditions when proposing each SNP as an instrument individually. However, when genetic risk scores were proposed as instruments, the instrumental inequalities detected violations of the MR conditions for 2 of the 6 exposures of interest.

Importantly, the instrumental inequalities are only able to falsify the 3 primary MR conditions. However, in order to obtain a point estimate, MR studies must also make one of a set of possible homogeneity conditions (Hernan and Robins, 2018). However, although historically unpopular, bounding approaches may provide information about the direction and plausible magnitudes of effect under the 3 primary MR conditions alone. In Chapter 6, we set out to investigate whether bounds can be usefully applied in the context of MR studies with multiple proposed instruments. We found that, when single SNPs were pro-

posed as instruments for the effect of maternal alcohol consumption during pregnancy on offspring ADHD, bounds on the average causal effect were wide, and barely improved on the assumption-free bounds. However, when larger numbers of SNPs were proposed as joint instruments, the bounds narrowed and were sometimes able to identify the direction of effect. In addition, the variation we observed across bounds calculated using different sets of proposed instruments in this study highlights how MR bounds can be used to compare and critically evaluate assumption sets used in MR. In our study, bounds computed using different sets of SNPs sometimes identified different directions of effect, an indicator of how strongly the conclusions of MR studies could change depending on the which genetic variants are assumed to be valid instruments.

In Chapter 6, we computed bounds using the same MR model in two separate study populations. Whenever results for the same research questions are available in multiple study populations, as in Chapter 6, it would be helpful to combine information across study populations. However, common meta-analytic techniques, such as fixed- and random-effects models, do not translate easily to the context of bounds. In Chapter 7, we describe an approach to pooling information from bounds calculated in different study populations, and apply it to the study results presented in Chapter 6. Using the same data and research question described in Chapter 6, we show that bounds computed in different study populations can be combined using the set-intersection methods described by Manski (Manski, 2020). However, the interpretation of these pooled bounds is complicated by the fact that these methods implicitly require an assumption of no effect modification by study population. The use of these bounds is further complicated by the fact that consistency (meaning that if $A_i = a$ then $Y_i^a = Y_i$ for every individual $i$ in the target population and the included study populations) may not truly hold across different study populations for our study question. This could occur if definitions of exposures differed across studies, or if categories of exposures contained multiple versions of treatment that varied in distribution across study populations. Beyond the potential utility of these pooled bounds for meta-analysis, this work highlights a broader issue with the interpretation of MR, that consistency and effect homogeneity across populations are implicitly assumed to hold any time MR estimates are used to conduct two-sample MR, to inform policy, or to triangulate effects in a broader population.

## 8.2 General Synthesis

### 8.2.1 Broader Implications

In Chapter 3, we found that prenatal MR is subject to a number of specific biases, which previous applications of this method have rarely discussed or attempted to mitigate. The effects of these biases on the MR literature may be severe. Indeed, the fact that the instrumental inequalities found violations of the MR conditions for several SNPs proposed as instruments for maternal vitamin D sufficiency and maternal alcohol consumption in multiple cohorts suggests that either the potential pleiotropic effects of these SNPs are stronger than was previously understood, or that the prenatal MR model is severely vulnerable to structural biases like selection on pregnancy, assortative mating, or pleiotropy via pre- or postnatal exposure status. This second possibility is a concerning indictment of prenatal MR in general, as authors of prenatal MR studies often describe these studies as superior to other causal inference designs (Allard et al., 2015; Alwan et al., 2012; Evans et al., 2019; Geng and Huang, 2018; Humphriss et al., 2013; Korevaar et al., 2014), and such studies are sometimes used as evidence to support policy recommendations (Dixon et al., 2020; Harrison et al., 2020; Kenny and Hedges, 2018; to the UK Chief Medical Officers, 2016). Of course, it is important to note that the results of the instrumental inequalities are always specific to the research question and study population to which they are applied. Therefore, the violations of the MR conditions detected in applications to maternal pregnancy vitamin D sufficiency and alcohol consumption (Chapters 4 & 6) are not necessarily a sign that all prenatal MR studies are biased, or even that all MR studies of these specific questions are biased. However, we detected violations of the MR conditions for multiple exposures and in multiple cohorts, including some violations when proposing SNPs as instruments individually, which generally indicates more severe violations of the MR conditions. This, alongside the smaller number of violations detected in adult analyses in UK Biobank (Chapter 5), suggests (but does not conclusively imply) that prenatal MR studies are more vulnerable to severe bias than was previously understood. Overall, our findings indicate that greater attention must be paid to the validity of the MR conditions in prenatal MR studies, especially for behavioral outcomes. Further research is needed to evaluate the impact of systematic biases specific to the prenatal MR context, including selection on pregnancy and assortative mating, on applied prenatal MR studies. It is also critical that researchers applying prenatal MR be made aware of the unique limitations of the prenatal MR design, and how to discuss or mitigate these potential biases in their studies.

Beyond the specific context of prenatal exposures, the studies presented here

suggest that both the instrumental inequalities and instrumental variable bounding approaches should be more broadly applied in instrumental variable studies, especially MR studies proposing multiple genetic variants as instruments. Historically, neither the instrumental inequalities nor instrumental variable bounds have been frequently applied in instrumental variable analyses. Indeed, to our knowledge, the study presented in Chapter 3 was the first time the instrumental inequalities had ever been applied to an MR study with multiple proposed instruments. As previously discussed, the unpopularity of the instrumental inequalities may be because previous work has suggested that, in the setting of dichotomous proposed instruments, only extreme violations of the MR conditions could be detected (Glymour et al., 2012). Similarly, bounding approaches may be unpopular because, in the case of binary proposed instruments, exposures, and outcomes, bounds on the average causal effect are often wide, and may not identify the direction of effect. Yet, the research presented here provides two clear lines of reasoning as to why both the instrumental inequalities and bounding approaches should be regularly applied in instrumental variable studies, especially in MR studies proposing multiple genetic variants as instruments. First, the studies presented in Chapters 4, 5, and 6 provide clear evidence that the instrumental inequalities are able to detect violations of the MR conditions when multiple SNPs are proposed as joint instruments, and are sometimes able to detect violations of the MR conditions when SNPs are proposed as instruments individually. Similarly, although the bounds computed proposing single SNPs as instruments in Chapter 6 were quite wide, bounds on the average causal effect narrowed substantially when multiple SNPs were proposed as joint instruments, and were sometimes able to identify directions of effect. These studies provide clear examples of cases when the instrumental inequalities and instrumental variable bounds provide meaningful information about an MR model, either by falsifying the model or by identifying a direction of effect without additional, potentially implausible assumptions.

Second, in the case of bounds, even if these approaches fail to identify a direction of effect, they still provide an opportunity to contextualize point estimates, both within individual studies (as shown in Chapter 6) and in meta-analyses (as shown in Chapter 7). To clarify this, let us first review the concept of bounding, which is described in more detail in Chapter 6. Put heuristically, studies generally combine data with strong assumptions in order to obtain a point estimate of a causal effect, a single number somewhere in the range of all possible values of the causal effect of interest. However, by combining data with weaker assumptions, investigators could instead obtain bounds, a range

of possible values of the causal effect of interest (rather than a single number). Importantly, bounds are distinct from confidence intervals, and indeed require their own confidence intervals. In contrast to confidence intervals, bounds will not typically collapse to a single point in an infinite dataset, as infinite data would only resolve the random error (attributable to sampling variability and stochastic counterfactuals), and would not allow us to observe all counterfactual outcomes for all participants. To explore the process of bounding, let us take as an example the research question presented in Chapter 6. There we were interested in estimating the average causal effect of binary maternal alcohol consumption on the binary presence of offspring ADHD symptoms, on the risk difference scale. Without any data, this average causal effect is already bounded by -1 (maternal alcohol consumption universally prevents offspring ADHD symptoms), and 1 (maternal alcohol consumption universally causes offspring ADHD symptoms). However, within a dataset, we already have information on the outcomes among individuals who were actually exposed and who were not actually exposed. Using the data and consistency alone, without any other assumptions, we can generate narrower bounds by imputing the missing counterfactual values for each individual to the most extreme possible outcome (e.g., assuming that everyone whose mother consumed alcohol during pregnancy would not have developed ADHD symptoms if their mother had not consumed alcohol during pregnancy, or that everyone whose mother did not consume alcohol during pregnancy would have developed ADHD symptoms if their mother did consume alcohol during pregnancy). These bounds, called the assumption-free bounds, will always have width 1, and will always include the null, meaning they cannot identify the direction of effect (Manski, 1990; Robins, 1989). Under the instrumental conditions, narrower bounds can be estimated, as described in Chapter 6. As we discuss in Chapter 6 and above, these narrow bounds can sometimes identify the direction of effect.

Specifically, because MR bounds rely on fewer assumptions than point estimation approaches in MR, and because they can be calculated for different assumption sets, bounds, even wide bounds, allow readers and investigators to understand how much information about a causal effect of interest is available in data alone, and how much relies on particular assumptions, either about the validity of a specific SNP proposed as an instrument, or about the heterogeneity of the effect across groups. In the results presented in Chapter 6, we can see that, when single SNPs are proposed as instruments, the bounds are very wide, and always cover the null. This would mean that conclusions from an MR analysis proposing one of those SNPs as an instrument would depend almost entirely on the homogeneity assumptions made by the investigators. When

larger numbers of SNPs were proposed as joint instruments, bounds from different sets of SNPs sometimes identified opposite directions of effect. Because the same average causal effect cannot be simultaneously negative and positive, this indicates at least one of the sets of proposed instruments are invalid, meaning the conclusions of such an analysis are also strongly dependent on the validity of each proposed genetic instruments. It is clear from both the studies presented here and previous research that bounds, especially wide bounds, have an inherent value for epidemiologic research. As Robins and Greenland put it, "wide bounds make clear the degree to which public health decisions are dependent on merging the data with strong prior beliefs" (Robins and Greenland, 1996). Causal inference always requires strong assumptions. But by presenting bounds alongside point estimates, and comparing bounds across multiple different assumption sets, investigators could evaluate how plausible causal effect sizes and directions change depending on the assumptions used. Incorporating bounds into a broader range of studies could therefore help to shift conversations around causal effects towards the question of what assumptions readers and investigators feel most confident in, and how strongly particular estimates and decisions rely on assumptions that feel suspect.

In an even more general sense, our findings argue for the importance of considering existing methods based on minimal assumptions, rather than trying to reinvent the wheel. Recent years have seen an explosion of sensitivity analyses and robust methods for MR (Bowden et al., 2015; Bowden et al., 2016; Cho et al., 2020; Hartwig et al., 2017; O'Connor and Price, 2018; Tchetgen et al., 2017; Verbanck et al., 2018; Zhu et al., 2018). Each of these methods was developed specifically to limit biases resulting from violations of the MR conditions, particularly the second and third MR conditions. Yet, almost all of these methods implicitly or explicitly require homogeneity conditions, even though biologic knowledge suggests this homogeneity is implausible for many exposures of interest and proposed instruments. Of course, MR estimates can be used as a test of the sharp null under the 3 primary MR conditions alone, and some researchers have argued that these alternative methods should be considered as tests of this sharp null. Yet, the results of such studies are typically interpreted as valid point estimates, rather than sharp null tests, and some estimators may even require homogeneity for sharp null testing. In contrast, neither the instrumental inequalities nor the instrumental variable bounding approach presented here require homogeneity assumptions. Despite this, and despite the fact that both methods were originally described well before MR became popular, neither has been meaningfully incorporated into the MR literature. Yet, our results show that these methods can be usefully applied to

MR studies. Beyond the implication that these specific methods should be applied to MR studies, these findings also suggest that it might be useful to more carefully consider the history of epidemiologic methods and existing epidemiologic methods when studying new research questions. There may be other methods based on minimal assumptions which have not been broadly accepted in epidemiologic research. Indeed, the concept of bounding in general has regularly been touted as a useful tool for epidemiologic research, but is rarely applied in practice (Cole et al., 2019; Robins and Greenland, 1996; Swanson and Hernán, 2013; Swanson et al., 2018). In addition, it may be that such methodologic advances have been ignored because they preceded the data and computing advances necessary for their productive use. Both the g-formula and g-estimation, for example, were limited in their initial applications because of the substantial computing power they required (Robins, 1986). The applications of the instrumental inequalities and bounds in MR shown in this dissertation would not have been possible prior to the current era of cheap and abundant genetic data. It is possible that there are many other methods that were simply ahead of their time, and could be productively applied with today's computing and data. While innovation is critical to science, our results suggest that older, less popular methods may still be useful, and should not be dropped by the wayside.

Our findings also suggest that the potential heterogeneity of effects of prenatal exposures is too often overlooked, especially in applications of MR. In general, we do not expect that the effects of an exposure will be exactly constant across an entire population. Indeed, that sort of constancy is not only biologically implausible, but also mathematically impossible in some cases (Hernan and Robins, 2018). In most cases, we expect that there will be some type of effect modification in measure present in a population. By effect modification in measure, we mean that for a given variable $V$, the average causal effect of $A$ on $Y$ varies across strata of $V$ on the relevant scale. For example, for the average causal effect of $A$ on $Y$, effect measure modification by $V$ on the additive scale is present if $\mathrm{E}(Y^{a=a} - Y^{a=a'}|V = v) \neq \mathrm{E}(Y^{a=a} - Y^{a=a'}|V = v')$. The existence of effect modification is not necessarily a problem in and of itself. After all, an average causal effect is an average, weighted accorded to the proportions of effect modifiers in the study population. However, as previously stated, in order to obtain a point estimate in MR, researchers must make one of a set of additional homogeneity assumptions. One of the most commonly used assumptions to estimate the average causal effect is that there is no effect modification of the average causal effect of $A$ on $Y$ by the proposed instrument $Z$ in any exposure category (Hernán and Robins, 2006). However, this

287

assumption is highly implausible for many prenatal exposures of interest. As described in Chapter 6, there is clear biologic evidence that this assumption will not hold when proposing alcohol dehydrogenase variants as instruments for the effect of maternal alcohol consumption of offspring outcomes. This is because offspring alcohol exposure during pregnancy depends both on the amount of alcohol consumed by the mother and the speed at which the mother is able to metabolize alcohol. Because alcohol metabolism is directly affected by alcohol dehydrogenase variants, the effect of pregnancy alcohol consumption on offspring will vary across levels of the maternal genetic variants proposed as instruments. This same violation of homogeneity conditions will likely occur whenever genetic variants related to the metabolism of a substance are proposed as instruments for the consumption of said substance during pregnancy. This homogeneity condition also poses a problem when non-metabolism related variants are proposed as instruments. Previous work has shown that, for a valid instrumental variable model, if any unmeasured confounder of the exposure $A$ and outcome $Y$ is also an effect modifier of the $A$-$Y$ relationship on the relevant scale, then the proposed instrument $Z$ will necessarily modify the effect of $A$ on $Y$ in some exposure category, violating the homogeneity condition described above (Hernán and Robins, 2006). It is likely that many key unmeasured confounders of maternal pregnancy exposures and offspring outcomes are also effect modifiers of that relationship. It is easy to imagine that, especially for offspring psychiatric outcomes, the effect of maternal pregnancy exposures are both confounded and modified by factors like family socio-economic status, parental mental health status, maternal health consciousness and interaction with the medical system, and substance use behaviors like smoking.

As an alternative to this assumption, researchers sometimes make an assumption of monotonicity. That is, the effect of the proposed instrument $Z$ on $A$ only works in one direction for every individual in the study population. Importantly, under monotonicity, one can only estimate the average causal effect within the compliers, the individuals for whom $A^{z=z} > A^{z=z'}$ for all $z > z'$. These compliers cannot actually be identified (though we can calculate the proportion of compliers in a study population) (Hernan and Robins, 2018; Swanson and Hernán, 2013). Because of the unidentifiable nature of compliers, and the fact that the nature and proportion of compliers is specific to the proposed instrument and population, several authors have argued that the complier average causal effect is not necessarily an especially useful measure (Deaton, 2010; Hernan and Robins, 2018). In addition, there are several challenges to the use of the monotonicity assumption for prenatal MR studies. First, while it is relatively easy to assume monotonicity holds in the context of a randomized

trial, the assumption is more difficult to justify in the context of observational studies, especially when the mechanisms by which a proposed genetic instrument affects the exposure of interest are unknown. Second, monotonicity is always specific to the particular genetic variant proposed as an instrument, and the populations of compliers for different genetic variants may not overlap at all. Thus, when multiple genetic variants are proposed as instruments, the estimated effect may actually be an average of effects within entirely disjoint subsets of the study population. Third, as discussed in Chapter 3, the definition of a complier in a prenatal MR study is slightly different than the definition in most MR designs. In most MR designs, the proposed instrument, exposure, and outcome are all measured within the same individual. However, for prenatal MR, the proposed instrument and exposure are measured in the mother, but the outcome is measured within the child. This means that, although the effect of interest is estimated within offspring, compliance status is based on the relationship between the mother's genetic variants and exposure status. Thus, for prenatal MR studies, monotonicity allows one to estimate an average causal effect among the subset of children whose mothers are compliers, even though the offspring themselves may not be compliers (as defined by their own genetic variants and exposure status).

The effects of violations of these homogeneity conditions are potentially serious. In Chapter 6, we found that bounds on the average causal effect were often wide, and covered the null, meaning the conclusions of an analysis of the same model using point estimation would depend almost entirely on the homogeneity conditions. In that study, we also found that point estimates for the average causal effect sometimes fell outside of the bounds, indicating a violation of the homogeneity conditions. Assuming the 3 primary MR conditions hold, this can occur when the bias resulting from heterogeneity of the effect is so large that it moves the point estimate outside of the valid bounds on the average causal effect. While violations of homogeneity were expected in the context of an MR study of maternal alcohol consumption, the fact that point estimates fell outside the bounds indicates that the bias resulting from these violations was severe. And yet, we found that prenatal MR studies almost never report their choice of point-identifying assumption. Only 4 of 22 studies reporting MR point estimates in Chapter 3 actually explicitly stated their point-identifying assumptions.

Beyond the challenges it presents to the internal validity of prenatal MR studies, thinking about effect heterogeneity is also critically important to interpreting results of MR studies, translating those results into recommended practice, and designing studies that answer the questions that are most important to us.

Causal effects can vary across populations, including those defined by location and by racial/ethnic ancestry. The vast majority of prenatal MR studies are conducted within white European ancestry populations located within North America or Western Europe (Chapter 3), and, given that the majority of GWAS studies have similar demographic characteristics (Haga, 2010), it is likely that the majority of MR studies more generally focus on this same population. Beyond data availability, the choice to limit study populations to white European ancestry participants is often a result of attempts to limit bias. MR requires that individuals at different levels of the proposed instrument are exchangeable with regards to counterfactual outcome, meaning that there cannot be any unmeasured common causes of the proposed instrument and outcome (Hernan and Robins, 2018; Hernán and Robins, 2006). However, different ancestry groups differ in their distribution of genetic variants, and may also differ in their distribution of outcomes for unrelated reasons. Thus, ancestry can be a confounder of the genetic variants proposed as instruments and the outcome, and thus can violate the MR conditions (VanderWeele et al., 2014). Researchers often attempt to mitigate this bias by restricting the study population based on self-reported race/ethnicity (in addition to the use of principal components for ancestry). Because of the demographics of the total study populations these analyses are often conducted in, white European ancestry individuals often make up the largest subgroup, and are therefore chosen for analysis.

However, as we show in Chapter 7, pooling bounds calculated in different study populations requires an assumption of no effect modification by study population on the relevant scale. This same assumption is required any time results from MR studies are generalized to a broader population. Yet, prenatal MR studies are used as evidence to support policy recommendations for diverse countries (Kenny and Hedges, 2018; to the UK Chief Medical Officers, 2016) and are interpreted as evidence of universal biologic effects (Murray et al., 2016; Zhang et al., 2015). This is concerning for multiple reasons. Although the presence of effect measure modification cannot be empirically verified (VanderWeele, 2012), evidence of effect modification has important implications for MR. Identifying populations where the effect varies, or is not present at all, can serve as a method of evaluating the MR assumptions (Glymour et al., 2012). Moreover, when MR is used as a means of exploring possible biologic relationships, ignoring effect modification could limit our ability to understand the mechanisms by which exposures impact outcomes. If effects are stronger or weaker, or not present at all, in specific subgroups, that could provide evidence about the relative importance of different mechanisms of action. Thinking about possible heterogeneity of effects is also vital when basing policy decisions

on MR. As previously mentioned, effects can easily vary across populations, and the same intervention could be ineffective or worse, actively harmful, if applied to a population that differs substantially from the study population the intervention was evaluated in. In addition to the lack of racial/ethnic diversity in many prenatal MR studies, the demographics of many influential cohorts differ substantially from the broader populations they were drawn from. For example, participants in the Avon Longitudinal Study of Parents and Children, based in Avon County, UK, were generally more likely to be married and of higher socioeconomic status than residents of the county or of the UK more broadly (Fraser et al., 2013). Assuming the effects estimated in MR studies of entirely white, socioeconomically advantaged individuals will be homogenous across the entire population of the UK seems difficult. It is not clear that interventions whose effects were estimated in prenatal MR studies within these types of cohorts would necessarily translate to broader populations. Further, in some cases, because of the potential for selecting on colliders affected by the proposed genetic instruments, this can result in biased estimates of the effect within the study population. But these issues are not insurmountable. Indeed, they are relatively easy to fix. First and foremost, we need to carefully consider whether the causal effects we are interested in might vary across populations of interest, or across potentially relevant variables. Second, we need to recruit, genotype, and maintain more geographically, racially, and socioeconomically diverse study populations for genetic epidemiologic research. Critically, this will also involve recruiting more diverse populations for the development of population reference panels for imputation of genetic data, and creating genomewide microarrays that provide adequate coverage in populations with different linkage disequilibrium patterns than European ancestry populations (Nelson et al., 2013; Peterson et al., 2019). Third, when making recommendations based on the results of MR studies, we need to explicitly define the target population to which such recommendations should be applied, and when necessary, apply methods for transporting effects from one population to another.

### 8.2.2  Future Directions

Overall, these findings have both substantive and methodologic implications, and suggest a need for further research in several areas. In particular, the results shown in this dissertation indicate a need for further applied research on the relationship between maternal genetic variants and offspring psychiatric outcomes, and how these relationships might bias MR estimates. Our findings also highlight a need for further methods development in several areas of MR.

Beyond these research directions, a critical aspect of future work in this area will be making both the instrumental inequalities and instrumental variable bounding approaches accessible to a broader audience.

### 8.2.2.1 Applied Genetic and Prenatal Epidemiology

The results of this dissertation suggest both that prenatal MR is subject to a number of unique biases that are rarely discussed within the prenatal MR literature, and that prenatal MR studies in general may be more vulnerable to structural biases than has been previously understood. Further research is needed to establish whether the assumption violations detected in the prenatal MR studies presented here affect other prenatal MR studies, and, more generally, how the forms of biases described in Chapter 3 impact prenatal MR studies in practice.

As previously mentioned, the instrumental inequalities cannot determine why the MR conditions are violated, only that they are violated. There are several reasons why the MR conditions might be violated in a particular prenatal MR study, some of which are detailed in Chapter 3. It is possible that the SNPs proposed as instruments in our analyses have a previously unknown pleiotropic relationship to offspring ADHD and ASD symptoms. For the case presented in Chapter 4, it is also possible that the MR conditions may be violated through pleiotropy via offspring genotype, if offspring vitamin D sufficiency also impacts their risk of ADHD or ASD symptoms. In the case of maternal alcohol consumption during pregnancy, offspring genotype is unlikely to be a source of bias, because alcohol dehydrogenase genes are not expressed in fetal life or early childhood (van Faassen and Niemelä, 2011). Future research might explore these potential issues through the application of existing methods for detection of these forms of bias (Bowden et al., 2015; Bowden et al., 2016; Cho et al., 2020; Evans et al., 2019; Hartwig et al., 2017; Tchetgen et al., 2017). However, it is important to note that these methods themselves rely on additional homogeneity assumptions, alongside other alternative conditions, which may not be met in all cases. This is especially true in the context of alcohol dehydrogenase-related SNPs proposed as instruments for the effect of maternal alcohol consumption during pregnancy, where both biologic knowledge and this dissertation suggest homogeneity assumptions commonly used for point estimation will not hold.

As we discuss in Chapter 3, the violations of the instrumental inequalities we detected in Chapters 4 and 6 could also result from variation in maternal

SNP-exposure relationships over time. Because we expect maternal genetic variants to impact maternal exposure status throughout the lifecourse, not only during pregnancy, such a situation could occur if maternal exposure status prior to pregnancy affected offspring outcomes (e.g. through mechanisms such as oocyte quality), or if maternal exposure status after pregnancy impacts offspring outcomes (e.g. through mechanisms like breast milk content, altered socio-economic status, or attachment style). Further, if the relationship between proposed genetic instruments and exposures vary over the course of pregnancy, MR estimates of the effect of prenatal exposure would be biased even if pre- or post-pregnancy exposure status did not affect the offspring. Future research could consider evaluating the potential impact of these issues by evaluating the extent to which relationships between maternal proposed genetic instruments and exposures of interest change over time, either within pregnancy or over the maternal life course. In addition, researchers could apply IV models that allow for time-varying exposure status (Robins, 1989; Tchetgen et al., 2017), though these methods come with their own assumptions, and as with the methods for evaluation of pleiotropy, require homogeneity assumptions.

It is also possible that the violations we observed in Chapters 4 and 6 were a result of assortative mating, a form of bias that can occur when parents in the study select mates based on particular traits, which can create a form of proposed instrument-outcome confounding. Essentially, if mating in the population is nonrandom, this could result in an open backdoor path between maternal genotype and offspring outcome via paternal genotype. Very little research has been conducted on the impact of assortative mating on prenatal MR, and existing prenatal MR studies rarely mention it. Out of the 43 studies evaluated in Chapter 3, only 3 mention potential bias due to assortative mating. Because assortative mating may affect all SNPs proposed as instruments, commonly used sensitivity analyses for prenatal MR, including MR-Egger and weighted median regression, may be unable to detect bias due to assortative mating. Assortative mating is a violation of Hardy-Weinberg equilibrium (HWE), and should theoretically be identified by the HWE tests typically conducted as part of quality control pipelines for genetic data. However, previous work has suggested that HWE tests are generally underpowered to detect violations of HWE, including assortative mating (Salanti et al., 2005). This is especially concerning for MR studies using point estimation, because weak instruments can amplify other biases. This could mean that even small violations of HWE could result in substantially biased MR point estimates. Further work is needed to establish the impact of assortative mating on prenatal MR studies in real data, and to evaluate possible mitigation strategies.

While some existing research has studied these biases in more general MR designs (Brumpton et al., 2020; Hartwig et al., 2018), it has not been evaluated in the specific context of prenatal MR.

One other understudied potential source of bias in our analyses is selection on pregnancy. It is only possible to measure exposures during pregnancy and offspring outcomes among women who actually become pregnant and carry the pregnancy to term. However, in prenatal MR studies, the genetic variants proposed as instruments are set at the mother's own conception. If either the genetic variant itself or pre-pregnancy exposure status impacts women's ability to become pregnant or carry to term, this could result in selection bias. Previous research has shown that several exposures of interest, including obesity, alcohol consumption, and smoking, have relatively strong associations with number of live births, number of pregnancies, and the presence of any pregnancy (Eggert et al., 2004; Weng et al., 2004; Wesselink et al., 2019). This suggests that selection on pregnancy may be a key form of bias for prenatal MR studies. This is especially concerning for two reasons. As with assortative mating, many existing methods to detect and limit bias in MR, such as MR-Egger and weighted median regression, would be unable to detect bias due to selection on pregnancy, because they rely on additional assumptions that would be violated by the presence of selection on pregnancy (Bowden et al., 2015; Bowden et al., 2016), which would generally result in an equal magnitude of bias across SNPs if the selection was due to differences in realized fertility across women with different exposure levels prior to pregnancy. Second, the majority of prenatal MR studies are conducted in study populations selected based on the presence of a pregnancy. This means that, although selection bias can often be resolved through the use of inverse probability of treatment weighting (Canan et al., 2017; Robins et al., 2000), inverse probability weighting cannot be applied in most prenatal MR studies, because no data is available to construct the weights. In this dissertation, we have suggested that researchers could consider sensitivity analyses using weights generated in external datasets to evaluate the potential impact of selection bias on their results, though this approach would require further assumptions about the transportability of weights across study populations. Despite the potential impact of this bias, almost no research has been conducted on selection on pregnancy, and out of the 43 prenatal MR studies evaluated in Chapter 3, only one mentioned selection on pregnancy. Further research is needed to evaluate the impact of selection on pregnancy in prenatal MR studies, to identify optimal bias reduction methods and sensitivity analyses, and to educate researchers and research consumers about the potential impact of this issue. Work on this topic would ideally include both

simulation studies and analyses in real data, in order to estimate the impact of this potential violation on real studies. However, evaluating selection on pregnancy in real data would require data on pregnancy exposures in women and offspring outcomes in samples that were not selected based on the presence of a pregnancy or interest in becoming pregnant, which may be difficult to obtain.

Even if we were able to obtain the type of dataset necessary to evaluate the potential impact of selection on pregnancy, evaluation of this bias also raises a bit of a philosophical question regarding the ideal choice of estimand. In some ways, not becoming pregnant can be viewed as a sort of competing event for the outcome of offspring ADHD symptoms, as women who do not have children necessarily cannot have children who develop ADHD symptoms. The inverse probability weighted estimate we propose may then be interpreted as an estimate of the effect of a joint intervention on pregnancy status and the exposure during pregnancy. In the simplest interpretation, we would be comparing the mean outcome had everyone in the population become pregnant and been exposed, relative to the mean outcome had everyone become pregnant but no one had been exposed. Formally, we would be estimating $E(Y^{a=a,p=1} - Y^{a=a',p=1})$, where $P$ denotes ever becoming pregnant. This estimand may not actually map well to an intervention of interest in the real world. Further research on the impact and nature of competing risks in prenatal MR could help to establish what causal quantities are of most interest, and the extent to which prenatal MR can estimate those quantities.

### 8.2.2.2 Methods Development

Throughout this dissertation, we have focused primarily on issues of identification. However, both instrumental variable bounds and the instrumental inequalities are impacted by uncertainty resulting from sampling variation. Because both the instrumental inequalities and bounding procedures described here involve minimum and maximum operations, the common techniques for estimating confidence intervals, such as the nonparametric bootstrap, will not generally produce valid confidence intervals for either procedure (Swanson et al., 2018). There is a growing body of literature surrounding the development of confidence intervals for partially identified parameters, but to this point, no single method has been determined to be preferable (Swanson et al., 2018; Tamer, 2010). Further research is needed to identify the best approach to estimating confidence intervals and developing statistical inference around partially bounded parameters, and to apply these methods to real data.

The expansion of research on this topic will be critical to expanded use of both the instrumental inequalities and instrumental variable bounds. One key use of bounds, which we have highlighted throughout this discussion, is their utility for policy decisions. As previously discussed, a key aspect of the use of MR for decision-making is careful consideration of what assumptions are required to generalize the results of a specific study to a broader population. In just the same way, it is important to recognize that the study populations in which we have calculated bounds are not infinite, and that it is critical to appropriately incorporate uncertainty into bounding. In addition, the development of confidence intervals for instrumental variable bounds would allow them to be more easily compared to point estimates, allowing researchers to more easily evaluate how dependent their results were on their point-identifying assumptions. The development of confidence intervals for the instrumental inequalities is similarly important, though their application may be slightly more complex. Even without confidence intervals, the instrumental inequalities can be used as a falsification test for the MR conditions within a specific sample. However, as previously stated, the instrumental inequalities cannot differentiate between random and structural violations of the MR conditions (see Chapter 4 for further details on this differentiation). Within a specific study, regardless of whether a violation is a result of structural or random biases, the result of a violation of the instrumental inequalities will be the same. The proposed instrument that violated the inequalities should not be used, and an alternative should be proposed. However, the development of confidence intervals for the instrumental inequalities could allow researchers to better differentiate between structural and random violations of the MR conditions, a delineation which could be useful in several ways. From a practical perspective, determining whether violations of the instrumental inequalities result from a structural violation would provide evidence for or against the use of the same MR model in other datasets. This could help to minimize time wasted on replication studies doomed to fail as a result of structurally invalid instruments (as opposed to random violations that would not necessarily occur in other datasets). In addition, differentiating between random and structural violations could help determine what cohort-specific results to include in meta-analyses. If violations of the MR conditions for a specific proposed instrument in a particular cohort were deemed to be random, results from that cohort should be eliminated from the analysis, but results for the same proposed instrument from other cohorts could still be meta-analyzed. In contrast, if violations of the instrumental inequalities were believed to result from structural biases, the invalid proposed instrument should be removed entirely from the meta-analysis, because the structural violation would impact results in all cohorts, even if the instrumental inequalities

did not falsify the model in other cohorts. Importantly, this strategy assumes that any identified bias is not due to study-specific selection bias, which would result in structural violations of the MR conditions, but would not necessarily cause bias in other study populations. If researchers believed selection bias was present, they could evaluate this assumption by conducting sensitivity analyses using inverse-probability weighting to account for selection bias conditional on the exposure and all variables that are believed to independently predict both selection into the study population and the outcome in each study sample (Hernan et al., 2004).

Developing confidence intervals for the instrumental inequalities could also help researchers to better understand potential pleiotropic effects of SNPs proposed as instruments, and thus the potential biologic mechanisms by which those SNPs affect various phenotypes. Under the assumption that individuals at different levels of the proposed instrument are exchangeable with regards to counterfactual outcome, meaning that there are no unmeasured confounders of the proposed instrument and outcome, and that there is no selection bias, the instrumental inequalities can actually be re-interpreted as bounds on the average controlled direct effect of the proposed instrument on the outcome when the exposure is held constant at a specific level (Cai et al., 2008). In the all binary case, for example, the instrumental inequalities can be interpreted as bounds on the effects $P(Y^{z=1,a=0} - Y^{z=0,a=0})$ and $P(Y^{z=1,a=1} - Y^{z=0,a=1})$, respectively. Although controlled direct effects cannot be used to identify indirect effects, controlled direct effect estimation has historically been of considerable interest to policy evaluation (VanderWeele and Vansteelandt, 2009). The use of the instrumental inequalities to bound controlled direct effects may therefore be of particular use to studies proposing policy instruments. Recent work has also shown that, if the instrumental inequalities fail to hold, the same proportions used to generate the instrumental inequalities can also be used to generate lower bounds on natural direct effects within each principal strata, even without additional monotonicity or cross-world assumptions (Zaidi and VanderWeele, 2020). Although we have previously discussed the various difficulties in the interpretation of principal strata effects when multiple SNPs are proposed as instruments, further exploration of these topics, especially the development of confidence intervals and statistical inference for these methods, could be a fruitful area of future exploration. Although the methods described by Zaidi and VanderWeele, 2020 and Cai et al., 2008 both have some limitations, they pose a useful way of thinking about how the instrumental inequalities could be used to identify genetic variants with pleiotropic effects, and about the extent to which such relationships can be used to evaluate possible

biologic pathways between genotypes and phenotypes. Further development of methods based on the instrumental inequalities may be especially useful when investigating relationships between traits that have not been evaluated in sufficiently large genome-wide association studies (GWAS), or where GWAS studies of the outcome are not available. In such cases, the lack of data would prevent researchers from using existing methods for detection of pleiotropy, such as linkage disequilibrium score regression (Bulik-Sullivan et al., 2015), that rely on the availability of results for both variables from existing GWAS. In these cases, the instrumental inequalities might be a useful way to evaluate possible pleiotropy, or at the very least to obtain useful information on potential pleiotropy from an MR study in which the model of interest has been falsified.

Another key area of future methods research is the development of methods for pooling bounds across multiple study populations when effects are heterogenous across study populations. As we discussed in Chapter 7, the use of set intersection methods for pooling bounds implicitly requires an assumption of no effect modification by study population (this assumption is distinct from the homogeneity assumptions required for point estimation). This is especially troublesome, because the average causal effect within a study population is a weighted average of the average causal effects within each strata of the effect modifiers, meaning that, if the distribution of effect modifiers of the exposure-outcome relationship differed across included study populations, this assumption would be violated. Essentially, in order to pool bounds across multiple study populations, one must assume that all included study populations were drawn randomly from the same theoretical super-population. This assumption is likely unreasonable in most contexts, and further research is needed to develop methods that do not require such strong assumptions. One possible direction for exploring this issue is expanding the transportability methods developed by Dahabreh et al. for point estimation of causal effects (Dahabreh et al., 2020). These methods use inverse probability of treatment weighting to allow for transport of estimates to a new target population, where they can subsequently be meta-analyzed. This approach requires an assumption of no effect modification conditional on the covariates included in the weights. If the weights included all measured effect modifiers, this would be a more reasonable choice of assumption. However, at present, such methods have only been developed in the context of point estimation, and have not been expanded to studies which calculate bounds on average causal effects. Moreover, in the setting of instrumental variables, the application of these methods may be especially complex if an effect modifier is downstream of the proposed instrument.

Broader use of the instrumental inequalities and instrumental variable bounds

in MR will also require the development of methods allowing for time-varying relationships between the proposed instrument and exposure. When the relationship between a proposed genetic instrument and an outcome changes over time, then MR estimates of the effect of the exposure will be biased (Labrecque and Swanson, 2019). Because genotypes are fixed at conception, and most exposures change over the lifetime, this is likely an issue for many MR studies. Prenatal MR is a relatively unique case, in that pregnancy is generally only 9 months long, meaning offspring are only directly exposed to maternal genotypes for those 9 months. We could therefore plausibly believe the relationship between maternal genetic variants and pregnancy exposures was constant through pregnancy, and that offspring are not affected by maternal pre- or post-pregnancy exposure status, meaning the effect would be isolated to pregnancy (under the assumption that other forms of bias, such as pleiotropy via offspring genotype or selection on pregnancy are not at play). As we discuss in Chapter 3, there are several cases where these assumptions will not necessarily hold (either because the gene-exposure relationships vary over the course of pregnancy, or because maternal exposure status before or after pregnancy continues to impact offspring outcomes), meaning prenatal MR is still vulnerable to bias from time-varying gene-exposure relationships. Nonetheless, the assumption of a constant gene-exposure relationship may be more plausible for prenatal MR studies than other MR designs. Valid use of instrumental variable bounding approaches in other MR designs will therefore require expansions allowing for time-varying associations between proposed instruments and exposures. For point estimates, it is possible to use extensions of structural mean models to estimate period or lifetime effects of exposures in instrumental variable studies with multiple measurements of the exposure, with parameters estimated using g-estimation (Robins, 1989). However, beyond the instrumental conditions, these methods still require potentially implausible homogeneity assumptions. To this point, methods for partial identification of this type of effect have not been established. Researchers might consider identifying bounds on time-varying effects either by considering how established methods for point identification of time-varying exposures might generalize to settings without effect homogeneity (Robins, 1989; Tchetgen et al., 2017), or by extending recent flexible methods for the identification of bounds to the context of a vector of time-varying exposures (Finkelstein and Shpitser, 2020; Poderini et al., 2020).

For MR models using a single exposure time point, the instrumental inequalities may in fact detect bias resulting from time-varying gene-exposure relationships. However, even bounds allowing for time-varying relationships between proposed instruments and exposures will still rely on the MR conditions. Thus it would

be helpful to develop extensions of the instrumental inequalities allowing for time-varying instrument-exposure relationships, in order to determine whether such models could be falsified. Because the instrumental inequalities are closely related to bounding approaches (Richardson and Robins, 2010), it would be logical to apply similar frameworks to extend both methods. Future research could therefore extend the instrumental inequalities to the setting of time-varying exposures using the same framework we previously discussed for bounds on causal effects of interest. One particular advantage of this approach is that it may allow for sharper inequality constraints than the Balke-Pearl inequalities.

Beyond time-varying SNP-exposure relationships, it would also be useful to explore the uses of further inequalities in MR studies with multiple proposed instruments, and to explore what can be inferred from patterns of violations across different combinations of SNPs proposed as instruments. Although the Balke-Pearl inequalities are sharp in the all binary setting (meaning they are the strongest possible constraints implied by the IV model, and no other constraint would be able to falsify an IV model that was not falsified by the Balke-Pearl inequalities), they are not sharp for proposed instruments with larger numbers of levels, or when multiple SNPs are proposed as joint instruments (Bonet, 2001; Richardson and Robins, 2010). This means that it may be possible to derive further constraints on the MR model when proposing larger numbers of genetic variants as instruments. Some previous research has described additional constraints on non-binary instrumental variable models, but some of these constraints have yet to be presented as implementable formula, and become increasingly computationally difficult in higher dimensional settings (Bonet, 2001; Evans, 2012). More recent research has established methods for identifying inequality constraints on models, including instrumental variable models, in more computationally simple ways (Finkelstein and Shpitser, 2020; Poderini et al., 2020), but these methods have not been applied to the context of MR with multiple proposed instruments. It is also interesting to note that the instrumental inequalities are themselves a special case of Bell's inequality (Pearl, 1995; Robins et al., 2015; Suppes, 1988), a concept that has been well explored within physics. It is possible that other modelling constraints identified and applied within physics, including entropic inequalities, might also be usefully applied to the setting of MR, or instrumental variable models more broadly.

The simulations we conducted in Chapter 4 suggest that, even when the instrumental inequalities hold for all SNPs proposed as instruments marginally, it may be possible to identify a pattern of violations of the instrumental inequalities across different subsets of the proposed instruments that are consistent

with the idea of a single "bad apple" – one SNP that responsible for all identified violations of the MR conditions. However, this approach has not yet been formalized, and further work is needed to clarify in what contexts this is actually possible, as well as what the benefits and limitations of such an approach are.

Along these same lines, it is also unclear whether it is possible to use the exact values of the instrumental inequalities to compare relative degrees of bias across models with different numbers of proposed joint instruments. As previously discussed, the instrumental inequalities can be considered bounds on the controlled direct effect of the proposed instrument on the outcome (Cai et al., 2008). In the all binary setting, it might be possible to compare values of the instrumental inequalities across different proposed instruments to determine which is the most biased. However, the Balke-Pearl inequalities are not necessarily sharp outside of the all binary case, and the sharpness of the bounds may differ when different numbers of SNPs are proposed as joint instruments (Richardson and Robins, 2010), which would complicate comparisons of the inequalities across sets. Further, the current lack of confidence intervals makes it difficult to compare values of the inequalities in settings where the number of individuals within strata of the proposed instrument might differ substantially. Take as an example the case presented in Chapter 4, in which 4 SNPs were proposed as instruments for the effect of a binary exposure. When each SNP was proposed as an instrument individually, each strata of the proposed instrument contained at least 10 individuals. However, when all 4 SNPs are proposed as joint instruments, 35 of the 81 strata of the proposed joint instrument contained less than 10 individuals. Several of the instrumental inequalities for latter case take the form $P(Y = y, A = a|Z = z) + P(Y = y', A = a|Z = z') \leq 1$. It is relatively easy to imagine a situation in which at least two strata of a proposed joint instrument with a large number of levels (e.g., a setting where a large number of SNPs are proposed as joint instruments) each contain only a single individual, or mother-child pair, as the case may be. If each of those two individuals reported particular values of the exposure and outcome, one could easily run into a situation where $P(Y = y, A = a|Z = z) = 1$ and $P(Y = y', A = a|Z = z') = 1$, meaning that the maximum value of the instrumental inequalities would be 2, the maximum possible value the inequalities could take when applied to a binary outcome. To be sure, this is still evidence of a violation of the MR conditions, and should be taken seriously. If the MR conditions were not violated, at least one of those strata would contain 0 individuals with that particular combination of the exposure and outcome. Yet, it is unclear that this violation, resulting from two people, would result in a

more biased estimate of a causal effect than a violation of the instrumental inequalities with a maximum value of 1.2 from a proposed joint instrument with hundreds of individuals within each strata of the proposed instruments. In the future, simulation studies may help to clarify the relationship of the magnitude of violations to the relative magnitude of bias in MR estimates when comparing proposed joint instruments incorporating different numbers of SNPs.

### 8.2.2.3 Teaching and Software Development for MR

The instrumental inequalities and instrumental variable bounding approaches were first described more than 20 years ago (Balke and Pearl, 1997; Manski, 1990; Pearl, 1995; Robins, 1989), though the specific bounding approach used in this dissertation was first published in 2014 (Richardson and Robins, 2014). Despite their age, and, in the case of bounds, several calls for their use in causal inference (Cole et al., 2019; Robins and Greenland, 1996; Swanson and Hernán, 2013), neither method has been broadly applied in MR, or instrumental variables analyses more generally. Indeed, to our knowledge, the studies included in Chapters 4, 5, and 6 are the first applications of the instrumental inequalities and instrumental variable bounds on the average causal effect to MR studies with multiple proposed instruments. In contrast, the MR-Egger method was first described in 2015 (Bowden et al., 2015), and the original paper has been cited more than 1200 times in the subsequent years. This rapid and wide uptake of MR-Egger is likely attributable both to the accessibility of simple tools for the implementation of MR-Egger, and the work of the original authors in expanding writing and teaching about MR-Egger and related methods. In the MR-Base package, MR-Egger regression can be run with a single line of code in R (Hemani et al., 2018). One key strength of this dissertation is our development and inclusion of adaptable R functions for the application and visualization of the instrumental inequalities and the Richardson-Robins bounds across larger numbers of proposed instruments. To our knowledge, prior to this dissertation, no software was available for the implementation of either method across combinations of proposed instruments. In order to increase the use of the instrumental inequalities and instrumental variable bounds in MR, we will need to improve the usability of this software, and to develop additional software for use of these methods in computing languages other than R. Further work is needed to improve the efficiency of the R functions we provide here, which will likely involve allowing for parallelization and potentially incorporating other methods to improve processing speed. In addition, it will be important in the future to incorporate these functions into an R package to

simplify the installation process for users, and to develop similar functions in other languages, including SAS, Stata, SPSS, and Python.

Hopefully, use of the instrumental inequalities and instrumental variable bounds will also expand as a result of further research into their properties and how these methods behave in real data. One reason researchers may be hesitant to use these methods is that they are distinct from other causal inference methods, and can be difficult to understand. The instrumental inequalities can be especially confusing to new users, because the reasoning that underlies them is not necessarily obvious, and they can seem like a bit of a black box. This opaqueness, alongside the fairly dense mathematical language used in papers describing the inequalities, could cause researchers to feel hesitant about using these methods. Ideally, the solution to this issue would be a mixture of explanatory articles and further methodological research on the instrumental inequalities and bounding approaches. In particular, simulation studies exploring the properties of the inequalities and bounds under different data-generating mechanisms, along with applications of the methods to real data, may help researchers and readers feel more confident in these types of analyses and their validity, even if their theoretical underpinning is difficult to understand.

## 8.3    Conclusions

While the use of the observational epidemiology can and should be focused on developing interventions to promote human health, it is inherently a risky endeavor, because it is based on unverifiable assumptions, and may give us the wrong answers. This is especially true in MR, because we have a limited understanding of the mechanisms at play. Unlike other causal inference methods, we also have a relatively weak conceptualization of how violations of the required assumptions translate into magnitudes of bias in estimates. Moreover, because of the relative makeup of genotyped cohorts and valid concerns regarding population stratification, MR studies are primarily conducted in white European ancestry populations, and it is unclear how such effects translate to other populations. To limit harm, it is critical that researchers incorporate a quality of humbleness into their work. Wider use of bounding approaches, falsification methods like the instrumental inequalities, and a clearer awareness of the impact of heterogeneity on analyses are key steps in this process. Adding falsification and bounding into standard MR would formalize an attitude of healthy skepticism, encouraging researchers to clarify the reliance of their conclusions

on the assumptions of their model and to refocus attentions away from what findings are simply eye-catching, and towards the plausibility of their models.

# References

Allard, C., Desgagné, V., Patenaude, J., Lacroix, M., Guillemette, L., Battista, M. C., Doyon, M., Ménard, J., Ardilouze, J. L., Perron, P., Bouchard, L., & Hivert, M. F. (2015). Mendelian randomization supports causality between maternal hyperglycemia and epigenetic regulation of leptin gene in newborns. *Epigenetics*, *10*(4), 342–351. https://doi.org/10.1080/15592294.2015.1029700

Alwan, N. A., Lawlor, D. A., McArdle, H. J., Greenwood, D. C., & Cade, J. E. (2012). Exploring the relationship between maternal iron status and offspring's blood pressure and adiposity: A mendelian randomization study. *Clin Epidemiol*, *4*(1), 193–200.

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176.

Bonet, B. (2001). Instrumentality tests revisited. *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 48–55.

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International journal of epidemiology*, *44*(2), 512–525.

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

Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. Å., Cho, Y., Howe, L. D., Hughes, A., & Boomsma, D. I. (2020). Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses. *Nature communications*, *11*(1), 1–13.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, *47*(3), 291–295.

Cai, Z., Kuroki, M., Pearl, J., & Tian, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, *64*(3), 695–701.

Canan, C., Lesko, C., & Lau, B. (2017). Instrumental variable analyses and selection bias. *Epidemiology (Cambridge, Mass.)*, *28*(3), 396.

Cho, Y., Haycock, P. C., Sanderson, E., Gaunt, T. R., Zheng, J., Morris, A. P., Smith, G. D., & Hemani, G. (2020). Exploiting horizontal pleiotropy to search for causal pathways within a mendelian randomization framework. *Nature communications*, *11*(1), 1–13.

Cole, S. R., Hudgens, M. G., Edwards, J. K., Brookhart, M. A., Richardson, D. B., Westreich, D., & Adimora, A. A. (2019). Nonparametric bounds for the risk function. *American journal of epidemiology*, *188*(4), 632–636.

Dahabreh, I. J., Petito, L. C., Robertson, S. E., Hernán, M. A., & Steingrimsson, J. A. (2020). Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology*, *31*(3), 334–344.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature*, *48*(2), 424–55.

Dixon, P., Hollingworth, W., Harrison, S., Davies, N. M., & Smith, G. D. (2020). Mendelian randomization analysis of the causal effect of adiposity on hospital costs. *Journal of Health Economics*, *70*, 102300.

Eggert, J., Theobald, H., & Engfeldt, P. (2004). Effects of alcohol consumption on female fertility during an 18-year period. *Fertility and sterility*, *81*(2), 379–383.

Evans, D. M., Moen, G. H., Hwang, L. D., Lawlor, D. A., & Warrington, N. M. (2019). Elucidating the role of maternal environmental exposures on offspring health and disease using two-sample mendelian randomization.

Evans, R. J. (2012). Graphical methods for inequality constraints in marginalized dags. *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 1–6.

Felix, J. F., & Cecil, C. A. M. (2019). Population dna methylation studies in the developmental origins of health and disease (dohad) framework. *Journal of developmental origins of health and disease*, *10*(3), 306–313.

Finkelstein, N., & Shpitser, I. (2020). Deriving bounds and inequality constraints using logical relations among counterfactuals. *Conference on Uncertainty in Artificial Intelligence*, 1348–1357.

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., & Ness, A. (2013). Cohort profile: The avon longitudinal study of parents and children: Alspac mothers cohort. *International journal of epidemiology*, *42*(1), 97–110.

Geng, T. T., & Huang, T. (2018). Maternal central obesity and birth size: A mendelian randomization analysis. *Lipids Health Dis*, *17*(1). https://doi.org/10.1186/s12944-018-0831-4

Glymour, M. M., Tchetgen Tchetgen, E. J., & Robins, J. M. (2012). Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American journal of epidemiology*, *175*(4), 332–339.

Haga, S. B. (2010). Impact of limited population diversity of genome-wide association studies. *Genetics in Medicine*, *12*(2), 81–84.

Harrison, S., Dixon, P., Jones, H. E., Davies, A. R., Howe, L. D., & Davies, N. M. (2020). Robust causal inference for long-term policy decisions: Cost effectiveness of interventions for obesity using mendelian randomization. *medRxiv*.

Hartwig, F. P., Davey Smith, G., & Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology*, *46*(6), 1985–1998.

Hartwig, F. P., Davies, N. M., & Davey Smith, G. (2018). Bias in mendelian randomization due to assortative mating. *Genetic epidemiology*, *42*(7), 608–620.

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., & Langdon, R. (2018). The mr-base platform supports systematic causal inference across the human phenome. *Elife*, *7*, e34408.

Hernan, M. A., Hernandez-Diaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 615–625.

Hernan, M. A., & Robins, J. M. (2018). *Causal inference.* Chapman; Hall/CRC.

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 360–372.

Humphriss, R., Hall, A., May, M., Zuccolo, L., & Macleod, J. (2013). Prenatal alcohol exposure and childhood balance ability: Findings from a uk birth cohort study. *BMJ Open*, *3*(6). https://doi.org/10.1136/bmjopen-2013-002718

Kenny, C., & Hedges, S. (2018). *Parental alcohol misuse and children* (tech. rep.). UK Parliament, Parliamentary Office of Science and Technology.

Korevaar, T. I. M., Steegers, E. A. P., Schalekamp-Timmermans, S., Ligthart, S., de Rijke, Y. B., Visser, W. E., Visser, W., de Muinck Keizer-Schrama, S. M. P. F., Hofman, A., & Hooijkaas, H. (2014). Soluble flt1 and placental growth factor are novel determinants of newborn thyroid (dys) function: The generation r study. *The Journal of Clinical Endocrinology and Metabolism*, *99*(9), E1627–E1634.

Labrecque, J. A., & Swanson, S. A. (2019). Interpretation and potential biases of mendelian randomization estimates with time-varying exposures. *American journal of epidemiology*, *188*(1), 231–238.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, *80*(2), 319–323.

Manski, C. F. (2020). Toward credible patient-centered meta-analysis. *Epidemiology*, *31*(3), 345–352.

Murray, J., Burgess, S., Zuccolo, L., Hickman, M., Gray, R., & Lewis, S. J. (2016). Moderate alcohol drinking in pregnancy increases risk for children's persistent conduct problems: Causal effects in a mendelian randomisation study. *Journal of child psychology and psychiatry*, *57*(5), 575–584.

Nelson, S. C., Doheny, K. F., Pugh, E. W., Romm, J. M., Ling, H., Laurie, C. A., Browning, S. R., Weir, B. S., & Laurie, C. C. (2013). Imputation-based genomic coverage assessments of current human genotyping arrays. *G3: Genes, Genomes, Genetics*, *3*(10), 1795–1807.

O'Connor, L. J., & Price, A. L. (2018). Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature genetics*, *50*(12), 1728–1734.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 435–443.

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C.-Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., & Brick, L. (2019). Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell*, *179*(3), 589–603.

Poderini, D., Chaves, R., Agresti, I., Carvacho, G., & Sciarrino, F. (2020). Exclusivity graph approach to instrumental inequalities. *Uncertainty in Artificial Intelligence*, 1274–1283.

Relton, C. L., & Davey Smith, G. (2012). Two-step epigenetic mendelian randomization: A strategy for establishing the causal role of epigenetic processes in pathways to disease. *International journal of epidemiology*, *41*(1), 161–176.

Richardson, T. S., & Robins, J. M. (2010). Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 415–444.

Richardson, T. S., & Robins, J. M. (2014). Ace bounds; sems with equilibrium conditions. *Statistical Science*, *29*(3), 363–366.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, *7*(9-12), 1393–1512.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

Robins, J. M., & Greenland, S. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, *91*(434), 456–458.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560.

Robins, J. M., VanderWeele, T. J., & Gill, R. D. (2015). A proof of bell's inequality in quantum mechanics using causal interactions. *Scandinavian Journal of Statistics*, *42*(2), 329–335.

Salanti, G., Amountza, G., Ntzani, E. E., & Ioannidis, J. P. A. (2005). Hardy–weinberg equilibrium in genetic association studies: An empirical evaluation of reporting, deviations, and power. *European journal of human genetics*, *13*(7), 840–848.

Suppes, P. (1988). Probabilistic causality in space and time. *Causation, chance and credence* (pp. 135–151). Springer.

Swanson, S. A., & Hernán, M. A. (2013). Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, *24*(3), 370–374.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., & Richardson, T. S. (2018). Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, *113*(522), 933–947.

Tamer, E. (2010). Partial identification in econometrics. *Annu. Rev. Econ.*, *2*(1), 167–195.

Tchetgen, E. J. T., Sun, B., & Walter, S. (2017). The genius approach to robust mendelian randomization inference. *arXiv preprint arXiv:1709.07779*.

to the UK Chief Medical Officers, G. D. G. (2016). Health risks from alcohol: New guidelines - list of supporting evidence.

VanderWeele, T. J. (2012). Confounding and effect modification: Distribution and measure. *Epidemiologic methods*, *1*(1), 55–82.

VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., & Kraft, P. (2014). Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.)*, *25*(3), 427.

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, *2*(4), 457–468.

van Faassen, E., & Niemelä, O. (2011). Biochemistry of prenatal alcohol exposure.

Verbanck, M., Chen, C.-Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nature genetics*, *50*(5), 693.

Weng, H. H., Bastian, L. A., Taylor Jr, D. H., Moser, B. K., & Ostbye, T. (2004). Number of children associated with obesity in middle-aged women and men: Results from the health and retirement study. *Journal of Women's Health*, *13*(1), 85–91.

Wesselink, A. K., Hatch, E. E., Rothman, K. J., Mikkelsen, E. M., Aschengrau, A., & Wise, L. A. (2019). Prospective study of cigarette smoking and fecundability. *Human Reproduction*, *34*(3), 558–567.

Zaidi, J. M., & VanderWeele, T. J. (2020). On the identification of individual level pleiotropic, pure direct, and principal stratum direct effects without cross world assumptions. *Scandinavian Journal of Statistics*.

Zhang, G., Bacelis, J., Lengyel, C., Teramo, K., Hallman, M., Helgeland, Ø., Johansson, S., Myhre, R., Sengpiel, V., Njølstad, P. å., Jacobsson, B., & Muglia, L. (2015). Assessing the causal relationship of maternal height on birth size and gestational age at birth: A mendelian randomization analysis. *PLoS Med*, *12*(8). https://doi.org/10.1371/journal.pmed.1001865

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., Robinson, M. R., McGrath, J. J., Visscher, P. M., & Wray, N. R. (2018). Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature communications*, *9*(1), 224.

**Chapter 9**

# Summary / Samenvatting

## 9.1 Summary

Previous work has suggested that the prenatal period is a potentially critical time for the development of psychiatric and behavioral outcomes in children. In particular, maternal substance use behaviors and micronutrient sufficiency may have an effect on offspring psychiatric health. However, conventional epidemiologic studies of such causal effects are vulnerable to bias from unmeasured confounding. The aim of this dissertation was therefore to explore how the validity of observational studies of the effect of maternal nutritional and substance use exposures on psychiatric symptoms in children could be improved. To do so, we investigated potential physiologic mechanisms by which prenatal exposures might impact psychiatric health in children, and explored the use of Mendelian randomization (MR) to study effects of pregnancy exposures on child outcomes.

In Chapter 2, we investigated the effect of maternal mid-pregnancy vitamin D insufficiency on offspring DNA methylation in cord blood in European ancestry mother-child pairs. In this study of 3,738 mother-child pairs across 7 European and North American cohorts, we did not find any evidence of associations between maternal vitamin D insufficiency in pregnancy and child cord blood DNA methylation at any measured site. Within this population, this suggests either that vitamin D insufficiency has no effect on offspring DNA methylation or that any such effects were too small to be detected in the current study.

The study in Chapter 2 relies on a strong assumption, that there was no unmeasured confounding of the exposure and outcome. One alternative to the outcome regression models used in Chapter 2 is MR. However, MR is vulnerable to certain unique biases, especially when applied to the prenatal context. To investigate the reporting of strengths and weaknesses of MR in the prenatal context, in Chapter 3 we conducted a systematic review of prenatal MR studies. We found that while researchers often reported sources of bias that affect all MR studies, they rarely discussed issues specific to the prenatal MR setting, including selection on pregnancy, pleiotropy via offspring genotype, and pleiotropy via pre- or post-conception exposure status. Although many studies reported point estimates, the additional assumptions required for point estimation were rarely mentioned.

In Chapter 4, we applied the instrumental inequalities, a falsification test for the instrumental variable model, to an MR study of the effect of maternal vitamin D insufficiency in pregnancy on child attention deficit-hyperactivity disorder (ADHD) and autism spectrum disorder symptoms. We found that, within our

dataset, the MR assumptions were violated for at least half of the 4 single nucleotide polymorphisms (SNPs) proposed as instruments. In simulations, we also found that the instrumental inequalities were more likely to detect violations of the MR assumptions when more genetic variants were proposed as instruments together, and when violations were larger.

To study whether the instrumental inequalities might be useful in detecting violations of the MR conditions in other settings, in Chapter 5 we applied the instrumental inequalities to MR models proposing SNPs as instruments for the effects of 6 exposures on cardiovascular disease in the UK Biobank. We did not detect any violations of the MR assumptions when proposing single SNPs as instruments. However, when proposing genetic risk scores as instruments, we found violations of the MR conditions for 2 of the 6 exposures.

In Chapter 6, we aimed to evaluate whether MR bounds, which do not require the additional assumptions needed for point estimation, could be usefully applied to MR studies. To do so, we computed bounds on the average causal effect of maternal alcohol consumption during pregnancy on child ADHD symptoms in two European cohorts. We found that, when proposing single SNPs as instruments, the bounds were very wide. When proposing multiple SNPs as instruments, the bounds narrowed and were sometimes able to identify a direction of effect.

In Chapter 7, we describe how bounds on the same causal effect in different study populations can be combined. We also apply this approach to bounds on the average causal effect of maternal alcohol consumption on offspring ADHD symptoms calculated in Chapter 6. All of the pooled bounds computed in this study covered the null. We discuss how bounds, even wide bounds, can be incorporated into scientific discourse.

Finally, in Chapter 8, we discuss overarching findings of this dissertation, the implications of those findings, and potential directions for future research on this topic.

## 9.2 Samenvatting

Eerdere onderzoeken suggereerden dat de prenatale periode van belang is bij de psychische- en gedragsontwikkeling van kinderen. Met name middelengebruik van de moeder tijdens de zwangerschap zou een effect kunnen hebben op de psychische gezondheid van het kind. Conventionele epidemiologische studies die dit soort causale effecten bestuderen zijn gevoelig voor ongemeten confounding. Om die reden was het doel van dit proefschrift om te onderzoeken hoe de validiteit van observationele onderzoeken die het effect bestuderen van voeding en middelengebruik van zwangere vrouwen op het voorkomen van psychiatrische symptomen bij kinderen, verbeterd kan worden. Om dit doel te bewerkstelligen onderzochten we potentiele fysiologische mechanismen waardoor blootstellingen tijdens de prenatale periode impact zouden kunnen hebben op de psychische gezondheid van kinderen en onderzochten we het gebruik van Mendeliaanse randomisatie (MR) als onderzoekmethode voor het bestuderen van de effecten van blootstelling aan risicofactoren tijdens de prenatale periode op uitkomsten in kinderen.

In hoofdstuk 2 onderzochten we het effect van vitamine D-insufficiëntie gedurende mid-pregnancy op veranderingen in DNA methylatie, gemeten in navelstrengbloed van Europese moeder-kind paren. Het onderzoek, waarin 3.738 moeder-kind paren uit 7 Europese en Noord-Amerikaanse cohorten werden bestudeerd, leverde geen bewijs op voor een associatie tussen maternale vitamine D-insufficiëntie tijdens de zwangerschap en veranderingen in DNA methylatie in navelstrengbloed. Dit suggereert dat, binnen deze onderzoekspopulatie, vitamine D geen effect heeft op veranderingen in DNA methylatie in het kind of dat deze effecten te klein zijn om gedetecteerd te kunnen worden.

Het onderzoek in hoofdstuk 2 berust op een sterke aanname, namelijk dat er geen ongemeten confounding is van de risicofactor en de uitkomstmaat. Een alternatief voor de regressiemodellen die in hoofdstuk 2 zijn gebruikt, is het toepassen van MR. MR is echter gevoelig voor unieke vormen van bias, vooral als deze wordt toegepast in de context van de prenatale periode. Om te bestuderen hoe de sterke en zwakke punten van MR in de context van de prenatale periode worden gerapporteerd in de literatuur, voerden we een systematische review uit waarbij we prenatale MR studies includeerden (hoofdstuk 3). We zagen dat onderzoekers vaak bronnen van bias rapporteerden die betrekking hadden op alle MR studies, maar dat er zelden discussie was over issues die specifiek zijn voor de prenatale MR setting, zoals het selecteren van deelnemers

op zwangerschap, pleiotropie via het genotype van het kind of pleiotropie door verschillen in pre- of postconceptionele blootstelling aan factoren. Hoewel veel onderzoeken puntschattingen rapporteren, werden de aanvullende aannames die zijn vereist bij het maken van puntschattingen zelden genoemd.

In hoofdstuk 4 pasten we de instrumentele ongelijkheden, een falsificatietest voor het instrumentele-variabele model, toe op een MR onderzoek naar het effect van maternale vitamine D-insufficiëntie. tijdens de zwangerschap op aandachtsdeficiëntie- en aandachtshyperactiviteitsstoornis (ADHD) en autisme spectrumstoornis-symptomen bij kinderen. We toonden aan dat, in onze dataset, de MR aannames waren geschonden bij ten minste de helft van de vier enkel-nucleotide polymorfismen (SNPs) die werden aangedragen als instrumenten. Uit simulaties bleek ook dat de instrumentele ongelijkheden eerder schendingen van MR aannames detecteren wanneer er meer genetische varianten gezamenlijk werden aangedragen als instrumenten en wanneer de schendingen grover waren.

Om te onderzoeken of de instrumentele ongelijkheden van pas komen om schendingen van aannames te detecteren in andere settings, pasten we in hoofdstuk 5 de instrumentele ongelijkheden toe op MR-modellen die SNPs aandroegen als instrumenten om de effecten van blootstellingen aan 6 factoren op hart- en vaatziekten in de UK Biobank te bestuderen. We vonden geen schendingen van de MR-aannames wanneer individuele SNPs werden aangedragen als instrumenten. Daarentegen vonden we bij 2 van de 6 factoren schendingen van de MR voorwaarden wanneer genetische risicoscores werden aangedragen als instrumenten.

In hoofdstuk 6 was ons doel om te evalueren of MR-bounds, welke geen aanvullende aannames vereisen die wel nodig zijn bij het maken van puntschattingen, toegepast konden worden op MR-onderzoek. Om dit te doen, berekenden we bounds van het gemiddelde causale effect van maternale alcoholconsumptie tijdens de zwangerschap op ADHD-symptomen bij kinderen in twee Europese cohorten. Ons onderzoek toonde aan dat wanneer individuele SNPs als instrumenten werden aangedragen, dit resulteerde in brede bounds. Wanneer meerdere SNPs werden aangedragen als instrumentele variabelen, waren de bounds smaller, zodanig dat er soms een richting van het effect aangetoond kon worden.

In hoofdstuk 7 beschrijven we hoe bounds van hetzelfde causale effect, verkregen uit verschillende studie populaties, gecombineerd kunnen worden. Tevens passen we deze methode toe op bounds van het gemiddelde causale effect van maternale alcoholconsumptie tijdens de zwangerschap op ADHD-symptomen

bij kinderen, berekend in hoofdstuk 6. Alle in dit onderzoek berekende gecombineerde bounds bevatten de waarde nul. We bespreken hoe bounds, zelfs brede bounds, geïncorporeerd kunnen worden in het wetenschappelijke discours.

Tot slot, in hoofdstuk 8, bediscussiëren we de overkoepelende bevindingen van dit proefschrift, de implicaties van deze bevindingen en potentiele richtingen voor toekomstig onderzoek over dit onderwerp.

# Appendix

# Curriculum Vitae

**Elizabeth Wicker Diemer**

Phone: (408)-564-9121
Email: ewdiemer@gmail.com

---

Phone: (408)-564-9121
Email: ewdiemer@gmail.com

EDUCATION

Erasmus University Medical Center, Rotterdam, The Netherlands
Doctor of Philosophy, Epidemiology, Expected 2021
Advisors: Sonja Swanson; Henning Tiemeier

Harvard T.H. Chan School of Public Health, Boston, Massachusetts
Master of Science, Epidemiology, May 2017
Advisors: Alison Field, Kathryn Terry

Washington University in St. Louis, St. Louis, Missouri
Bachelor of Arts, Major: Psychology, Minor: Public Health, May 2014

RELEVANT EXPERIENCE

Summer Research Fellow, University of North Carolina at Chapel Hill, Center of Excellence for Eating Disorders, June 2016-August 2016 – Supervisor: Jessica Baker, PhD

Epidemiology Data Analyst Intern, The Fenway Institute, Boston, MA, November 2015-June 2016. Supervisor: Julia Coffey-Esquivel, MPH

Research Assistant, Washington University, Department of Psychiatry, St. Louis, MO, June 2012 to July 2014. Supervisor: Alexis Duncan, PhD, MPH

HONORS AND AWARDS

| | |
|---|---|
| Dean's List | January 2011 to January 2014 |
| College Honors | May 2014 |
| John U Monro Fellowship | September 2015 |
| John Bruce Nichols and Margaret L | |

GRANTS

Academy Ter Meulen Grant                    €3,900                    2018-2019

PEER-REVIEWED ARTICLES

1. **Diemer EW**, Neumann A, Labrecque JA, Tiemeier HW, Swanson SA (2021). Mendelian randomization approaches to the study of prenatal exposures: a systematic review. *Paediatric and Perinatal Epidemiology* .35(1):130-142.

2. Wonuola AA, Hammerschlag AR, Shahid E, Allegrini AG, Karhunen V, Sallis HM, Ask H, Askeland RB, Baselmans BML, **Diemer EW**, Hagenbeek FA, Havdahl A, Hottenga JJ, Mbarek H, Tesli M, van Beijsterveldt CEM, PGC Bipolar Disorder Working Group, PGC Schizophrenia Working Group, Breen G, Lewis C, O'Donovan M, Thapar A, Boomsma DI, Kuja-Halkola R, Reichborn-Kjennerud T, Magnus P, Rimfeld K, Ystrom E, Jarvelin MR, Lichtenstein P, Lundstrom S, Munafo MR, Plomin R, Tiemeier H, Nivard MG, Bartels M, Middeldorp CM (2020). Genetic associations between childhood psychopathology and adult depression and associated traits in 42,998 individuals: a meta-analysis. *JAMA Psychiatry.* 77(7):715-728.

3. **Diemer EW**, Labrecque JA, Tiemeier H, Swanson SA (2020). Application of the instrumental inequalities to a Mendelian randomization study with multiple proposed instruments. *Epidemiology.* 31(1):65-74.

4. Watson HJ, **Diemer EW**, Zerwas S, Gustavson K, Knudsen GP, Torgersen L, Reichborn-Kjennerud T, Bulik CM (2019). Prenatal and perinatal risk factors for eating disorders in women: A population cohort study.*Int J Eat Disord.* 52(6):643-651.

5. **Diemer EW**, Reisner SL, White-Hughto JM, Gordon AR, Austin SB (2018). Beyond the Binary: Differences in eating disorder prevalence by gender identity in a transgender sample. *Transgend Health.* 3(1):17-23.

6. Watson HJ, Zerwas S, Torgersen L, Gustavson K, **Diemer EW**, Knudsen GP, Reichborn-Kjennerud T, Bulik CM (2017). Maternal eating disorders and perinatal outcomes: A three-generation study in the Norwegian Mother and Child Cohort Study. *Journal of Abnormal Psychology*, 126(5), 552-564. PMID: 28691845

7. **Diemer EW**, Grant JD, Munn-Chernoff MA, Patterson D, Duncan AE (2015). Gender identity, sexual orientation, and eating-related pathology in a national sample of college students. *Journal of Adolescent Health*, 57, 144-149. PMID: 25937471. PMCID: PMC4545276.

8. Duncan AE, Sartor CE, Jonson-Reid M, Munn-Chernoff MA, Eschenbacher MA, **Diemer EW**, Nelson EC, Waldron M, Bucholz KK, Madden PAF, Heath AC (2015). Associations between body mass index, posttraumatic stress disorder, and child maltreatment in young women. *Child Abuse & Neglect*, 45, 154-162. PMID: 25770346. PMCID: PMC4470860

MANUSCRIPTS IN PREPARATION AND UNDER REVIEW

1. **Diemer EW**, Havdahl A, Andreassen OA, Munafo MR, Njolstad PR, Tiemeier H, Zuccolo L, Swanson SA (revise and resubmit). Bounding the average causal effect in Mendelian randomization studies with multiple proposed instruments: an application to prenatal alcohol exposure and attention deficit hyperactivity disorder.*American Journal of Epidemiology*

2. **Diemer EW**, Swanson SA, Eddy KT, Chavarro J, Field AE (in preparation). Atypical anorexia nervosa and "significant weight loss" in a large longitudinal sample of adolescents and young adults.

3. **Diemer EW**, Neumann A, Heinonen K, Tuhkanen JM, Yeung E, Suderman M, Briggs C, Hivert MF, Felix JF, Tiemeier H (in preparation). Maternal mid-pregnancy vitamin D sufficiency and offspring cord blood DNA methylation: Epigenome-wide consortium meta-analysis.

CONFERENCE PRESENTATIONS

1. Bounds on the average causal effect of prenatal alcohol exposure on ADHD – a Mendelian randomization study. *Society for Epidemiologic Research Annual Meeting*. December 2020. Virtual. (poster)

2. Application of the instrumental inequalities to a Mendelian randomization study with multiple proposed instruments. *Causality Reading*

*Group.* February 2020. Leiden University Medical Center, Leiden, Netherlands.

3. A genetic and epigenetic approach to studying the effects of lifestyle behaviors during pregnancy on offspring outcomes. *Generation R Research Meeting.* November 2019. Erasmus MC, Rotterdam, Netherlands.

4. What went wrong? Planning a simulation study to evaluate the sensitivity of the instrumental inequalities to specific bias structures. *Center for Quantitative Methods.* October 2019. Erasmus MC, Rotterdam, Netherlands.

5. Application of the instrumental inequalities to a Mendelian randomization study with multiple proposed instruments. *Society for Epidemiologic Research Annual Meeting.* June 2019. Minneapolis, Minnesota. (poster)

6. Application of the instrumental inequalities to a Mendelian randomization study with multiple proposed instruments. *European Causal Inference Meeting.* March 2019. Bremen, Germany.

7. Bounds on the average causal effect of prenatal alcohol exposure on ADHD – a Mendelian randomization study. *MRC IEU Research Meeting.* March 2019. MRC Integrative Epidemiology Unit at University of Bristol, Bristol, UK.

8. Application of the instrumental inequalities to a Mendelian randomization study with multiple proposed instruments. *Masterclass on Mendelian randomization and health inequalities.* November 2018. Erasmus MC, Rotterdam, Netherlands

9. Application of the instrumental inequalities to a Mendelian randomization study with multiple proposed instruments. *Center for Quantitative Methods*, October 2018. Erasmus MC, Rotterdam, Netherlands.

10. Gambling, Purging Behaviors, and Eating Disorder Diagnosis among US College Students. *Annual Meeting of the Eating Disorders Research Society*, October 2014, San Diego, California. (poster)

11. Intimate Partner Violence and Eating Disorders among College Students: Moderation by Gender. *International Conference on Eating Disorders*, March 2014, New York City, New York. (poster)

12. Relations among Gender Identity, Sexual Orientation, and Eating Disorders in College Students. *International Conference on Eating Disorders*, March 2014, New York City, New York. (poster)

13. Inappropriate Compensatory Behaviors among African- and European-American Male Adolescents and Young Adults. *International Conference on Eating Disorders*, May 2013. Montreal, Quebec, Canada. (poster)

TEACHING ACTIVITIES

**Confounding Control: A Component of Causal Inference**   2018-2021

Harvard T.H. Chan School of Public Health
Teaching Assistant

This course is part of the blended online/on-campus MPH-EPI program. As a TA for this course, I lead office hours and discussion sessions, monitor online discussion boards, grade class assignments, and assist in updating class materials based on student performance and feedback.

**Principles in Causal Inference**                           2018-2020

Netherlands Institute for Health Sciences
Teaching Assistant

This course is part of the first year required courses for masters students. As a TA for this course, I grade class assignments and provide support for students on class projects.

EDITORIAL ACTIVITIES

Ad hoc Reviewer
- International Journal of Epidemiology
- Lancet Psychiatry
- Journal of Adolescent Health
- European Journal of Epidemiology
- Epidemiology

# PhD Portfolio

**Date:** Monday, 1 February 2021
**Name PhD student:** Elizabeth W. Diemer
**Erasmus MC Department:** Child and Adolescent Psychiatry
**Research School:** NIHES
**PhD period:** August 2017 - February 2021
**Promotor:** Henning Tiemeier
**Copromotor:** Sonja Swanson

## 1. PhD training

| MSc-program Genetic Epidemiology, NIHES: | Year | Grade | ECTS |
|---|---|---|---|
| Principles of Research in Medicine and Epidemiology | | XMT | 0.7 |
| Introduction to Medical Writing | 2020 | AP | 2 |
| Principles of Genetic Epidemiology | 2017 | AP | 0.7 |
| Genomics in Molecular Medicine | 2017 | AP | 1.4 |
| Advances in Genomics Research | 2017 | AP | 0.4 |
| Genome-Wide Association Analysis | 2017 | AP | 1.4 |
| Study Design | | XMT | 4.3 |
| Genetic-Epidemiologic Research Methods | 2018 | 7.4 | 5.1 |
| Linux for Scientists | 2018 | AP | 0.6 |
| SNP's and Human Diseases | 2017 | AP | 1.4 |
| Biostatistical Methods I: Basic Principles | | XMT | 5.7 |
| Biostatistical Methods II: Classical Regression Models | | XMT | 4.3 |
| Advances in GWAS | | XMT | 1.4 |
| Family-based Genetic Analysis | | XMT | 1,4 |
| An Introduction to the Analysis of Next-Generation Sequencing Data | 2020 | 7.9 | 1.4 |
| **Elective courses, NIHES:** | | | |
| History of Epidemiologic Ideas | 2018 | AP | 0.7 |
| Markers and Prediction Research | 2018 | AP | 0.7 |
| Mendelian Randomization | 2019 | 8.7 | 0.9 |
| Child Psychiatric Epidemiology | 2019 | AP | 0.9 |
| CAPICE seminar on drug development and target validation | 2019 | AP | 1 |
| CAPICE seminar on latest developments in statistical genetics | 2020 | AP | 0.7 |

**Symposia, Conferences, and Workshops**

| Event | Year | ECTS |
|---|---|---|
| CAPICE workshop on genetics and developmental psychology, Kings College London | 2018 | 1.0 |
| CAPICE workshop on statistical analysis of genome wide association studies, Imperial College London | 2018 | 1.0 |
| CAPICE workshop on longitudinal and multivariate modelling, Universiteit Twente | 2019 | 1.0 |
| MRC IEU Research Meeting, University of Bristol | 2019 | 0.2 |
| European Causal Inference Meeting, Bremen (oral presentation) | 2019 | 0.2 |
| CAPICE workshop on Mendelian randomization, University of Bristol | 2019 | 1.0 |
| Society for Epidemiologic Research Annual Meeting, Minneapolis, Minnesota (poster presentation) | 2019 | 0.2 |
| Causality Reading Group, Leiden University MC | 2020 | 0.2 |
| CAPICE workshop on career development in academia and industry (virtual) | 2020 | 1.0 |
| Center for Quantitative Methods Meetings, Erasmus MC | 2018-2021 | 0.2 |
| Generation R Research Meetings, Erasmus MC | 2018-2020 | 0.2 |

**2. General Scientific Activities**

| Activity | Year | ECTS |
|---|---|---|
| Teaching assistant for course: Principles of Causal Inference EP01), NIHES | 2018-2020 | 0.5 |
| Teaching assistant for course: Confounding Control: A Component of Causal Inference (EPI524), Harvard T.H. Chan School of Public Health | 2018-2021 | 6.0 |
| Generation R General Tasks | 2017-2020 | |
| Peer Review (Epidemiology, International Journal of Epidemiology, Journal of Adolescent Health, European Journal of Epidemiology, Lancet Psychiatry) | 2018-2021 | 3.0 |

# Acknowledgements

I could fill a much larger book than this with acknowledgements, and it still wouldn't be an adequate explanation of how much this dissertation owes to the people around me, or how blessed I feel to have those people in my life. Truly, this PhD is as much theirs as it is mine – I would have never even been in a position to start these projects, let alone finish this dissertation, without the help of all the wonderful friends, family, mentors, and colleagues that surround me.

First, thank you to my supervisors. Henning, thank you for your support, and for allowing me the freedom to explore my own interests throughout my PhD.

Sonja, thank so much for going on this journey with me. Thank you for hiring me, guiding me to exciting research questions, pushing me to explore new topics, patiently wading through proofs with me, reaching out to experts for me when I wasn't brave enough to ask alone, and always making sure our work focused not only on what is methodologically correct but also what is most important to the people our research aims to help. I feel privileged that, through working with you as a student and as a teaching assistant, I've gotten to watch you model what it is to be a good epidemiologist and a good teacher of epidemiology. Especially in the times we live in, I couldn't have asked for a better mentor.

I would also like to thank my committee. Arfan, Maria, Niels, Janine, Eric, Vanessa; thank you for taking the time to read this not-particularly short thesis, and for all your questions and suggestions, which have certainly improved the work.

Thank you to all the members of the causal inference group, the department of child and adolescent psychiatry, the department of epidemiology, and the department of biostatistics, both long term and temporary, but most especially to Kelly, Paloma, and Jeremy. Kelly, thank you for letting me be a part of your masters thesis, and for your excellent translations. This thesis quite literally would not be complete without you, and I feel honored to be a part of your growth as a researcher. Jeremy, thank you for pushing me to clarify and sharpen my thinking on both epidemiologic issues and science more broadly, for improving my R code, and for much needed hockey talk. Paloma, thank you for being my deskmate, sounding board, R/medical stuff teacher, co-TA, and board game opponent. I feel so lucky that I can count the three of you as not only colleagues but friends.

More broadly, thank you to all the friends who talked and laughed and danced and explored and ate and just generally existed with me through a pandemic, attempted coups, and all the smaller weirdnesses of being a stranger in a strange land. Lizzy, Annie, Claire: I don't think I would've made it without your calls and visits. Thank you to Kerina, Boots, and Andy for being so welcoming and for always being up for an adventure. Max, thank you for the copy-editing. Dino, Irma, thank you for the games, the food, and for understanding how frustrating but also great PhD life (or PhD student adjacent life) can be.

And of course, I would like to thank my family. From the beginning, you are the ones who fostered my creativity, curiosity, and determination (though you might call it stubbornness). It isn't an easy task to comfort someone 5,000 miles away, but you consistently manage to make me feel safe, supported, and loved, no matter where I am.

Finally, I owe my thanks to all the essential workers who worked tirelessly throughout the current pandemic. The fact that I was able to remain safe and healthy enough to complete this project is because of them.