University *of*
Leicester

# EXTENSIONS TO MENDELIAN RANDOMIZATION

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by


Thomas Michael Palmer

*BSc (Hons), MSc*

Centre for Biostatistics and Genetic Epidemiology
Department of Health Sciences
University of Leicester


March 23, 2009

**Extensions to Mendelian randomization**

**Thomas Michael Palmer**

The Mendelian randomization approach is concerned with the causal pathway between a gene, an intermediate phenotype and a disease. The aim of the approach is to estimate the causal association between the phenotype and the disease when confounding or reverse causation may affect the direct estimate of this association. The approach represents the use of genes as instrumental variables in epidemiological research and is justified through Mendel's second law.

Instrumental variable analysis was developed in econometrics as an alternative to regression analyses affected by confounding and reverse causation. Methods such as two-stage least squares are appropriate for instrumental variable analyses where the phenotype and disease are continuous. However, case-control and cohort studies typically report binary outcomes and instrumental variable methods for these studies are less well developed.

For a binary outcome study three estimators of the phenotype-disease log odds ratio are compared. An adjusted instrumental variable estimator is shown to have the least bias compared with the other two estimators. However, significance tests of the adjusted estimator are shown to have an inflated type I error rate, so the standard estimator, which had the correct type I error rate, could be used for testing.

A single study may not have adequate statistical power to detect a causal association in a Mendelian randomization analysis. Meta-analysis models that extend existing approaches are investigated. The ratio of coefficients approach is applied within the meta-analysis models and a Taylor series approximation is used to investigate its finite sample bias.

The increasing awareness of the Mendelian randomization approach has made researchers aware of the need for instrumental variable methods appropriate for epidemiological study designs. The work in this thesis viewed in the context of the research into instrumental variable analysis in other areas of biostatistics such as non-compliance in clinical trials and other subject areas such as econometrics and causal inference contributes to the development of methods for Mendelian randomization analyses.

# Acknowledgements

I would like to thank my supervisors John Thompson and Martin Tobin for all their help throughout my PhD. I have greatly benefitted from their expertise and also from their enthusiasm for research. In particular I enjoyed collaborating with John on the `winbugsfromstata` package and Martin's detailed knowledge of genetics has been invaluable.

I would also like to thank Paul Burton and Nuala Sheehan for their collaboration on the work on the adjusted instrumental variable estimator, Santiago Moreno for collaboration on the `winbugsfromstata` package and Alex Sutton for giving me the opportunity to be involved with the work on contour enhanced funnel plots.

I received funding through a Medical Research Council capacity building studentship in genetic epidemiology (G0501386). I won a Student Conference Award from the International Society of Clinical Biostatisticians which funded my attendance at the $27^{\text{th}}$ meeting of the society in Alexandroupolis in Greece in August 2007.

Some of the computational work in this thesis was performed using the University of Leicester Mathematical Modelling Centre's computer cluster which was purchased through the HEFCE Science Research Investment Fund. I would like to thank Stuart Poulton from the Department of Physics and Astronomy for help in using the computer cluster.

# Contents

# List of Tables

# List of Figures

# Publications

I have contributed to the following publications during my PhD studies:

1. Thompson, J. R., Palmer, T., & Moreno, S. 2006. Bayesian Analysis in Stata using WinBUGS. *The Stata Journal*, 6(4), 530-549.

2. Maznyczka, A., Mangino, M., Whittaker, A., Braund, P., Palmer, T., Tobin, M., Goodall, A. H., Bradding, P., & Samani, N. J. 2007. Leukotriene B4 production in healthy subjects carrying variants of the arachidonate 5-lipoxygenase-activating protein gene associated with a risk of myocardial infarction. *Clinical Science*, 112, 411-416.

3. Tobin, M. D., Tomaszewski, M., Braund, P. S., Hajat, C., Raleigh, S. M., Palmer, T. M., Caulfield, M., Burton, P. R., & Samani, N. J. 2008. Common Variants in Genes Underlying Monogenic Hypertension and Hypotension and Blood Pressure in the General Population. *Hypertension*, 51, 1658–1664.

4. Palmer, T. M., Thompson, J. R., Tobin, M. D., Sheehan, N. A., & Burton, P. R. 2008. Adjusting for bias and unmeasured confounding in the analysis of Mendelian randomization studies with binary responses. *International Journal of Epidemiology*, 37, 1161–1168.

5. Palmer, T. M., Peters, J. L., Sutton, A. J. & Moreno, S. G. 2008. Contour enhanced funnel plots for meta-analysis. *The Stata Journal*, 8 (2), 242–254.

6. Palmer, T. M., Thompson, J. R. & Tobin, M. D. 2008. Meta-analysis of Mendelian randomization studies incorporating all three genotypes. *Statistics in Medicine*, 27, 6570–6582.

7. Tobin, M. D., Timpson, N. J., Wain, L. V., Ring, S., Jones, L. R., Emmett, P. E., Palmer, T. M., Ness, A. R., Samani, N. J., Davey Smith, G. & Burton, P. R. 2008. Common variation in the WNK1 gene and blood pressure in childhood: the Avon Longitudinal Study of Parents and Children. *Hypertension*, 52, 947–979.

I have presented the following talks and posters at conferences during my PhD studies:

- Incorporating measures of study similarity in a meta-analysis. ISCB 26, Geneva, August 2006 (poster).

- Meta-analysis of Mendelian randomization studies. Young Statisticians meeting, UWE, Bristol, April 2007 (talk).
- Meta-analysis of Mendelian randomization studies. ISCB 27, Alexandroupolis, August 2007 (Student Conference Award, talk).
- An adjusted instrumental-variable model for Mendelian randomization. International Genetic Epidemiology Society Conference, York, September 2007 (poster).
- Performing Bayesian analysis in Stata using WinBUGS. UK Stata Users Group Meeting, Cass Business School, London, September 2007 (talk).

# Chapter 1

# Introduction

## 1.1 Aims

The aim of this thesis is to investigate statistical aspects of the application of the Mendelian randomization approach in epidemiology. The Mendelian randomization approach is concerned with the causal pathway including a gene, an intermediate phenotype and a disease. For example, Minelli *et al.* (2004) examined the pathway involving polymorphisms of the MTHFR gene, homocysteine (the intermediate phenotype) and coronary heart disease. The aim of a Mendelian randomization analysis is to estimate the association between the intermediate phenotype and the disease in a way that is robust to the possible presence of confounding or reverse causation. As such the approach now represents the use of subject's genotypes as an instrumental variable (IV) in order to estimate this association between the intermediate phenotype and the disease. The idea behind the Mendelian randomization approach has been around for about twenty years. However, it is only since the growth of the field of genetic epidemiology that the approach has been implemented in applied studies.

Instrumental variable analysis has largely been developed in the fields of econometrics and causal inference and there are a number of statistical models for such analyses. The appli-

cation of instrumental variable analysis to epidemiological studies presents some specific problems. Hence, there is a need to evaluate existing and novel statistical methods for instrumental variable analyses appropriate for epidemiological study designs. Additionally, the practice of performing a meta-analysis is now common in epidemiological research in order to increase the statistical power of an analysis. Therefore, the investigation of meta-analysis models implementing the Mendelian randomization approach is also particularly relevant.

## 1.2 Background & motivation

One of the aims of epidemiological research is to identify modifiable causes of common diseases of public health interest. Typically, epidemiological studies are observational and differ from randomized controlled trials in that subjects are not randomly allocated to each phenotype group at the start of the study. Such studies have merit because often they are the only ethical or practical way to assess a research question concerning human health. For a reported association from an observational study to be considered robust it should be replicated in other similar studies and preferably corroborated by findings from other types of studies such as randomized controlled trials (RCTs).

There have been notable epidemiological findings which have been successfully replicated such as the well known association between smoking and lung cancer (Doll & Hill, 1952). However, there have also been findings which have not been confirmed in randomized controlled trials or other studies. One example is the finding that hormone replacement therapy was protective for cardiovascular disease (Lawlor *et al.*, 2004; Rossouw *et al.*, 2002). Such spurious findings in observational research are most likely caused by confounding by social, behavioural or physiological factors which are difficult to control for, or indeed to measure accurately. In econometrics, the branch of economics concerned with statistical analysis, and causal inference instrumental variable analysis has been proposed as a method to overcome some of these potential problems.

The Mendelian randomization approach was proposed by Katan (1986) who wanted to determine the association between low cholesterol and the risk of cancer. In particular Katan was concerned whether reported associations between low cholesterol and cancer were causal (Keys *et al.*, 1985; McMichael *et al.*, 1984). Katan's idea was to investigate the distribution of the apolipoprotein E (apo E) genotypes within cases and controls, since apo E has a role in the clearance of cholesterol from blood plasma. Katan hypothesised that if the association between cholesterol and cancer was causal then the E-2 allele should be more common amongst cases. Davey Smith & Ebrahim (2003) restarted interest in the Mendelian randomization approach and then Thomas & Conti (2004) noted that Katan's idea allowed the use of genetic polymorphisms as instrumental variables.

## 1.3 Epidemiological and genetic concepts

### 1.3.1 Epidemiological concepts

A basic definition of a confounder is a variable which affects both the phenotype variable, $X$, and the outcome variable, $Y$, but is not itself affected by either of these variables (Rothman *et al.*, 2008). Failing to adjust for a confounder in the estimation of the association between the phenotype and the outcome will typically result in a biased estimate. A variable thought to be a confounder should not lie on the causal path between $X$ and $Y$, since adjustment for such a variable could bias the estimated association between $X$ and $Y$, which is sometimes referred to as collider bias (Weinberg, 1993).

Adjustment for confounding variables can be performed in a number of ways. One method is to include the potential confounder, along with the phenotype of interest, in the linear predictor of a generalised linear model (GLM). This approach to controlling for confounding views the confounding in terms of explained variation, since the confounder captures some of the variation in the outcome also explained by the phenotype. There is a close analogy with the fact that analysis of variance and covariance can be performed within the GLM framework. A method to adjust for confounding in case-control studies is the

Mantel-Haentzel odds ratio (Mantel & Haenszel, 1959) which relies upon stratifying by levels of the confounder.

The collapsibility or noncollapsibility of various different measures of risk such as risk differences, relative risks and odds ratios is also relevant when discussing confounding. It has been commented that, "much of the statistics literature does not distinguish between the concept of confounding as a bias in effect estimation and the concept of noncollapsibility" (Greenland *et al.*, 1999b). Collapsibility refers to the property of a measure of association that is constant across the strata of another variable and the observation that the odds ratio can be non-collapsible is due to Miettinen & Cook (1981). GLMs with identity or log-links are generally said to be collapsible whereas those with logit links for binary outcomes are said to be noncollapsible (Wermuth, 1987).

The following Table 1.1 demonstrates the concept of collapsibility which is adapted from Jewell (2003, Table 8.6). On the right hand side is the pooled $2 \times 2$ table of results for a study. In the table the variable D represents disease status, being diseased ($D$) or not diseased ($\overline{D}$) and the variable E represents a phenotype, subjects are either exposed ($E$) or unexposed ($\overline{E}$). This data is also presented stratified by a third variable C with two levels $C$ and $\overline{C}$. The relative risk (RR) and odds ratio (OR) for each of the tables is also given.

| | $C$ | | $\overline{C}$ | | Pooled | |
|---|---|---|---|---|---|---|
| | $D$ | $\overline{D}$ | $D$ | $\overline{D}$ | $D$ | $\overline{D}$ |
| $E$ | 120 | 280 | 14 | 86 | 134 | 366 |
| $\overline{E}$ | 60 | 340 | 7 | 93 | 67 | 433 |
| | $RR = 2.00$ | | $RR = 2.00$ | | $RR = 2.00$ | |
| | $OR = 2.43$ | | $OR = 2.16$ | | $OR = 2.37$ | |

Table 1.1: Collapsibility of the relative risk and the non-collapsibility of the odds ratio.

In Table 1.1 variable C is not a confounder since the relative risk in the pooled data is the same as in each stratum of C. However, despite that C is not a confounder the odds ratio has different values in the pooled table and each stratum of C. It is this difference in the odds ratio in the strata of a variable which does not affect disease status which is

referred to as the non-collapsibility of the odds ratio. Greenland *et al.* (1989, Table 1) is an example where the odds ratio is collapsible across strata but where the relative risk are not indicating the presence of confounding.

Another important epidemiological concept is reverse causation, which refers to the situation where an individual's disease status affects the levels of the phenotype. This phenomenon can therefore bias the estimated association between the phenotype and the disease. Reverse causation is a design issue in retrospective studies such as case-control studies.

### 1.3.2 Basic introduction to genetic terminology

Human genetic information is encoded in genes which are located on chromosomes made up of DNA (Deoxyribonucleic acid). There are 23 pairs of chromosomes in the human genome consisting of 22 pairs of autosomes and 1 pair of sex chromosomes. Somatic cells have two copies of each of the autosomes and two sex chromosomes. There are approximately $3 \times 10^9$ DNA base pairs in the human genome and the loci where DNA varies between individuals are termed polymorphic. Traditionally, a single nucleotide polymorphism (SNP) is defined as a polymorphism at a single base that occurs with a minor allele frequency of 1%, although with advances in bioinformatics database such as dbSNP (`http://www.ncbi.nlm.nih.gov/projects/SNP`) now include SNPs that have lower allele frequencies than 1%. There are estimated to be between $3 \times 10^6$ and $10 \times 10^6$ SNPs and approximately 30,000 genes in the human genome (Day *et al.*, 2001). At a genetic locus individuals with two copies of the same allele are called homozygous, while individuals with two different alleles at a locus are called heterozygous.

The obvious effects of genetic variation are seen in Mendelian disorders, in which a disease is attributable to a genetic mutation. The genetic model or mode of inheritance determines the way that a trait (or disease) is expressed with respect to the genotype. Mendelian traits are described as dominant when one copy of the mutant (risk) allele is sufficient to cause the disease, hence in this instance the disease is present in heterozygotes. Traits are

described as recessive if the disease is present in only those individuals with two copies of the risk allele. Traits can also be inherited as co-dominant, which describes a relationship where the phenotypes caused by each allele both manifest themselves when both alleles are present. The additive genetic model is a special case of co-dominance and assumes that there is an equal increase in the risk of disease per copy of the risk allele.

Assuming the common allele is denoted $g$ and the risk allele $G$, there are three possible genotypes; the common homozygotes $gg$, the heterozygotes $Gg$ and the rare homozygotes $GG$. Hardy-Weinberg equilibrium (Hardy, 1908) states that if the risk allele has frequency $q$ then the frequencies of the three genotypes should be; $(1-q)^2$, $2q(1-q)$ and $q^2$ respectively.

Apart from Mendelian traits genetic polymorphisms can affect the risk of more complex diseases. These diseases are considered to be complex because the genetic factor is only one among many factors that could possibly affect the risk of disease. Indeed there may even be many genetic factors that contribute to the risk of disease and these genetic factors may contribute in different ways. For example, these genes may act in combinations of additive, multiplicative and epistatic ways. An additive effect means each gene contributes equally in explaining the risk of disease, a multiplicative effect means that the presence of two factors increases the risk of disease by more than two times the risk associated with the first factor. An epistatic interaction means that one gene must be present in a particular form for a second gene to have an effect. Therefore, complex diseases often do not have clear patterns of inheritance and common diseases of public health relevance often have complex patterns of inheritance. It is these diseases and their respective gene-disease associations which are of particular interest in large scale genetic association studies, such as the seven diseases (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type I diabetes, type II diabetes) studied in The Wellcome Trust Case Control Consortium (2007).

A phenotype is an observable characteristic of an organism such as a biochemical or physical property. Phenotypes are commonly affected by both genetic and environmental

factors and cover a very diverse range of examples including; blood group, level of C-reactive protein in the blood plasma and bone mineral density.

There are two main types of genetic epidemiological study, linkage and association analyses. Mendelian randomization analyses use genetic association analyses. Genetic association studies more closely resemble traditional epidemiological studies in which individuals with and without the disease of interest are compared in order to assess the relative importance of a proposed genetic risk factor. A genetic association is inferred if the allele or genotype of interest is more frequent than would be expected by chance in a group of affected individuals than in a group of non-affected individuals.

Some additional genetic terminology is defined in Appendix A. The following section describes Mendel's laws, of which the second law provides the basis for the Mendelian randomization approach.

## 1.4 Mendel's laws

In 1865 Gregor Mendel (1822–1884) published experiments investigating the properties of heredity (Mendel, 1865). Mendel studied a number of heritable traits in peas including seed colour. Mendel was particularly interested in traits that were inherited randomly (half from each parent) and from these experiments he postulated the existence of what are now known as genes (Speed & Zhao, 2007). Mendel's work has come to be known as the two laws of inheritance; the law of segregation and the law of independent assortment.

### 1.4.1 The law of segregation

The law of segregation, also known as Mendel's first law, was derived from observing how two gene variants controlled seed colour and how this trait was inherited. The law of segregation has three parts:

(i) The concept of alleles; different versions of genes, made up of alleles, account for

variations in inherited characteristics.

(ii) For each characteristic an individual inherits two alleles, one from each parent.

(iii) The two alleles for each characteristic segregate at gamete production.

### 1.4.2 The law of independent assortment

The law of independent assortment, also known as the 'inheritance law' or Mendel's second law, was derived from observing how gene-pairs for seed shape and colour were jointly inherited. The law of independent assortment states that during gamete formation the segregation of one gene-pair is independent of other gene-pairs. When two gene-pairs segregate each haploid gamete is equally likely to have each of the possible combinations of genotypes (Speed & Zhao, 2007). This means that the inheritance pattern of one trait will not affect the inheritance pattern of another, to quote Mendel (1865),

> *the behaviour of each pair of differentiating characteristics in hybrid union is independent of the other differences between the two original plants, and, further, the hybrid produces just so many kinds of egg and pollen cells as there are possible constant combination forms.*

Mendel's second law implies that for a given genetic polymorphism individuals are randomized to a particular genotype because each allele is inherited at random from each parent (Wijsman, 2002). It has been argued that this process is akin to that of a randomization process such as would be implemented in a randomized controlled clinical trial (Davey Smith & Ebrahim, 2003). The consequence of this randomization is that the individual's genotype for a given polymorphism is assigned independently of other factors. Therefore, at the population level the distribution of genotypes will be independent of the distribution of the possible confounders and, since this mirrors the conditions for an instrumental variable, a genotype can therefore be suitable candidates for use as an instrumental variable.

Importantly, the discovery that some loci are more likely to be inherited together, which can be true even loci which are separated by large genetic distances on chromosomes, is expressed in the concept of linkage disequilibrium. Hence, Mendel's second law only applies to loci in linkage equilibrium. On its own Mendel's second law is not sufficient to allow the use of genes as instrumental variables and ensure the validity of Mendelian randomization analyses, other conditions are also required which are discussed throughout this thesis.

## 1.5 Statistical methods

This section introduces the concepts underlying the statistical methods discussed in this thesis, including; instrumental variable analysis, weak instruments, directed acyclic graphs (DAGs) and meta-analysis models.

Instrumental variable analysis was proposed and has largely been developed within econometrics. Instrumental variable analysis has also received attention in causal inference because of its ability to draw causal inferences in certain circumstances. Taking the example of linear regression, the aim of an instrumental variable analysis is to overcome the bias in ordinary least squares parameter estimates caused by the inclusion of covariates which are correlated with the error terms. This correlation between the covariate and the error term violates an assumption of linear regression analysis and is known in econometrics as endogeneity. Arguably endogeneity could be caused by confounding and reverse causality.

In causal inference graphs have been used to represent causal links between variables (Greenland *et al.*, 1999a). In a graph any line or arrow connecting two variables is called an arc or an edge. Two variables in a graph are adjacent if they are connected by an edge. Single headed arrows represent direct links from causes to effects and points on the graph representing variables are called nodes or vertices. A path through a graph is any unbroken route traced out along or against arrows connecting adjacent nodes. A directed path from one node to another is one that can be traced through a sequence of single

headed arrows, such a path can also be called a causal path.

A variable $X$ is said to be a cause of another variable $Y$ if there is a directed path of arrows leading out of $X$ into $Y$, in which case $Y$ is said to be a descendant of $X$. $X$ is said to be a parent of $Y$ if there is a single headed arrow from $X$ to $Y$, in which case $Y$ is a child of $X$. A two-headed arrow connecting two variables in a graph is used to indicate that two variables share one or more ancestors, i.e. they have a common cause. A graph is directed if all edges between variables have either a single or double-headed arrow. A graph is acyclic (or recursive) if no directed path in the graph forms a closed loop. If a graph satisfies both of these conditions it is a directed acyclic graph (DAG).

Figure 1.1 shows the assumed relationships between the genotype, phenotype, confounder and disease in a Mendelian randomization analysis.

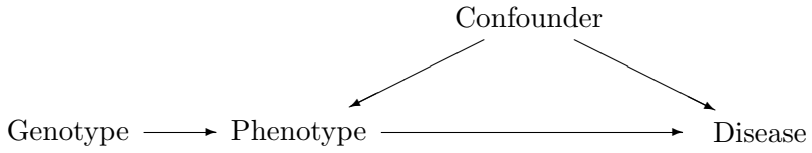



Figure 1.1: The relationship between the variables in a Mendelian randomization analysis.

For a variable to be used or to qualify as an instrumental variable it should fulfill certain conditions which are expressed in Figure 1.1 when this figure is interpreted as a DAG. These 'core' conditions for Mendelian randomization are (Didelez & Sheehan, 2007b; Lawlor *et al.*, 2008d):

(i) The genotype should be associated with the phenotype of interest.

(ii) The genotype should be independent of confounding factors that confound the association between the phenotype and the disease.

(iii) The genotype should be independent of the disease given the phenotype and the confounding factors.

An additional condition for the analysis to have a causal interpretation is that:

(iv) All of the associations are linear and unaffected by statistical interactions.

The other implication of Figure 1.1 is that if the gene-disease association is non-significant then the phenotype-disease causal association will also be non-significant (Sheehan *et al.*, 2008).

In econometrics there are several methods for instrumental variable analysis that produce equivalent estimates when all the relationships between the variables are linear without interactions. Two such methods are known as the two-stage least squares and the ratio of coefficients approaches (Greene, 1999; Stewart & Gill, 1998). Two-stage least squares involves two linear regressions; the first of the phenotype on the instrument which produces predicted values of the phenotype. In the second stage the disease outcome variable is regressed on the predicted values of the phenotype. The ratio of coefficients approach first estimates the associations between the gene and the disease and between the gene and the phenotype. The instrumental variable estimate of the association between the phenotype and the disease is then given by the ratio of these estimates.

An important concept in instrumental variable analyses is whether a chosen instrumental variable is a 'weak' instrument. A weak instrument is loosely defined as an instrument which produces an $F$-statistic less than 10 in the first stage regression of the phenotype on the genotype (Lawlor *et al.*, 2008d). Weak instruments were first discussed in econometrics by Bound *et al.* (1995) and if the chosen instrument is 'weak' this can cause bias in the IV estimate of the association between phenotype and the disease (Staiger & Stock, 1997).

There are some differences in the terminology used in statistics and econometrics that relates to instrumental variable analysis. In econometrics a variable is said to be exogenous in a linear regression if it is not correlated with error term (Wooldridge, 2002, p 50). For instrumental variable estimation econometricians sometimes say that exogenous variables act as their own instruments since they are unrelated to instrumental variables. In econometrics an endogenous variable is a variable which is correlated with the error term in a linear regression. The presence of such endogenous variables will bias the ordinary least squares parameter estimates and instrumental variable analysis was proposed as a method

to overcome the problem of endogenous variables (Wooldridge, 2002, p 50).

Meta-analysis is now common in epidemiology and aims to provide a quantitative summary of the evidence on a particular research question (Sutton *et al.*, 2000). The meta-analysis of genetic association studies is becoming more common since the instigation of HuGE reviews through the Human Genetic Epidemiology Network (HuGENet) (Ioannidis *et al.*, 2006; Little & Higgins, 2006). Meta-analyses of genetic association studies are potentially very powerful (Kavvoura & Ioannidis, 2008). However, these meta-analyses face some of the same difficulties as meta-analyses of epidemiological studies and clinical trials such as; publication bias, between study heterogeneity and varying baseline risk. Meta-analyses of genetic association studies also face specific limitations including the assessment of Hardy-Weinberg equilibrium within each study (Thompson *et al.*, 2008).

## 1.6 Outline of thesis

The outline of this thesis is as follows; Chapter 2 is a review of the literature relating to the Mendelian randomization approach. The literature review covers relevant methodology and the application of the Mendelian randomization approach within epidemiology.

Chapters 3 and 4 are concerned with the application of Mendelian randomization analysis to a study with a binary outcome. Three estimators of the phenotype-disease log odds ratio and their theoretical properties are considered in Chapter 3 which are investigated through simulations in Chapter 4. Appendix B discusses related theory for some of the other commonly used generalized linear models. Appendix C presents simulation results for some of these other GLMs.

Chapters 5 considers meta-analysis models for Mendelian randomization studies. This work builds on previously proposed multivariate meta-analysis models. Chapter 6 uses a Taylor series approximation to investigate the finite sample bias in the ratio of coefficients approach.

Chapter 7 is the discussion and conclusion. The glossary in Appendix A outlines relevant genetic terminology. R and Stata code for some of the analysis is given in Appendix D and two publications arising from this thesis are given at the end of the thesis.

# Chapter 2

# Review of the literature relating to the Mendelian randomization approach

## 2.1 Introduction

The aim of this literature review is to identify and evaluate the strengths and limitations of the Mendelian randomization approach. It is notable that the Mendelian randomization draws from several subject areas such as econometrics and causal inference as well as biostatistics, so a wide selection of reference material is included to compile this review.

To identify relevant literature, the ISI Web of Science (`http://wok.mimas.ac.uk/`) was searched using the topic identifier;
`TS=(mendelian randomization OR mendelian randomisation)`. This search produced 164 references up to July 2008. Of these the majority were original research articles (94) with the remainder representing review articles, conference proceedings, letters and editorial commentaries. Using the same terms as keywords in a search of Medline (`http://www.ncbi.nlm.nih.gov/`) produced 56 references. Comparing the two sets of search

results there were some possibly spurious references in the Web of Science set because the word 'Mendelian' is given as a keyword in many articles in genetic epidemiology not specifically on the topic of Mendelian randomization and the Medline results did not include conference proceedings such as Youngman *et al.* (2000).

Additional references were identified using search tools such as Google Scholar (`http://scholar.google.co.uk/`) and Scirus (`http://www.scirus.com`). Notably, the Mendelian randomization approach has also been discussed in several textbooks on epidemiology (Elwood, 2007), statistical genetics (Lauritzen & Sheehan, 2007) and causal inference (Didelez & Sheehan, 2007a) which are not found by the above search tools.

This review covers the statistical methodology, the initiation of the Mendelian randomization approach, its assumptions and limitations and its application within epidemiology.

## 2.2 Initiation and advantages of the approach

This section outlines the initiation, adoption and advantages of the Mendelian randomization approach. The disadvantages of the approach are covered in Section 2.4.

### 2.2.1 Initiation of the approach

The aim of epidemiological research is to investigate the aetiology of disease within human populations. Notable epidemiological discoveries include identifying the associations between smoking and lung cancer, between asbestos and mesothelioma and between intrauterine radiation and childhood leukaemia (Vandenbroucke, 2004). However, epidemiology relies upon observational evidence and as a consequence there have been reported associations which have not been replicated in later studies or confirmed in clinical trials (Davey Smith & Ebrahim, 2002). It is therefore important to explain these false positive findings.

The gold standard for medical research is a randomized controlled trial. In an RCT

the randomization of subjects to treatment means that the statistical comparison of the treatment groups should be free from confounding factors and represent an unbiased comparison of the treatments. As discussed in Chapter 1, in an observational study such as a case-control or cohort study it is possible to apply statistical methods to control for other measured covariates (Greenland & Morgenstern, 1989). However, this leaves the possibility that a result could be affected by confounding factors that were unmeasured or other mechanisms such as reverse causation. It is argued that the Mendelian randomization approach confers some of the benefits of randomization to observational studies, to quote Bubela (2006),

> *The Mendelian randomization approach hopes to revitalize the discipline of epidemiology by strengthening causal inferences about environmentally modifiable risk factors.*

The use of genetic data to test the relationship between a quantitative intermediate phenotype and a disease in a way that is recognisable as Mendelian randomization was first described by Katan (1986). The approach was not immediately widely adopted because the use of genetic data in epidemiology was then uncommon. The first use of the term 'Mendelian randomization' was made by Gray & Wheatley (1991) although the application was in the context of a clinical trial rather than an observational study. The approach has also been referred to as 'Mendelian deconfounding' (Tobin *et al.*, 2004), to avoid confusion with the underlying biological process, and 'Mendelian triangulation' (Bautista *et al.*, 2006), to reflect the way in which the phenotype-disease association is inferred from the gene-disease and gene-phenotype associations. It is now generally accepted that Mendelian randomization represents the use of genes as instrumental variables in epidemiological research (Lawlor *et al.*, 2008d; Wehby *et al.*, 2008).

The number of articles citing the term "Mendelian randomisation/randomization" in the ISI Web of Science is shown in Figure 2.1. There was minimal reference to the approach until Davey Smith & Ebrahim (2003) restarted interest and there are now upwards of 30 articles per year published referencing the topic. In particular the topic of Mendelian

randomization and specifically Katan's letter to the Lancet was the subject of a "Reprints and Reflections" (Katan, 2004) in the International Journal of Epidemiology (volume 33, number 1, 2004), which explains the increase in the number of papers published in that year.

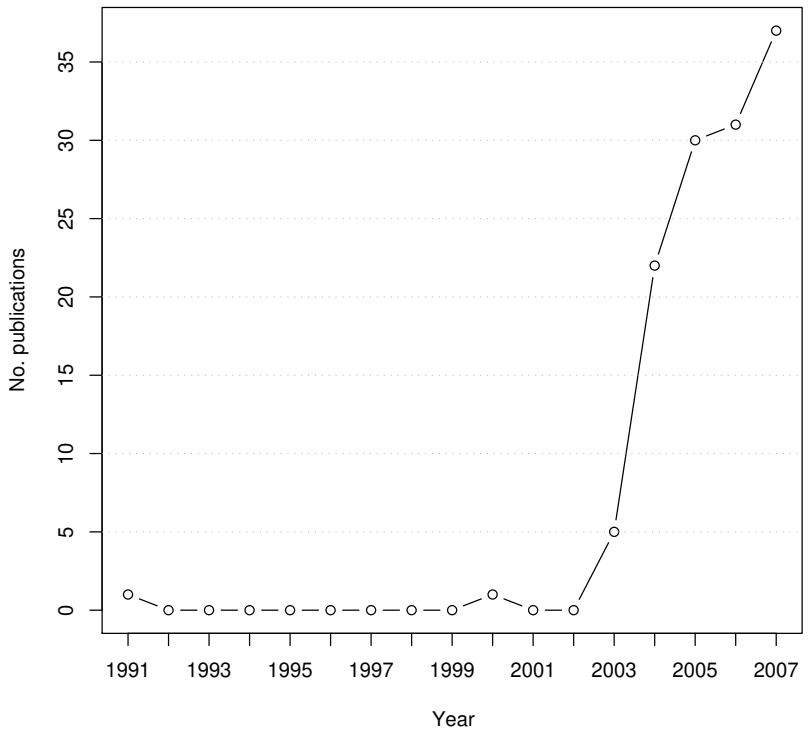



Figure 2.1: The number of articles citing the term "Mendelian randomization/randomisation" as recorded by ISI Web of Science.

There have also been a number of foreign language articles on the subject of Mendelian randomization (Bammann & Wawro, 2006; Norby, 2005; Novotny & Bencko, 2007; Olsen & Thulstrup, 2005), which indicates that the approach is becoming more well known. The rise in the number of articles can also be explained due to the rise in the number of replicated gene-disease associations which can be exploited within applied Mendelian randomization studies.

### 2.2.2 Katan's original idea

The proposal of Katan (1986) which instigated the Mendelian randomization approach has been summarised by Elwood (2007). An association was seen between low cholesterol levels and increased cancer rates in observational studies such as McMichael *et al.* (1984). It was suggested that this could be either a causal effect, with a reduction in cholesterol causing an increase in cancer risk (hence requiring the treatment of high cholesterol to be reconsidered), or due to pre-symptomatic cancers causing a reduction in cholesterol levels.

Katan was aware that conventional studies of the association between cholesterol and cancer could be limited by many factors associated with cholesterol levels such as dietary factors. Katan therefore proposed to compare cancer risks in people with different polymorphisms of the apolipoprotein E gene. Individuals with the E2 allele have lower levels of cholesterol because their genotype gives them greater efficiency in removing cholesterol from blood plasma. Therefore, if low cholesterol causes an increased risk of cancer, people with the E2 allele should have a higher risk of cancer. Also a comparison of subjects with different genotypes should be free of confounding as the genotype would be distributed randomly with respect to cholesterol levels. However, Katan's idea was not immediately pursued, although as Davey Smith & Ebrahim (2003) comment there are now a few reports about the risk of cancer and apolipoprotein E.

### 2.2.3 Justification of the approach

The key argument behind the Mendelian randomization approach is that Mendel's laws justify the use of a subject's genotype as an instrumental variable. The majority of papers using the approach have followed the argument of Davey Smith & Ebrahim (2003) by justifying the use of a genotype as an instrumental variable through the use of Mendel's second law.

### 2.2.4 Controlling for unmeasured confounding

One of the main advantages of an instrumental variable analysis is the ability to control for confounding. It is common practice in biostatistical research to control for measured confounding variables within a statistical model. This typically results in a less biased estimate of the effect of the variable of interest. However, many confounders of observed associations may be unknown or unquantifiable. It has therefore been argued that it is better to use modelling approaches that directly guarantee the unbiasedness of the phenotype-disease such as Mendelian randomization (Vandenbroucke, 2004), or at least to compare the two approaches (Bautista *et al.*, 2006). It has been commented by Lauritzen & Sheehan (2007) that,

> *Mendelian randomisation has been proposed as a method to test for, or estimate, the causal effect of an exposure or phenotype on a disease when confounding is believed to be likely and not fully understood.*

It is helpful to view confounding in terms of the variation in the outcome captured by variables in the statistical model. A confounding variable captures some of the variability in the distribution of the outcome variable explained by the variable of interest. This in turn distorts the observed effect of the phenotype variable of interest (Pearl, 2001). If the criticism that unmeasured confounding is present in a study then its presence should be both, "biologically and quantitatively plausible" (Clayton, 2007).

### 2.2.5 A lifelong effect estimate

The use of genetic information in a Mendelian randomization analysis makes use of a lifelong marker for the disease of interest (Brennan, 2004). Hence, a strength of Mendelian randomization analyses is that the estimate of the phenotype-disease association reflects a causal effect of lifelong mean differences in the phenotype (Lawlor *et al.*, 2008c).

However, it is difficult to disentangle the issue of whether a Mendelian randomization es-

timate is of a different lifelong effect compared with the direct estimate of the phenotype-disease association from an observational study. For instance, Lawlor *et al.* (2008c) explained the differences between instrumental variable and direct estimates of the association between cholesterol and coronary heart disease using this lifelong effect estimate argument.

## 2.3 Parallels with randomized controlled trials

Davey Smith & Ebrahim (2003) argue that the Mendelian randomization approach confers some of the benefits of randomization, as used in randomized controlled trials, to epidemiological analyses. This argument is often presented with a diagram similar to that shown in Figure 2.2.

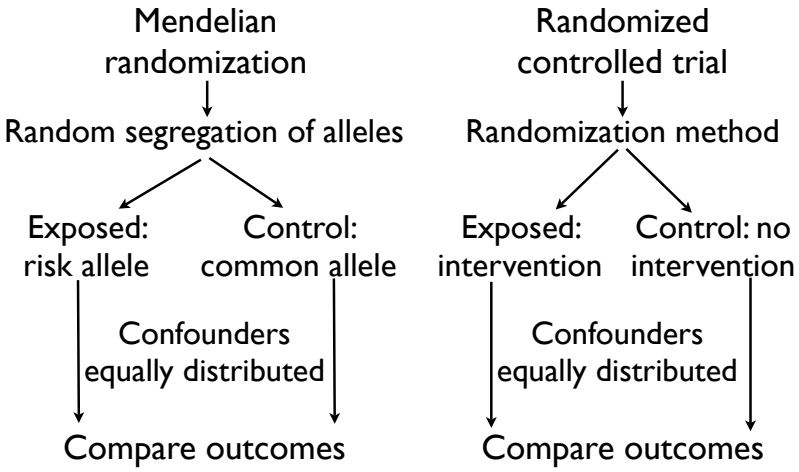



Figure 2.2: Comparison of Mendelian randomization and an RCT adapted from Davey Smith & Ebrahim (2005, Figure 1).

A modified version of Figure 2.2 has been used by other authors in subsequent articles which is shown in Figure 2.3. In this modified version it is slightly clearer that it is the association between the phenotype and the disease that is of primary interest rather than the gene-disease association.

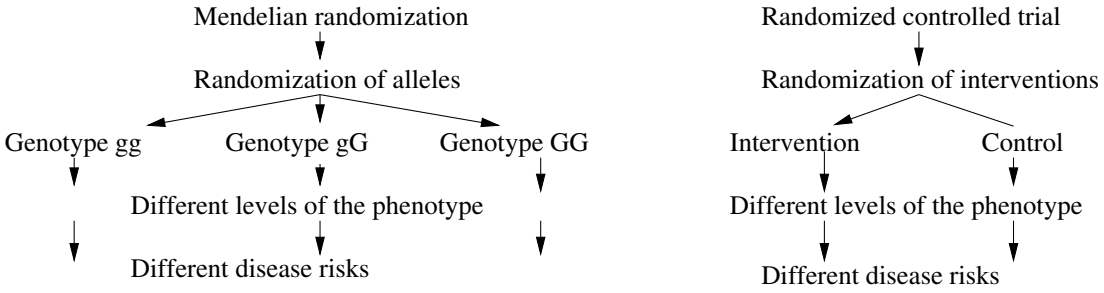



Figure 2.3: Alternative comparison of Mendelian randomization and an RCT adapted from Hingorani & Humphries (2005, Figure 2) and CRP CHD Genetics Collaboration, (2008, Figure 1).

### 2.3.1 Gray and Wheatley's approach to Mendelian randomization

Gray & Wheatley (1991) were the first authors to use the term 'Mendelian randomization' however their application of the approach in this and subsequent papers (Wheatley, 2002; Wheatley & Gray, 2004) was slightly different from the epidemiological applications. They proposed to use the approach in a clinical setting in which it was neither possible nor ethical to randomize patients to treatment groups.

Gray and Wheatley's approach to Mendelian randomization is that there are some situations in which it is not possible to perform a randomized controlled trial. Their main example is the assessment of the efficacy of allogenic stem cell transplantation (SCT) in leukaemia. They report that the ongoing Medical Research Council trial AML 15 has been designed to evaluate SCT using a donor versus no-donor comparison, those without a donor receive Conventional Intensive Consolidation Chemotherapy (CCT), based on their approach (Burnett *et al.*, 2005). The motivation behind Gray and Wheatley's idea was that haematologists believed that SCT should be the treatment of choice for younger leukaemia patients who have a matched sibling (in terms of human leucocyte antigen, known as HLA) available as a bone marrow donor. One drawback of the approach is that theoretically for a given child there is only a 25% chance that one of their siblings will have the matching tissue type.

Gray and Wheatley's approach has been criticised from the perspective of mathematical

genetics (Curnow, 2005). In particular, Curnow argued that patients with a compatible sibling will have HLA genotypes in different proportions to those without a compatible sibling. However, this would only be important if the effectiveness of SCT was related to patients' genotypes.

## 2.4 Assumptions and limitations of the approach

Given that theoretically it is possible for a genotype to fulfill the conditions of an instrumental variable the next stage is to assess whether this is reasonable for a particular study. The 'core' conditions, as they have been described by Didelez & Sheehan (2007b), for a genotype to be an instrumental variable were given in Chapter 1. The conditions state that the genotype should be associated with the phenotype, independent of factors that may confound the phenotype-disease association and independently distributed from the disease outcome variable given the phenotype, i.e. the genotype should only act through the phenotype to affect disease risk. The first and second conditions can be investigated using standard statistical tests, an example is given by Lawlor *et al.* (2008d). The third condition is not easy to assess and relies heavily on background knowledge of the genetics of the example.

In addition to the 'core' instrumental variable conditions there are other genetic and environmental factors which have been discussed in order to establish the validity of Mendelian randomization analyses. In total Davey Smith & Ebrahim (2004) and Nitsch *et al.* (2006) list nine conditions with a further three conditions suggested by Bochud *et al.* (2008) for Mendelian randomization analyses to be valid in applied studies. These have been described as necessary conditions for the use of Mendelian randomization to infer causality in observational epidemiology. The conditions adapted from Bochud *et al.* (2008) are given below and are then discussed in turn:

1. Sufficient sample size to establish reliable genotype-phenotype, or genotype-disease associations.

2. Absence of confounding due to linkage disequilibrium.

3. Absence of confounding due to population stratification.

4. Absence of pleiotropy (the multiple function of individual genes).

5. Absence of canalization or developmental compensation (a functional adaption to a specific genotype influencing the expected genotype-disease association or social pressures on behaviours affected by genotype).

6. A suitable genetic variant exists to study the phenotype of interest.

7. The association between gene and phenotype is strong.

8. The effects of a gene on a disease outcome acts only via the phenotype.

9. The genetically determined phenotype has a similar impact on disease risk as the phenotype.

10. Absence of segregation distortion, or transmission ratio distortion (TRD), at the locus of interest.

11. Absence of selective survival due to the genetic variant of interest.

12. Absence of parent-of-origin effect.

Large sample sizes (condition 1) are required for Mendelian randomization analyses and is one of the reasons that the meta-analysis of Mendelian randomization studies has been suggested (Lawlor *et al.*, 2008d) and performed (Lewis & Davey Smith, 2005). More generally, in genetic epidemiology it is recognised that genetic effects typically explain a small proportion of the variation in a phenotype (Frayling *et al.*, 2007a). Non-replication of the results of genetic association studies was common until recently (Hirschhorn *et al.*, 2002). Probable reasons for this include a lack of statistical power coupled with reporting and publication bias (Cardon & Bell, 2001; Little & Khoury, 2003). In recognition of this fact several journals have recently changed their publication policy in that any significant

gene-disease associations must be reproduced in a second independent study. For example, Zeggini *et al.* (2007) provide replication of their association between certain polymorphisms and type II diabetes.

The CRP CHD Genetics Collaboration, (2008, Figure 4) (CRP: C-reactive protein; CHD: coronary heart disease) present sample size calculations for a Mendelian randomization analysis which showed that typically not less than 10,000 cases are required for a minimum odds ratio of 1.2 and minor allele frequency of 5%. Sample size calculation is an area in econometrics which has not received attention because econometric analyses are usually performed on existing official studies with large sample sizes.

Mendelian randomization requires the lack of effect of linkage disequilibrium (condition 2) at the locus of interest. Linkage disequilibrium occurs when two genetic polymorphisms are associated, most commonly because they are close to one another in the genome as a consequence of being inherited together over many generations. Mendel's second law, that genes segregate independently, fails to hold when two genetic polymorphisms are in linkage disequilibrium with one another. An association between a genotype and a disease might therefore be biased due to the omission of other SNPs in linkage disequilibrium with the SNP of interest. Also, differences in patterns of linkage disequilibrium between populations may partly account for differing estimates from gene-disease association studies (Little & Khoury, 2003).

Didelez & Sheehan (2007b) provide a number of DAGs to show how some of these conditions might affect a Mendelian randomization analysis, some of which are included in this review. In the DAGs, $G$ represents the genotype variable, $X$ the phenotype variable, $Y$ the disease outcome variable and $U$ a confounding variable.

Figure 2.4(a) shows that if the gene of interest $G_1$ is in LD with another gene $G_2$ which has an effect on disease risk not through the phenotype then the core IV condition that the gene should be independent of the disease is violated. Figure 2.4(b) shows that if $G_1$ is in LD with $G_2$ and $G_2$ is associated with the confounders then the IV condition that the gene should be independent of the confounders is violated. Figure 2.4(c) shows a situation

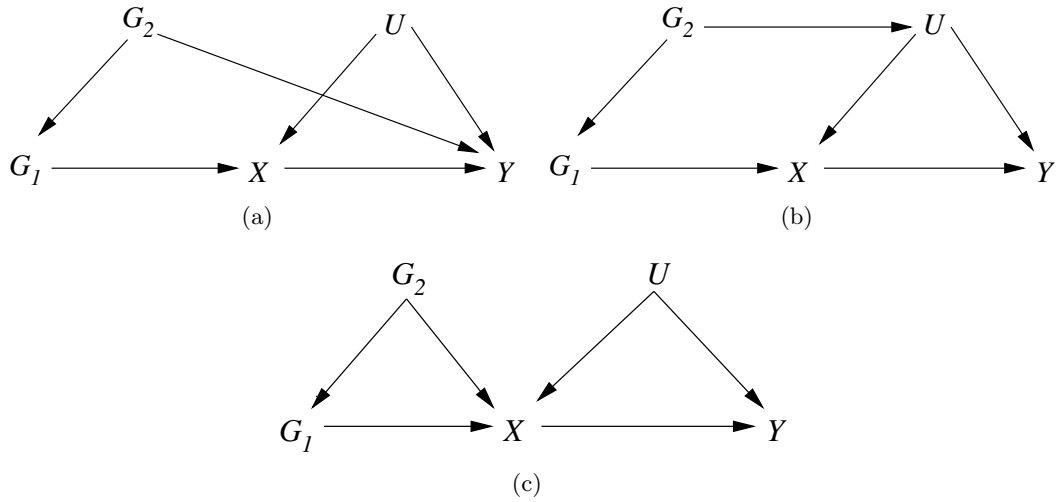



Figure 2.4: DAGs for possible ways linkage disequilibrium may occur in a Mendelian randomization analysis (Didelez & Sheehan, 2007b, Figure 5).

in which linkage disequilibrium does not violate the core IV conditions because $G_2$ is only associated with the disease through the phenotype.

There is a concern that an observed association between a gene and a disease may be affected by population stratification (condition 3) (Little *et al.*, 2003) and that this in turn threatens the reliability of Mendelian randomization analyses (Thomas & Witte, 2002). The effects of population stratification in genetic associations studies is sometimes described as confounding at the genetic level because different genetic subgroups within a sample carry different risks of disease. If the presence of population stratification is known in a study it can be controlled for by performing a stratified analysis (Ardlie *et al.*, 2002; Cardon & Palmer, 2003; Thomas & Witte, 2002; Wacholder *et al.*, 2002). Indeed within genetic epidemiology caution over population stratification has led to instances where family-based associations have been used in place of case-control studies (Ziegler & König, 2006, Chapter 10).

However, population stratification may be harder to control for if there has been recent admixture in a population (Knowler *et al.*, 1988). Typically, admixture is only problematic in certain populations such as in the US, whereas UK based studies such as the The Wellcome Trust Case Control Consortium (2007) have been found to be relatively free of

population stratification. Davey Smith (2006) states that the Mendelian randomization approach should be applied in populations of homogeneous origin.

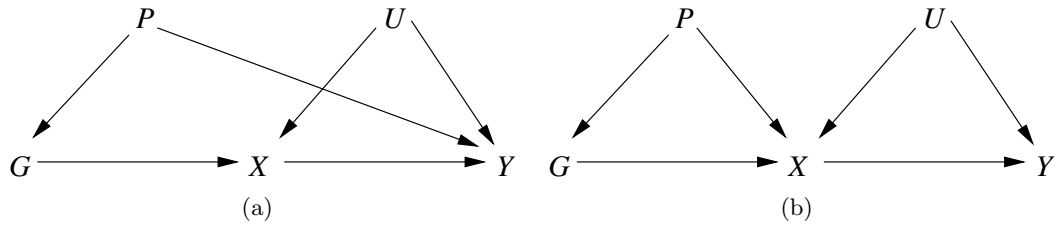



Figure 2.5: DAGs for possible ways population stratification may affect a Mendelian randomization analysis (Didelez & Sheehan, 2007b, Figure 8).

Figure 2.5(a) shows that population stratification can violate the IV condition that the genotype should be independent of the disease. Figure 2.5(b) shows a situation in which population stratification does not violate any of the core IV conditions, and it has been commented that such instances may strengthen inferences about the genotype-disease association (Didelez & Sheehan, 2007b).

Pleiotropy (condition 4) can be problematic for a Mendelian randomization analysis if another compound being metabolized by a gene also affects the risk for the disease under question. This is shown in Figure 2.6(a) and demonstrates that gene may no longer be independent of the risk of disease. Figure 2.6(b) shows the situation where the second compound is linked with the confounders which means the gene is no longer independent of them, which also violates a core IV condition. It may be possible to adjust for the effects of this second compound if its function is known (Brennan, 2004).

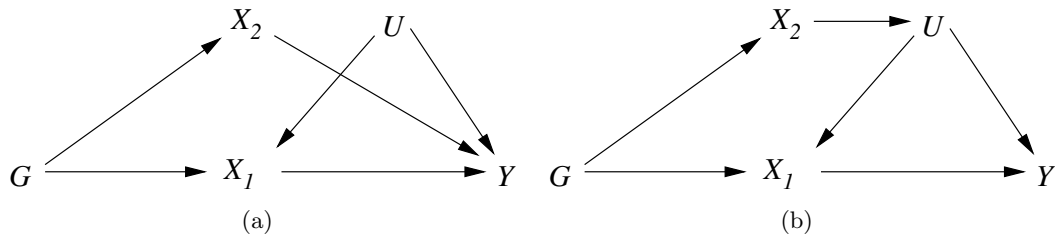



Figure 2.6: DAGs for possible ways pleiotropy may affect a Mendelian randomization analysis (Didelez & Sheehan, 2007b, Figure 6).

Canalization or developmental compensation (condition 5) refers to processes which reflect

developmental buffering against the effect of a polymorphism during fetal or possibly post-natal development. As Davey Smith & Ebrahim (2004) point out certain Mendelian randomization analyses are more prone to canalization than others. For example, if the phenotype is associated with a behaviour that is only adopted after development has ceased then canalization should not affect these analyses.

As noted by Vineis (2004) a Mendelian randomization analysis requires considerable knowledge on the part of the researcher since knowledge about suitable genetic variants and their function is required (condition 6). Brennan (2004) reinforces this by stating that the problem with the Mendelian randomization approach is that the gene responsible for the environmental or lifestyle agent must be known a priori to the analysis, also knowledge of disease mechanisms is required (Little & Khoury, 2003).

The requirement for a strong gene-phenotype association (condition 7) is to avoid the possibility that the genotype may be a 'weak instrument' and to try to ensure the resulting estimate of the phenotype-disease association is precise. If the instrument is 'weak' then the two-stage least squares estimate, for use with a continuous disease variable, will be biased (Staiger & Stock, 1997).

That the effect of the gene on the disease outcome should be only via the phenotype (condition 8) is one of the core IV conditions. Biological background knowledge is required in order to judge whether this condition holds.

That the genetically determined phenotype has a similar impact on disease risk as the environmental phenotype investigated (condition 9) was suggested by Bochud *et al.* (2008). It is not actually clear that this condition required by instrumental variable theory, for example it is more important to examine the strength of the instrumental variables.

Segregation distortion or meiotic drive (condition 10) describes transmission ratio distortion (TRD) during meiosis. This occurs when the distribution of alleles at a particular locus differs from the distribution of alleles at a particular locus in the surviving offspring from that expected in Mendelian proportions. After meiosis transmission ratio distortion

may result from selective survival between conception and birth or later. Davey Smith & Ebrahim (2008) argued that TRD would only affect a Mendelian randomization analysis if it induced a correlation between the genotype and a confounding factor which these authors considered to be unlikely at a population level.

Absence of selective survival due to the genetic variant of interest (condition 11) is possibly a reference to the concern that differential genotype outcomes could bias the results of genetic association studies.

A parent-of-origin effect (condition 12) occurs when the effect of an allele on a phenotype is dependent upon which parent the allele was inherited from. For example, some genes are functionally active depending on whether a particular variant is inherited from the mother or father. The presence of a parent-of-origin effect implies that the effect on the phenotype conferred by a specific genetic variant is not homogeneous in the population (Bochud *et al.*, 2008). Davey Smith & Ebrahim (2008) stated that this was unlikely to be problematic at a population level and concluded that more empirical data is required before it is known whether the extra conditions of Bochud (conditions 10–12) are of practical importance.

### 2.4.1 Comparison with traditional epidemiological methods

It is interesting to compare direct estimates of a given phenotype-disease association with Mendelian randomization estimates, which is akin to the idea behind the Hausman test. Such a study has been performed by Bautista *et al.* (2006) who compared direct phenotype-diseases associations from observational studies with associations derived from Mendelian randomization analyses. Bautista *et al.* (2006) argue that whilst Mendelian randomization estimates are supposedly unbiased, they may be inaccurate if the sample size is too small. As such they propose a method that gives a range of plausible values for unbiased odds ratios for a Mendelian randomization analysis. The authors argue that their approach is more informative than a statistical test between the results of the different studies. However, Bautista *et al.* (2006) encountered a number of practical difficulties, for example it was not always possible to compare the direct and IV estimates from the same study,

and their approach has been criticised for being statistically naive (Thomas *et al.*, 2007).

A similar comparison of direct and Mendelian randomization phenotype-disease estimates was made in a review evaluating the role of fibrinogen and C-reactive protein as risk factors for cardiovascular disease by Davey Smith *et al.* (2004). The findings of the two previous reviews of Kamath & Lip (2003) and Hirschfield & Pepys (2003) were not in agreement with Mendelian randomization analyses. However, it was argued that the original studies were under-powered because the genotype-phenotype effect was modest compared with the variability of the phenotype in the population. It can be noted that the CRP CHD Genetics collaboration has been set up to help resolve questions in this research area. Also, Kolz *et al.* (2008) have found direct estimates that were adjusted for confounding factors which were in agreement with the Mendelian randomization analysis of Timpson *et al.* (2005).

### 2.4.2 Other limitations

Nitsch *et al.* (2006) were cautious about Mendelian randomization from the perspective of causal inference because Mendelian randomization studies are observational and there is a general issue about how to make causal inferences from observed associations. For example, making causal inferences using instrumental variable analysis is not included in the Bradford Hill causality criteria (Bradford Hill, 1965). However, this probably reflects that instrumental variable methods were not well known when the causality criteria were written.

Figure 2.7 illustrates genetic heterogeneity, the situation when more than one gene affects the phenotype. As represented here, genetic heterogeneity would not violate any of the core IV conditions assuming that neither $G_2$ nor $G_3$ affected the confounders or the risk of disease in a way other than through the phenotype. Didelez & Sheehan (2007b) argue that genetic heterogeneity could weaken an observed gene-phenotype association and therefore genetic heterogeneity could be an explanation if a genotype was found to be a weak instrument. Bochud (2008) argues that genetic heterogeneity could weaken an association

in a Mendelian randomization study since the level of genetic heterogeneity underlying many complex traits is unclear. They argue that most genetic association studies rely on the assumption the "common disease, common variant hypothesis" which implies that the same variants are causal in affected individuals. Therefore, if this assumption did not hold for a particular analysis then genetic heterogeneity could weaken an association.

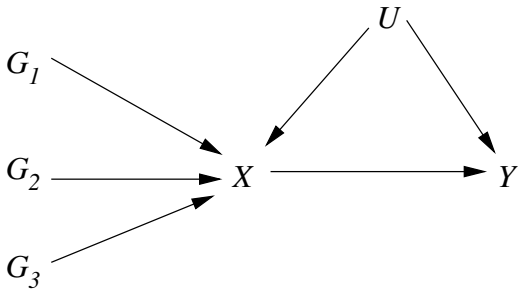



Figure 2.7: DAG demonstrating genetic heterogeneity in a Mendelian randomization analysis (Didelez & Sheehan, 2007b, Figure 7).

Additionally, Mendelian randomization does not inform strategies for genetic screening for disease risk or targeting therapy (Davey Smith, 2006). Also Sheehan *et al.* (2008) point out that causal inference is not the primary interest in prognostic research.

For the data from a genetic-association study to be used in a Mendelian randomization analysis the genotypes of the controls should be in Hardy-Weinberg equilibrium (HWE). Only the controls are specified since selection by disease status may affect HWE (Ziegler *et al.*, 2008a). HWE in the controls indicates that the data is a representative sample from the population in order to robustly infer gene-disease associations in genetic association studies (Salanti *et al.*, 2005a). Rodriguez *et al.* (2009, Table 6) consider the impact of deviations from Hardy-Weinberg in controls on the gene-disease association $p$-value. They present a sensitivity analysis by adding hypothetical missing observations to put a study in perfect HWE and argue that these missing observations can strengthen the gene-disease association. Important causes of departure from Hardy-Weinberg equilibrium include genotyping error, which may be differential between cases and controls (Clayton, 2007; Clayton *et al.*, 2005). In this regard it is also desirable that those undertaking the genotyping are blind to the case-control status of the samples to avoid selective checking of

the results or biased genotype calling if there is overlap between genotypes on the intensity plot for a particular polymorphism (Ziegler *et al.*, 2008a).

Mendelian randomization has been criticised that it over-simplifies the underlying biology of causal pathways of disease, Jousilahti & Salomaa (2004) comment that,

> *"Mendelian randomisation" is, in most cases, a gross oversimplification of the underlying biology of a complex, multifactorial disease. We suspect that its applicability is likely to be rare and limited to a few special occasions.*

However, as pointed out by Davey Smith (2006) this criticism misses the point that Mendelian randomization is essentially instrumental variable analysis with genetic instruments as clarified by Wehby *et al.* (2008). As noted by Thomas & Conti (2004), those in favour of Mendelian randomization analyses do not doubt that biological pathways are more complex than the simple triangulation relationship (of the ratio of coefficients approach) implies.

Ziegler *et al.* (2008b) argued that the assumptions of Mendelian randomization may not hold for the applied example used by Lawlor *et al.* (2008d). Ziegler et al.'s argument was biological rather than statistical, specifically, they argued that the core condition that the genotype should be independent of confounding factors may not hold for the example. However, Ziegler et al.'s comments do not refer to an identical causal pathway since their citations found an association between tumour necrosis factor A (TNFA) genotypes and sepsis (Little *et al.*, 2006; Menges *et al.*, 2008) and CRP genotypes were used in the example. It is therefore uncertain whether the argument of Ziegler et al. would affect the example. Lawlor *et al.* (2008a) also commented that the evidence cited against their work was also not fully available in published literature.

## 2.5 Instrumental variable methods

This section discusses statistical methods for instrumental variable analysis. As noted by Thomas & Conti (2004), when Katan proposed the idea of Mendelian randomization he did not explicitly reference the method of instrumental variable analysis. However, this is the theory on which such analyses are based. The theory of instrumental variable analysis has largely been developed and applied within econometrics (Bowden & Turkington, 1984) and the first use of the term 'instrumental variable analysis' is attributed to Reiersol (1941, 1945). Theory relating to instrumental variable analysis has also been developed in causal inference such as Pearl (2000).

The Mendelian randomization approach is not the first application of instrumental variable analysis in epidemiology. For example, Grootendorst (2007) details non-genetic variables that have been used as instrumental variables in epidemiological analyses. Also, introductions to instrumental variable analysis have been written for epidemiologists such as Zohoori & Savitz (1997) and Greenland (2000).

### 2.5.1 Econometric methods for continuous outcome variables

There are three main forms of instrumental variable analysis performed in econometrics, these are known as the two-stage least squares, ratio of coefficients and control function approaches. All three methods produce equivalent parameter estimates when all variables are continuous.

Two-stage least squares was proposed by Theil (1953) and Basmann (1957) and involves a series of two linear regressions. For a Mendelian randomization analysis the phenotype is first regressed on the genotype. From this equation the predicted values of the phenotype are generated. The second stage regression is of the disease outcome on the predicted phenotype variable. It is important to correct the standard errors of the parameter estimates after the second stage of two-stage least squares, especially when the sample size is small (Baltagi, 1998, page 280). The application of two stage least squares has been carefully

described with respect to the 'core' conditions for an instrumental variable by Lawlor *et al.* (2008d), Wehby *et al.* (2008) and Sunyer *et al.* (2008). The resulting parameter estimate for the predicted phenotype variable is interpreted as the phenotype-disease association, which is assigned a causal interpretation in the absence of interactions. The intercept from two-stage least squares has the same interpretation as the intercept in an ordinary least squares regression, which is the expected value of the disease outcome variable when all other explanatory variables are set to zero.

The ratio of coefficients approach is attributed to Wald (1940) and is sometimes referred to as the Wald estimator. In the context of Mendelian randomization the method divides the estimate of the gene-disease association by the estimate of the gene-phenotype association. Carroll *et al.* (2006, Chapter 6) has given a mathematical proof of the Wald estimator.

The control function approach is not as precisely defined in the literature compared with the other two methods. The description given here is that of Nichols (2006), but some econometrics textbooks such as Cameron & Trivedi (2005) are not particularly clear in their description of the method. The name of the control function approach derives from the fact that it is designed to control for confounding and in general terms it is a method designed to approximate the influence of omitted variables. An example of the use of a control function approach for continuous variables is given by Heckman & Hotz (1989). Referring to the stages from two-stage least squares, one way to derive control function estimates for a continuous disease outcome variable is to include the residuals from the first stage regression in the regression of the disease outcome variable on the phenotype (Nichols, 2006). The use of the control function approach has been suggested to help overcome some of the difficulties in applying instrumental variable methods to non-continuous outcome measures (Karaca-Mandic & Train, 2003; Petrin & Train, 2003).

To help assess whether a proposed instrumental variable fulfills the 'core' conditions there are a number of statistical tests that have been proposed within econometrics. Two well known econometric tests are the Hausman test (Hausman, 1978) and the Sargan test (Sargan, 1958). For a parameter in a linear regression the Hausman test examines

the difference between the ordinary least squares estimate and the instrumental variable estimate. This test is also referred to as the Durbin-Wu-Hausman test (Martens *et al.*, 2006). The Hausman test is also sometimes referred to as a test of endogeneity in the sense that if there is a statistically significant difference between the two estimates then there is reason to support the presence of endogeneity. The Sargan test evaluates the quality or strength of the chosen instrumental variable, it is referred to as a test of over-identifying restrictions. The J-test (Hansen, 1982) is closely related to the Sargan test. There are also several other statistical tests relating to instrumental variable analysis which are discussed by authors such as Baum *et al.* (2007).

### 2.5.2 Fieller's Theorem

Fieller's Theorem (Fieller, 1954) is relevant for instrumental variable analyses because it is concerned with the ratio of two normally distributed random variables and hence it relates to the ratio of coefficients approach. Fieller's Theorem states that the ratio of two independent standard normal random variables is a standard Cauchy random variable, hence the Cauchy distribution is a ratio distribution. The standard Cauchy distribution arises as a special case of Student's t distribution with one degree of freedom and the mean and variance of the Cauchy distribution are undefined. Fieller's Theorem has been discussed for the case when the two normally distributed random variables are correlated (Hinkley, 1969, 1970). Marsaglia (1965) and Marsaglia (2006) provide further discussion about the ratio of normally distributed random variables.

The possible implications of Fieller's Theorem for the ratio of coefficients approach in the context of Mendelian randomization have been discussed by Thompson *et al.* (2003). This report focused on deriving appropriate confidence intervals for the ratio of coefficients estimate of the phenotype-disease association. If uncertainty in the gene-phenotype association is taken into account then the confidence interval for the phenotype-disease association may not be continuous. This has led to comments that Mendelian randomization analyses are most efficient when the genotype-phenotype relationship has high precision

(Thomas & Conti, 2004).

### 2.5.3 The Method of Moments and the Generalised Method of Moments

In econometrics M-estimation is taken to mean 'maximum-likelihood-like' estimation and covers maximum likelihood and non-linear least squares estimation techniques (Cameron & Trivedi, 2005, page 119). Method of Moments (MM) estimation specifies a set of population moment conditions and solves the corresponding sample moment conditions. MM estimation may not be feasible in certain circumstances due to the over-identification of a system of equations, i.e. there might be more moment conditions to solve than the number of parameters. The Generalised Method of Moments (GMM) technique (Hansen, 1982) was proposed to accommodate this complication (Cameron & Trivedi, 2005, Chapter 6).

Under GMM different population moment conditions specify different GMM estimators in the same way that different distributions specify different models in the generalised linear models framework (Johnston *et al.*, 2008). Foster (1997) reviewed GMM techniques including models using a logit link of the same form as those discussed by Johnston *et al.* (2008) which is relevant for epidemiological study designs. These moment conditions are discussed in Section 3.2.7.

## 2.6 Causal inference

Causal inference is a form of statistical modelling which is concerned with making causal interpretations from statistical analyses (Pearl, 2001). Causal inference is relevant to the discussion about Mendelian randomization since instrumental variable analyses can have a causal interpretation. The question of whether the estimate of a phenotype-disease association from a Mendelian randomization analysis has a causal interpretation will depend on the type of data being analysed and the methods used to perform the analysis (Didelez & Sheehan, 2007b). Statisticians are very familiar with the axiom that 'correlation does not

imply causation'. This means that an observed correlation between two variables does not imply that there is a cause-and-effect relationship between them. However, correlation between two variables is required for a causal relationship between two variables to be proved, although well designed studies are required (Holland, 1986). The following related quote is taken from Pearl (2001).

> *One cannot substantiate causal claims from associations alone, even at the population level, behind every causal association there must lie some causal assumption that is not testable in observational studies.*

From the perspective of causal inference the biological process of Mendelian randomization at meiosis has been described as a 'minimal' condition because the unique identification of the causal effect of the phenotype on disease is only possible in the presence of assumptions (Didelez & Sheehan, 2007b). These assumptions require all dependencies in the modelling framework to be linear and additive. However, it is possible to compute bounds for the causal effect of the phenotype on the disease when all variables are binary (Balke & Pearl, 1994; Manski, 1990). Bounds can be computed for the phenotype-disease association in case-control studies if the disease prevalence is known (Didelez & Sheehan, 2007b).

A central issue for Mendelian randomization analyses is whether the phenotype-disease association can be assigned a causal interpretation under the full class of GLMs (Didelez & Sheehan, 2007b). In this regard some epidemiologists may indeed have been 'fast and loose' with the use of the term 'causality' in relation to Mendelian randomization analyses. In particular when there is a non-linear relationship between a phenotype and a disease the parameter estimates from a Mendelian randomization analysis may not have a causal interpretation. However, this matter is non-trivial, which has led some authors to argue that it may be more appropriate to describe estimates derived for non-linear relationships between the phenotype and disease as free from confounding rather than causal (Tobin *et al.*, 2004).

Structural equation modelling is a form of causal inference (Pearl, 2000). Grassi *et al.* (2007) investigated the association between homocysteine and ischaemic heart disease

using Mendelian randomization. The authors described their approach as a structural equation modelling approach even though their approach was in fact the ratio of coefficients approach. Grassi *et al.* (2007) stated that in Mendelian randomization the phenotype should be a complete mediator of the genotype and in such circumstances the ratio of coefficients approach is a structural equation modelling approach.

The use of direct acyclic graphs, from causal inference, to describe Mendelian randomization analyses is particularly informative (Didelez & Sheehan, 2007b; Lauritzen & Sheehan, 2007). In particular, the typical DAG for a instrumental variable analysis, as per Figure 1.1, has been used in several articles on Mendelian randomization (Didelez & Sheehan, 2007b; Lawlor *et al.*, 2008d). Didelez & Sheehan (2007a) provide DAGs and their accompanying moral graphs for a number of limitations for Mendelian randomization analyses. A moral graph is the equivalent undirected form of a DAG with additional edges connecting nodes that have a common child (Cowell *et al.*, 1999). The purpose of a moral graph is to show which nodes are conditionally independent, as indicated by any nodes that are unconnected on the moral graph.

## 2.7 Meta-analysis methods

Meta-analyses combine the results of multiple studies to provide a quantitative summary of the evidence on a research question and are often presented to summarise a systematic review. It has been argued that it is unlikely that a single Mendelian randomization study will have a large enough sample size to have adequate power to detect a small effect estimate (Lawlor *et al.*, 2008d). Meta-analyses have higher power than individual studies to detect effect size estimates, since the power of a meta-analysis is a function of the total sample size of the included studies. It has also been argued that medical research and clinical practice should be based upon, "the totality of relevant and sound evidence" (Sutton & Higgins, 2008), hence meta-analyses should be performed for Mendelian randomization analyses.

Meta-analysis models estimate a pooled effect estimate from a set of studies and are described as either fixed or random effects models depending on the assumptions made about the underlying pooled effect (Sutton *et al.*, 2000). Fixed effects models assume that each study's effect estimate comes from a single underlying pooled effect. In fixed effects models the pooled estimate is derived as a weighted average with the weights inversely proportional to the variance of each study's effect estimate (Fleiss, 1993). Random effects meta-analysis models assume that each study has a different underlying effect estimate, the mean of which is the true underlying effect (DerSimonian & Laird, 1986). In a random effects model the 'between-study' variance of the distribution of the underlying effects has to be accommodated in the weights used to derive the pooled estimate, hence these models typically produce wider confidence intervals about the pooled estimate (Jones, 1995).

Meta-analysis models for Mendelian randomization using the ratio of coefficients approach incorporate two sets of variables and it is therefore possible to apply methods for multivariate meta-analysis (van Houwelingen *et al.*, 1993, 2002) as demonstrated by Minelli *et al.* (2004) and Thompson *et al.* (2005). The strength of multivariate meta-analysis models is that they can accommodate the correlation between the multiple outcome measures.

The limitations that apply to the meta-analysis of genetic association studies are relevant to Mendelian randomization meta-analyses. These limitations include: population structure, linkage disequilibrium, conformity to Hardy-Weinberg equilibrium, bias, population stratification, statistical heterogeneity, epistatic (the interaction between genes) and environmental interactions, and the choice of statistical models used in the analysis (Salanti *et al.*, 2005b). That these factors may vary between studies has prompted authors to argue that meta-regression techniques could be helpful for Mendelian randomization meta-analyses (Salanti *et al.*, 2005b).

Mendelian randomization meta-analyses can also benefit from HuGE Net (Human Genetic Epidemiology Network) (`http://www.cdc.gov/genomics/hugenet/`) reviews which seek to report robust gene-disease associations (Little *et al.*, 2003). There are also a number of relevant databases cataloging gene-disease associations, such as the Genetic Asso-

ciations Database (`http://geneticassociationdb.nih.gov/`) and the National Human Genome Research Institute catalogue of published genome wide association studies (`http://www.genome.gov/26525384`). There are also databases cataloging gene-phenotype associations such as the GEN2PHEN project (`http://www.gen2phen.org/index.html`) and the Human Genome Variation database of Genotype-to-Phenotype (HGVbaseG2P) information (`http://www.hgvbaseg2p.org/index`).

## 2.8 Epidemiological analyses applying the approach

Davey Smith (2007) described several applied examples of Mendelian randomization analyses. In this and another article (Ebrahim & Davey Smith, 2007) it was found that Mendelian randomization has been applied to the following risk factors and diseases; milk and osteoporosis, alcohol and coronary heart disease, sheep dip and farm workers' compensation neurosis, folate and neural tube defects and C-reactive protein and coronary heart disease, to name a few. For successful application of the approach good study design principles need to be followed such as a simple well defined phenotype and a large sample size to provide adequate power for the analysis.

In the following subsections Tables 2.1, 2.2 and 2.3 summarise some of the applied examples of individual studies and meta-analyses that have applied the Mendelian randomization approach. The tables are not exhaustive and are meant to provide examples of the types of analyses that have been performed. The tables list the disease, phenotype and genotype investigated, the findings of the study and the statistical methods used for the analysis. These tables are adapted from Sheehan *et al.* (2008, Table 1). In the tables the following abbreviations are used; CHD: coronary heart disease, CRP: C-reactive protein, FTO: fat mass and obesity related, GD: gene-disease association, GP: gene-phenotype association, MR: Mendelian randomization, PD: phenotype-disease association, T2D: type II diabetes, TSLS: two-stage least squares.

### 2.8.1 Example individual studies

Table 2.1 lists studies that reported continuous outcome measures and Table 2.2 gives studies that reported binary outcome measures.

| Disease/outcome | Phenotype | Genetic variant | Findings | Analysis | References |
|---|---|---|---|---|---|
| Carotid intima media thickness (c-IMT) | CRP | CRP gene (haplotypes derived from 5 SNPs) | MR evidence does not support a causal role for CRP in the development of a thickened intima media (and potentially later CHD) | TSLS & Durbin-Wu-Hausman test. | Kivimäki *et al.* (2007) |
| c-IMT | Lifetime BMI | FTO polymorphism rs9939609 used as IV. | MR estimates although larger not significantly different from OLS estimates of BMI-c-IMT association. | OLS, TSLS & Hausman test. | Kivimäki *et al.* (2008) |
| Metabolic phenotypes | CRP | CRP gene +1444 C to T polymorphism | CRP has been associated with metabolic phenotypes in observational studies, but MR evidence from this study does not support a causal relationship between CRP levels and any of the metabolic phenotypes studied. | TSLS & Durbin-Wu-Hausman test. | Timpson *et al.* (2005) |
| Type 2 diabetes | Macrophage migration inhibitory factor (MIF) | MIF gene (4 SNPs) | MR evidence supports a causal role for MIF in the development of T2D in women. | Compared mean MIF levels between those with and without T2D. | Herder *et al.* (2008) |
| Fat mass | Maternal BMI | FTO gene rs9939609 polymorphism | MR evidence does not support the hypothesis that maternal BMI during pregnancy affects fat mass in children aged 9–11 years. | TSLS & Durbin-Wu-Hausman test. | Lawlor *et al.* (2008b) |

Table 2.1: Studies applying Mendelian randomization reporting continuous outcomes.

| Disease/outcome | Phenotype | Genetic variant | Findings | Analysis | Reference |
|---|---|---|---|---|---|
| Blood pressure and hypertension | C-reactive protein (CRP) | CRP gene 1059G/C polymorphism | Evidence does not support a causal relationship between CRP levels and blood pressure or hypertension. | BP used TSLS & hypertension used `qvf` program in Stata. | Davey Smith *et al.* (2005a) |
| Physical function in 65–80 year-olds | Interleukin-18 (IL-18) | Four IL-18 gene polymorphisms | Supports the hypothesis that high IL-18 levels are a cause rather than a consequence of disability in the elderly. | Tested association between phenotype and disease under an IV analysis. | Frayling *et al.* (2007b) |
| Lung cancer | Cruciferous vegetables | GSTM1 and GSTT1 status | Some chemopreventive effects were seen in individuals who were GSTM1 null or null for both genes. | Not true IV; GD associations stratified by phenotype. | Brennan *et al.* (2005) |
| Still-birth | Caffeine intake | Slow oxidizer status (CYP1A2), slow acetylator status (NAT2), and low activity of GSTA1 | No link between any single genotype and the risk of stillbirth. An association between a combination of genotypes and stillbirth was discovered but replication required. | GD odds ratios reported. | Hammer Bech *et al.* (2006) |
| Type 2 diabetes | CRP | CRP haplotypes | Non-statistically significant increased risk of T2D with certain CRP haplotypes | GD odds ratios reported to test significance of PD association. | Dehghan *et al.* (2007) |
| Type 2 diabetes | Interleukin-6 | 10 SNPs for the IL6R gene | IL6R genotypes were not significantly associated with the risk of T2D in women. | The `qvf` program in Stata used to estimate odds ratios for diabetic risk per unit change of log(IL-6). | Qi *et al.* (2007) |

Table 2.2: Studies applying Mendelian randomization reporting binary outcomes.

### 2.8.2 Example meta-analyses

| Disease/outcome | Phenotype | Genetic variant | Findings | Analysis | Reference |
|---|---|---|---|---|---|
| Coronary heart disease | Fibrinogen | $\beta$-fibrinogen G-455 to A and C-148 to T polymorphisms | Evidence suggests that observational plasma fibrinogen-CHD associations are explained by confounding or reverse causation. | MA of relative risks & MA of odds ratios. | Davey Smith *et al.* (2005b); Keavney *et al.* (2006) |
| Stroke | Homocysteine | MTHFR C677T polymorphism | MR evidence consistent with causal association between homocysteine concentration and stroke | MA of GD odds ratios & GP mean differences. | Casas *et al.* (2005) |
| Myocardial infarction | C-reactive protein (CRP) | CRP gene +1444 C to T polymorphism | MR evidence does not support a causal role for CRP in non-fatal myocardial infarction. | MA of GD odds ratios & GP mean differences. | Casas *et al.* (2006) |
| Blood pressure | Alcohol intake | ALDH2 *2 allele | MR evidence supports the hypothesis that (even modest) alcohol intake increases blood pressure. | MA of GD odds ratios & GP mean differences. | Chen *et al.* (2008) |
| CHD | Homocysteine | MTHFR 677 C to T polymorphism | No strong evidence to support an association of the MTHFR 677 C to T polymorphism and CHD. | MA of GD odds ratios & GP mean differences. | Lewis *et al.* (2005) |

Table 2.3: Meta-analyses applying Mendelian randomization.

## 2.9 Discussion

This review has discussed the initiation and advantages of the Mendelian randomization approach; the parallels with randomized controlled trials; the assumptions and limitations of the approach; instrumental variable, causal inference and meta-analysis methods; and epidemiological analyses applying the approach.

This review demonstrates that there is a consensus that the use of genetic data can make a valuable contribution to epidemiological research in addition to the contribution of more standard genetic epidemiological analyses such as genetic association studies (Khoury *et al.*, 2005). Davey Smith (2007) argue that genetic evidence should be used in epidemiology not only to identify the cause of disease but also to identify modifiable environmental risk factors in order to better prevent and treat disease which is the aim of the Mendelian randomization approach.

This review has highlighted that although a relatively small number of applied Mendelian randomization analyses have been published there has been considerable discussion of the potentials and limitations of the approach within the epidemiological literature. It is a strength of the approach that it can appeal to Mendel's second law of genetics in order to justify the instrumental variable assumptions. This contrasts with the use of non-genetic instrumental variables in other areas of epidemiology and in other subject areas such as economics in which the same instrumental variable assumptions are justified using prior observational research.

The review has discussed research into instrumental variable methods from biostatistics, causal inference and econometrics. However, as Lawlor *et al.* (2008c) discussed there are barriers to overcome in terms of the different terminologies used in these subject areas. One example is the econometric concept of endogeneity which encompasses the biostatistical and epidemiolical concepts of confounding and reverse causation.

It has been shown that Mendelian randomization analyses are based on two sets of assumptions; one set which relates to the core instrumental variable conditions and another

set which relates to the practical implementation of a study on a particular research question. Therefore, in a Mendelian randomization analysis it is important to provide a thorough discussion of whether these assumptions have been met both theoretically and practically (Keavney *et al.*, 2004). The Mendelian randomization approach is also increasingly being cited in review articles as a method to explore for further research. For example, Ducimetière & Cambien (2007) referred to the approach in this way with respect to research into coronary heart disease aetiology.

In general it is becoming more common that methods for causal inference are being applied to observational data to compare the results with those from randomization controlled trials. For example, Hernán *et al.* (2008) re-analysed data from the Nurses' Health Study using causal models and found results more similar to the Women's Health Initiative trial whereas the original analysis of the Nurses' Health Study found results in the opposite direction to the trial. The argument of these authors is that when applying causal models to observational data it is important to specify the correct causal question in order to provide answers comparable with those from randomized trials.

The Mendelian randomization approach is relatively new and hence this review has found that it is currently not referenced in several of the collaborative sets of guidelines about study quality. In particular, the Mendelian randomization approach could be referenced in the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (Ebrahim & Clarke, 2007) and the Meta-analysis of Observational Studies in Epidemiology (MOOSE) statement (Stroup *et al.*, 2000). With respect to the Bradford Hill causality criteria Cox & Wermuth (2001) argue that these criteria were designed as a series of conditions which make the hypothesis of causality more convincing but none of which are either necessary or sufficient to prove causality. Hence, it is not necessary that the Bradford Hill causality criteria make reference to specific statistical methods for causal inference.

With respect to the areas of research subsequently investigated in this thesis, PHOEBE (2007) comment that,

> *The method of Mendelian randomization definitely has promise for observational epidemiology but, in order to widen its applicability, general methods for the non-linear case are required and different causal parameters could be considered.*

This highlights that there is a gap in the literature on instrumental variable models in that the equivalent of the full class of models encompassed by the family of generalised linear models has not been clearly defined. Hence, there is scope to investigate instrumental variable estimators appropriate for the analysis of epidemiological studies such as case control and cohort studies and this provides the motivation for the work in chapters 3 and 4.

It has been noted that the use of meta-analysis are becoming more common in epidemiology and genetic epidemiology, in particular the meta-analysis of genetic association studies (Verzilli *et al.*, 2008). Meta-analysis has a number of advantages for Mendelian randomization analyses. It may be the case that for small effect sizes a meta-analysis is the only way to provide conclusive evidence on a particular research question. Also, under the ratio of coefficients approach the phenotype-disease association can be derived from gene-disease and gene-phenotype associations. Hence, Mendelian randomization meta-analyses could be undertaken on studies that did record all three outcomes and were not originally designed to be part of a Mendelian randomization analysis. Therefore, there is scope to investigate meta-analysis methods using the Mendelian randomization approach and this provides the motivation behind the research in chapters 5 and 6.

In conclusion, this review demonstrates that the Mendelian randomization approach represents an example of the use of causal analysis in observational studies. The main limitation in terms of statistical modelling approaches for Mendelian randomization analyses is that methods for the instrumental variable analysis of binary and categorical outcome data are not well developed, hence there is scope for investigating statistical models for both individual studies and meta-analysis in this area.

# Chapter 3

# An adjusted instrumental variable estimator: theory

## 3.1 Introduction

The aim of this chapter is to investigate instrumental variable models for the analysis of epidemiological studies such as case-control or cohort studies that report binary outcomes. Instrumental variable theory is well developed for continuous outcome variables and Mendelian randomization analyses can be applied to continuous outcome variables using the methods of two-stage least squares or the ratio of coefficients approaches. However when the outcome variable is categorical or binary as in case-control or cohort studies estimation is more problematic, prompting comments such as PHOEBE (2007),

> *However, the strong additional parametric assumptions such as linearity of all relationships and no interactions required for calculation of the average causal effect are usually not justifiable for epidemiological applications where a binary disease outcome is commonly of interest. In the non-linear / interaction case, even the specification of the causal parameter is not obvious and determination of its relationship to the relevant regression parameters that can be estimated*

*from the data is not straightforward..*

Instrumental variable theory has not been fully generalized to non-linear situations (Pearl, 2000) so the practical implications of such a violation of the core assumptions have not yet been clearly defined. The specification of the relevant causal parameter and identification of how it relates to what can be estimated in observational studies are not generally straightforward (Didelez & Sheehan, 2007b). However, the two-stage least squares approach has been adapted for a non-linear effect of the phenotype, an approach known as non-linear two-stage least squares (Amemiya, 1974). However, this method has limitations in its application to the analysis of case-control and cohort studies.

Two examples of Mendelian randomization studies reporting instrumental variable analysis of binary outcomes are given by Davey Smith *et al.* (2005a) and Qi *et al.* (2007). Both of these studies used the user written Stata program `qvf` (Hardin & Carroll, 2003a) (QVF: quasilikelihood and variance function). However, the estimator implemented in this program was proposed to use instrumental variables to correct for measurement error and it has not been demonstrated that it is the optimal method for a Mendelian randomization analysis (Carroll *et al.*, 1995, Chapter 5), and unusually for a user-written Stata program the source code is not available.

The first section of this chapter explains the method of two-stage least squares, the correction that is required to the standard errors at the second stage and the method of non-linear two-stage least squares. The next section outlines an adjusted estimator for binary outcome studies that is described as an instrumental variable estimator because it follows the control function approach. The expected parameter values from the adjusted and two other estimators are discussed in terms of the theory relating marginal and conditional parameter estimates from generalized linear mixed models (GLMMs). The adjusted IV estimator is also discussed in terms of causal inference and in particular with respect to Pearl's back-door criterion and the literature on adjusting for non-compliance in RCTs.

## 3.2 Two-stage least squares

This section outlines the derivation, following Cameron & Trivedi (2005), of the two-stage least squares approach for continuous outcome variables and the correction required to the standard errors of the parameter estimates. This theory is relevant because it informs subsequent discussion about the adjusted IV estimator.

Denoting a continuous outcome variable $Y$, phenotype variable $X$ and instrumental variable $Z$, where these are $(n \times 1)$ vectors where $n$ is the number of observations. In two stage least squares interest lies in recovering the parameter $\beta$ from (*iid*: independent and identically distributed),

$$y_i = \beta x_i + \nu_i, \quad \nu_i \overset{iid}{\sim} N(0, \sigma^2), \ i = 1 \dots n. \tag{3.1}$$

Typically parameters could be estimated using ordinary least squares (OLS). The motivation for the instrumental variable approach comes from the situation where the $x_i$ and $\nu_i$ are correlated. In this situation the OLS estimate of $\beta$, $\widehat{\beta}$, is biased because,

$$\widehat{\beta} = \beta + \frac{\sum_i x_i \nu_i}{\sum_i x_i^2}. \tag{3.2}$$

For the above expression when $X$ and $\nu$ are uncorrelated the OLS estimator is unbiased and consistent, when $X$ and $\nu$ are correlated the OLS estimator is biased and inconsistent. Taking expectations conditional on the instrumental variable $z$ gives,

$$E(Y|Z) = \beta E(X|Z) + E(\nu|Z) \tag{3.3}$$

which implies that,

$$\widehat{\beta}_{IV} = \beta + \frac{\sum_i z_i \nu_i}{\sum_i z_i x_i}. \tag{3.4}$$

As $Z$ and $\nu$ become independent as the sample size increases, therefore instrumental variable estimators are consistent for valid instrumental variables.

### 3.2.1 Econometric instrumental variable conditions

It should be noted that one of the 'core' conditions for an instrumental variable is defined slightly differently in econometrics. In Chapter 1 the first condition for a variable to be an instrument was expressed as: the instrument must be associated with the phenotype, or that the instrument is correlated with the phenotype, i.e. $\text{cov}(Z, X) \neq 0$. In econometrics the preferred way to express this condition is that the instrument should be partially correlated with the phenotype (Wooldridge, 2002, page 84). In econometrics the conditions for $Z$ to be an instrumental variable are given as:

(i) $\text{cov}(z_i, \nu_i) = 0$,

(ii) $\alpha \neq 0$ where $x_i = \alpha z_i + e_i$ with $E(e_i) = 0$.

The second condition here is the equivalent of the first condition from Chapter 1. This econometric form of the condition is the reason it is possible to include other exogenous variables in both stages stage of the two stage least squares procedure (Baltagi, 1998, p 278). As given here the first condition is not testable because the error term, $\nu$, is unobserved whereas it is possible to test the second condition.

### 3.2.2 Derivation of the two-stage least squares estimator

The first stage of two-stage least squares regresses the phenotype, $X$, on the instrument, $Z$, to generate the predicted values of the phenotype $\widehat{X}$. At the second stage the disease outcome, $Y$, is regressed on the predicted levels of the phenotype $\widehat{X}$. Given that $Z$ fulfills the conditions to be an instrumental variable, the system of equations defined by Equation 3.1 and condition (ii) above is identified because $\beta$ can be expressed in terms of the expectations of the other variables.

To follow Wooldridge (2002), writing Equation 3.1 in matrix form and then multiplying

through by Z and taking expectations gives,

$$E(Z'Y) = \beta E(Z'X), \tag{3.5}$$

which has a unique solution if and only if $E(Z'X)$ has full rank. The solution is given by,

$$\beta = \left(E(Z'X)\right)^{-1} E(Z'Y) \tag{3.6}$$

$$\widehat{\beta} = (Z'X)^{-1}Z'Y \tag{3.7}$$

$$= \frac{(Z'Z)^{-1}Z'Y}{(Z'Z)^{-1}Z'X}.$$

Which can immediately be recognised as the Wald estimator, or ratio of coefficients approach. Replacing $E(Z) = \widehat{X}$ gives,

$$\widehat{\beta} = (\widehat{X}'X)^{-1}\widehat{X}'Y, \tag{3.8}$$

which is the two stage least squares estimator, where,

$$\widehat{X} = Z(Z'Z)^{-1}Z'X \tag{3.9}$$

In econometrics the matrix, $P$, is called the projection matrix where $P = Z(Z'Z)^{-1}Z'$ and therefore,

$$\widehat{X}'X = X'PX = (PX)'PX = \widehat{X}'\widehat{X} \tag{3.10}$$

and hence Equation 3.6 becomes,

$$\widehat{\beta} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'Y \tag{3.11}$$

which is the expression that would be expected given the name two stage least squares. The two stage least squares estimate is consistent which means it converges in probability to the true value. However, a finite sample bias may occur when the instrument is weak or when the instrument is not strictly independent of the error terms in Equation 3.1.

### 3.2.3 Correction of the standard errors

After the second stage of estimation in two-stage least squares the standard errors of the parameter estimates need to be corrected because predicted values of $X$, $\widehat{X}$, were used in the second stage rather than observed values (Wooldridge, 2002, page 91). This was shown graphically by Thomas *et al.* (2007, Figure 1) who plotted the gene-disease outcome ($y$-axis) versus the gene-phenotype outcome ($x$-axis) and showed the need for error bars in both dimensions. Wiggins (2000) explains that the correction to the standard errors applies the correct mean squared error to the variance-covariance matrix of the two-stage least squares parameter estimates. Indeed that the variance-covariance matrix of the parameter estimates from the second stage of two-stage least squares needs adjustment was noted by Pagan (1984, top of page 227).

Further, following Wooldridge (2002, page 95) there is an important difference between the residuals from the second stage regression, $\nu_i$, and what are known as the two-stage least squares residuals, denoted $\widehat{\nu}_i$. The two-stage least squares residuals are not just the predicted values of the residuals from the second stage regression. The second stage residuals use the predicted values of $X$ whereas the two-stage least squares residuals use the observed values of $X$ such that,

$$\nu_i = y_i - \widehat{x}_i\widehat{\beta}, \tag{3.12}$$

$$\widehat{\nu}_i = y_i - x_i\widehat{\beta}. \tag{3.13}$$

Following Baltagi (1998, page 280), Wiggins (2000) and Wooldridge (2002, pages 139–141) an estimator of the asymptotic variance-covariance matrix of $\widehat{\beta}$ is given by,

$$\widehat{S} = \widehat{\sigma}^2 \left( \sum_{i=1}^{n} \widehat{x}_i' \widehat{x}_i \right)^{-1}. \tag{3.14}$$

where $\sigma^2$ is the variance of the $\nu_i$ and $\widehat{\sigma}^2$ is given by,

$$\widehat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\widehat{\nu}_i^2, \tag{3.15}$$

where $k$ is the number of instrumental variables. In fact it is debatable whether a degrees of freedom correction should be made to this variance since the results for IV estimation are asymptotic. So alternatively $n$ could be used as the denominator, however the above form is conservative and hence generally preferred.

Baum (2006, p189) states that inferences drawn using the second stage residuals $\nu$ will be inconsistent since the predicted values of the phenotype $\widehat{X}$ are not true explanatory variables. A full mathematical explanation of this is given by Greene (2008, p317–319) who states that estimates of $\sigma^2$ based on $\widehat{\nu}$ are consistent since if,

$$\widehat{\nu} = Y - X\widehat{\beta} \tag{3.16}$$

$$= Y - X(Z'X)^{-1}Z'Y \tag{3.17}$$

$$= \left(I - X(Z'X)^{-1}Z'\right)\nu \tag{3.18}$$

then,

$$\widehat{\sigma}^2 = \frac{\widehat{\nu}'\widehat{\nu}}{n} \tag{3.19}$$

$$= \frac{\nu'\nu}{n} + \frac{\nu'Z}{n}\left(\frac{X'Z}{n}\right)^{-1}\left(\frac{X'X}{n}\right)\left(\frac{Z'X}{n}\right)^{-1}\left(\frac{Z'\nu}{n}\right) - 2\left(\frac{\nu'X}{n}\right)\left(\frac{Z'X}{n}\right)\left(\frac{Z'\nu}{n}\right). \tag{3.20}$$

and the second and third terms in the above expression converge to 0 since,

$$\text{plim}\left(\frac{Z'\nu}{n}\right) = \text{plim}\left(\frac{Z'Y}{n}\right) - \text{plim}\left(\frac{Z'X\beta}{n}\right) = 0, \tag{3.21}$$

where plim() denotes converges in probability to. Convergence in probability means that for a sufficiently large sample size there is a very high probability that the term in the

brackets will be a certain value and a consistent estimator is one which converges in probability to its true value. Therefore, $\widehat{\sigma}^2$ based on $\widehat{\nu}$ is consistent. This also concurs with Davidson & MacKinnon (2004, p324) who state that the estimators of $\sigma^2$ based on the different residuals are not asymptotically equivalent and that the unadjusted residuals do not tend to the true IV residuals as the sample size increases. The correction to the standard errors of two-stage least squares estimates is used in the `tsls` function in the `sem` package in R (Fox, 1979, 2008) and in the `ivregress` and `ivreg2` commands in Stata.

### 3.2.4 Murphy-Topel standard errors

Murphy & Topel (1985) proposed a correction to the covariance matrix of the parameter estimates after the second stage of a two-stage modelling procedure in which the first stage model is linked to the second stage model. The rationale for this correction is the same as the correction applied to the standard errors after the second stage of two-stage least squares in that they argued that the second stage estimate of the covariance matrix of the parameter estimates needs to be corrected for the variability in the predicted values of the phenotype incorporated in the second stage.

The Murphy-Topel correction to the standard errors of a two-stage modelling process has been described by Hardin (2002) and implemented in the user-written command `qvf` (Hardin & Carroll, 2003a,b). The calculation of Murphy-Topel standard errors for two-stage models using standard statistical software has also been described by Hole (2006) who demonstrated a simplified procedure to calculate the Murphy-Topel variance covariance matrix of a two stage model for the following pairs of models used at the first and second stages respectively; logistic-Poisson, logistic-probit, probit-Poisson, linear-Poisson, logistic–negative-binomial. It can also be noted that since the third edition of his textbook Greene (2008) has noted that the Murphy-Topel covariance matrix is the asymptotic covariance matrix of two-stage maximum likelihood models.

The Murphy-Topel correction to the variance-covariance matrix of the second stage parameter estimates is complex and involves the first stage model's score function evaluated

at the first stage parameter estimates. Hardin (2002) notes that the Murphy-Topel covariance matrix has a similar form to the sandwich estimate of the second stage covariance matrix since two-stage models are a special case of partial M-estimators (Binder, 1983; Huber, 1967; Stefanski & Boos, 2002). The correction effectively results in an inflation of the second stage standard errors which is affected by the functional form of the first stage model. In particular, if linear regression is used for the first stage model then the Murphy-Topel correction inflates the standard error of all the second stage parameter estimates by the same factor, which is not the case for other models used at the first stage.

### 3.2.5 Weak instruments

Lawlor *et al.* (2008d, Section 4.10) note that investigating weak instruments using the $F < 10$ condition requires careful interpretation. This is because for a first stage regression with a single instrument an F = 10 implies a $P = 0.0015$ for the $t$-test of the coefficient on the instrument. Hence, for a single instrument there is a range, $0.0015 < P < 0.05$ for which the instrument is statistically significant in the first stage but yet still weak.

Indeed Staiger & Stock (1997) note that $F < 10$ may not be conservative enough if there are many more instrumental variables than intermediates. And if the instruments are weak but there are no other variables available to use then Mikusheva & Poi (2006) and Andrews *et al.* (2007) argue that it may be more appropriate to use estimation methods such as limited information maximum likelihood (LIML) which may have better finite sample properties.

The problem with weak instruments is that the parameter estimate can be biased and also that the associated standard error can be larger than the ordinary least squares regression, whose parameter estimate is also biased. Therefore, Wald tests of $\beta_{IV}$ can be wrong. In finite samples for weak instruments $\beta_{IV}$ may not be centred of the true value of $\beta$ even though $\beta_{IV}$ may be a consistent estimator. How large the sample size should be before the finite sample bias is negligible does not have a simple answer.

In particular, when are many weak instruments heuristically it might be assumed that using all available instruments should provide the most efficient instrumental variables estimate. However, in practice this may not be the case since the finite sample bias of the instrumental variables estimate can increase with the number of instruments. In the case where there are more instruments than intermediates the Hansen test can be performed to test the validity of these over-identifying instruments (Cameron & Trivedi, 2009; Hansen, 1982). When performing this test the instrumental variables estimate should be estimated using the generalized method of moments technique.

Additionally, when the IV assumptions hold assuming a continuous outcome measure it can be shown that the correlation between the instrument and phenotype is given by (Martens *et al.*, 2006),

$$|\rho_{Z,X}| = \sqrt{1 - \frac{\rho^2_{X,\nu}}{\rho^2_{\nu,\epsilon}}}. \tag{3.22}$$

This implies that there is a limit on the strength of the instrumental variable and that the strength of the instrumental variable decreases as the variance of the error term increases.

### 3.2.6 Non-linear two stage least squares

The theory for two-stage least squares regression has been generalized to allow the recovery of the instrumental variable estimate when there is a non-linear association between the phenotype and the outcome. This is referred to as the non-linear two-stage least squares approach (Amemiya, 1974; Bowden & Turkington, 1981). This method is appropriate for the situation when there is a non-linear function of the phenotype in the second stage regression. For example to follow Cameron & Trivedi (2005, page 198–199) the non-linear two-stage least squares approach can recover a consistent estimate of $\beta$ in the following system of equations,

$$x_i = \alpha z_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \tag{3.23}$$

$$y_i = g(X\beta) + \nu_i, \quad \nu_i \sim N(0, \sigma^2) \tag{3.24}$$

where $g()$ is some non-linear function.

Similarly to two-stage least squares, the standard errors of the parameter estimates at the second stage of non-linear two-stage least squares should be corrected using a similar correction to that discussed in Section 3.2.3.

However, in the non-linear two-stage least squares approach does not appear to be appropriate for binary or categorical outcomes since the variability is not modelled in terms of the probability of the different categories of the outcome variable because the outcome measure is still assumed to be continuous. Therefore, different approaches to non-linear two stage least squares are investigated in the next section.

### 3.2.7 Generalised Method of Moments

The moment conditions are based around the assumption that the instrument ($Z$) should be independent of the error term in the association between the phenotype ($X$) and the disease ($Y$). Where $h()$ is the inverse link function of the corresponding generalised linear model the GMM moment conditions as given by Johnston *et al.* (2008) are formed by assuming,

$$Y = h(X'\beta) + \epsilon, \tag{3.25}$$

where $E(\epsilon) = 0$. Then the moment condition is given by,

$$\begin{aligned} E\left(Z\epsilon\right) &= 0 \\ \Rightarrow E\left(Z(Y - h(X'\beta))\right) &= 0. \end{aligned} \tag{3.26}$$

Therefore, for a log-link the GMM moment condition is given by,

$$E\left(Z(Y - \exp(X'\beta))\right) = 0. \tag{3.27}$$

Due to the functional form of $\epsilon$ in Equation 3.25 these moment conditions are sometimes referred to as additive moment conditions. For the GMM equivalent of Poisson regression an alternative to the additive GMM moment condition has been proposed by Mullahy (1997). Mullahy's moment condition is termed as a multiplicative moment condition because it assumes that there is an exponentially distributed error term that therefore has an additive effect in the linear predictor. The multiplicative moment condition of Mullahy (1997) is derived for the log link by dividing Equation 3.25 by $\exp(X'\beta)$ so,

$$\begin{aligned} Y\exp(-X'\beta) &= 1 + \epsilon\exp(-X'\beta) \\ \Rightarrow \epsilon\exp(-X'\beta) &= Y\exp(-X'\beta) - 1. \end{aligned} \tag{3.28}$$

It is this term, $U = Y\exp(-X'\beta) - 1$, that is then used as the residual in the moment condition such that,

$$E(ZU) = 0 \tag{3.29}$$

$$\Rightarrow E\left(Z(Y\exp(-X'\beta) - 1)\right) = 0 \tag{3.30}$$

$$\Rightarrow E\left[Z\left(\frac{Y - \exp(X'\beta)}{\exp(X'\beta)}\right)\right] = 0. \tag{3.31}$$

Windmeijer & Santos Silva (1997) argued that if $X$ is endogeneous then either $E(ZU) = 0$ or $E(Z\epsilon) = 0$, however these statements cannot both be true. The implication of this is that the argument in the second half of Johnston *et al.* (2008, Appendix A) in which the multiplicative moment condition is approximated by a Taylor series expansion and shown to be approximately equivalent to the additive moment condition may be misleading for some situations. The multiplicative moment condition has recently been implemented in a user written program for Stata called `ivpois` (Nichols, 2007b). This moment condition is

closely related to the structural nested mean models approach of Hernán & Robins (2006) and its use would estimate a causal relative risk parameter for a cohort study.

An approach to estimate $\beta$ from these moment conditions is to minimize the squared moment condition using a Newton-type algorithm.

## 3.3 Statistical models for a binary outcome Mendelian randomization study

This section outlines three estimators for use with a binary outcome Mendelian randomization study. The three estimators are termed direct, standard IV and adjusted IV. The direct estimator is the conventional epidemiological approach, the second is the estimator that might be used following the principle of two-stage least squares using an appropriate GLM at the second stage and the third is an adjustment to this. The work in this section and the simulation results in the next chapter are described in Palmer *et al.* (2008a) which is included at the end of the thesis.

### 3.3.1 Estimators for Mendelian randomization studies with binary responses

The key variables in describing the Mendelian randomization model are; the disease status $Y$, intermediate phenotype $X$, genotype $G$ and confounder $U$. The assumed relationship between these variables is shown in Figure 3.1.

For the $i^{th}$ subject in a cohort let: $y_i$ represent the binary disease status, $p_i$ represent the probability of having the disease, $x_i$ represent the level of the biological phenotype and $g_i$ represent the genotype, which is coded 0, 1 and 2 to indicate the number of copies of the relevant risk allele. Typically there will be many unmeasured confounders, so it is assumed that they can be represented by a single variable, $U$, that captures their combined effect. This confounding variable is arbitrarily assumed to be standardised to have a mean of

zero and a standard deviation of one. For simplicity, an additive effect of the genotype on the intermediate phenotype is assumed, although the following arguments would apply equally to any known mode of inheritance. It is also assumed that the confounder acts additively in the linear predictors of the associations between the genotype and phenotype and between the phenotype and the disease.

The coefficients in the regression of phenotype on genotype are denoted by $\alpha$*'s* so that,

$$x_i = \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i, \text{ with } \epsilon_i \sim N(0, \sigma^2_\epsilon), \tag{3.32}$$

and $\epsilon$ represents the effects of measurement error and unmeasured factors that are not confounders because they do not influence disease. The coefficients in the linear predictor between phenotype and disease are denoted by $\beta$*'s*, so that the disease status follows a Bernoulli distribution,

$$y_i \sim \text{Bern}(p_i), \text{ with } \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i + \beta_2 u_i. \tag{3.33}$$

Implicit in the notation is the idea that $\epsilon_i$ and $u_i$ are independent of one another. The primary interest in this framework is to recover the phenotype-disease log odds ratio, $\beta_1$.

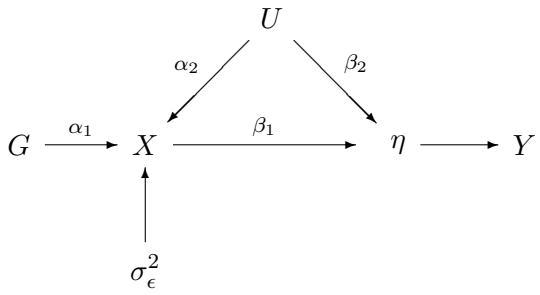



Figure 3.1: The relationship between the variables ($\eta$ is the linear predictor of the logistic regression).

If both regressions were linear, ignoring the confounder in the instrumental variable analysis would not bias the estimate of $\beta_1$ although it would change $\beta_0$, but this is not the case

for a non-linear relationship between phenotype and disease (Didelez & Sheehan, 2007b). Substituting the formula for $x_i$ in Equation 3.32 into the logistic regression in Equation 3.33 gives,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1(\alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i) + \beta_2 u_i. \tag{3.34}$$

The coefficient of $g_i$ in this relationship is $\beta_1\alpha_1$ while the coefficient of $g_i$ in the linear regression in Equation 3.32 is $\alpha_1$. In principle the ratio of the estimates of these coefficients should give an estimate of $\beta_1$ (Thomas & Conti, 2004), which is the effect of the phenotype on disease risk after adjusting for confounding. Unfortunately $u_i$ and $\epsilon_i$ are unknown, so the estimate of $\beta_1\alpha_1$ is taken from the logistic regression without those terms, thus in effect replacing the true conditional model with a marginal model which averages over the unknown terms, $u_i$ and $\epsilon_i$.

An alternative to the ratio estimate of $\beta_1$ is obtained by taking the predicted values of the intermediate phenotype from the first regression ignoring the confounding,

$$\widehat{x}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 g_i \approx \alpha_0 + \alpha_1 g_i \tag{3.35}$$

and substituting those into the logistic regression in (3.33), in which case,

$$\log \frac{p_i}{1-p_i} \approx \beta_0 + \beta_1(\widehat{x}_i + \alpha_2 u_i + \epsilon_i) + \beta_2 u_i. \tag{3.36}$$

In this two-stage approach the estimate of interest is just the coefficient of the predicted phenotype $\widehat{x}_i$, but the biases will be similar to those that occur for the ratio estimator.

In an attempt to correct for this difference between marginal and conditional parameter estimates, and thus improve upon the standard instrumental variable estimator an adjusted IV estimator is applied. The estimated residuals from the first stage linear regression in Equation 3.32 are,

$$r_i = x_i - \widehat{x}_i. \tag{3.37}$$

These estimated residuals capture some of the variability contained in the unknown con-

founder, $u_i$, and the phenotype error term, $\epsilon_i$. This information can be used in the second regression by fitting,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \widehat{x}_i + \beta_r r_i. \tag{3.38}$$

The information about the confounding contained in the residuals should, in part, compensate for the missing terms in the marginal form of the logistic regression model and therefore reduce the difference between the conditional and marginal estimates of $\beta_1$.

Three estimators of $\beta_1$ are considered, first the direct estimator, that does not use Mendelian randomization but performs a logistic regression of disease status on the intermediate as in a traditional epidemiological study. The direct estimator of $\beta_1$ is derived from the logistic regression of the disease status variable on the phenotype,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i. \tag{3.39}$$

The standard IV estimator uses Mendelian randomization and is the logistic regression of the disease status on the predicted phenotype,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \widehat{x}_i. \tag{3.40}$$

The third estimator is the adjusted IV estimator obtained from the logistic regression of the disease status on the predicted phenotype and the first stage residuals as in Equation 3.38.

### 3.3.2 Wooldridge's procedure 15.1

Procedure 15.1 of Wooldridge (2002, page 274) which is the two-stage conditional maximum likelihood procedure of Rivers & Vuong (1988) is essentially the same as the adjusted IV estimator described above except that the observed values of the phenotype are used instead of the predicted values. These authors also used a probit regression at the second stage, replacing this with a logistic regression the second stage of procedure 15.1 takes the

following form;

$$\log \frac{p_i}{1-p_i} = \gamma_0 + \gamma_1 x_i + \gamma_2 r_i \tag{3.41}$$

$$= \gamma_0 + \gamma_1 \widehat{x}_i + (\gamma_1 + \gamma_2) r_i. \tag{3.42}$$

Comparing this formulation with the adjusted IV estimator gives,

$$\gamma_1 = \beta_1 \tag{3.43}$$

$$\gamma_2 = \beta_r - \beta_1. \tag{3.44}$$

Wooldridge (2002) states that the usual $t$-statistic testing the null hypothesis that $\gamma_2 = 0$ is a valid test of the presence of confounding. The equivalent test for the adjusted IV estimator is therefore a test of the null hypothesis that $\beta_1 = \beta_r$.

The use of the first stage residuals was also suggested by Nitsch *et al.* (2006) who commented that,

> *Another, equivalent way to obtain IV estimates is to save the residuals from the regression of X on Z and then include them in the regression of Y on X. Such residuals act as unbiased estimates of the unmeasured confounders in U and therefore lead to unbiased estimates of the causal effect from X to Y only if the regression model for the regression of X on Z is appropriately specified, however. If it is not, biased estimates will be obtained.*

## 3.4 Theoretical values of the three estimators

This section describes the theoretical values of the parameters from the direct, standard IV and adjusted IV estimators. Firstly, as background the parameter estimates from a threshold model are discussed.

### 3.4.1 Parameter estimates from Maddala's threshold model

Maddala (1983, page 244) considers a system of two linear structural equations, the second outcome variable of which is observed as dichotomous, subject to the original continuous outcome exceeding some threshold (i.e. $Y$ is observed as 1 if the value of the linear predictor exceeds a certain value and 0 otherwise). Maddala notes that from this two-stage estimation procedure the estimates from the second equation are the true parameter values divided by the standard deviation of the error terms from the second equation. Wooldridge (2002) also notes that the parameter estimates from procedure 15.1, as described in Section 3.3.2, are scaled by the square root of a residual variance term.

### 3.4.2 Theory from GLMs with random coefficients

To investigate the theoretical values of the three estimators the theory relating population and subject specific parameter estimates in generalised linear models with a random intercept, a form of generalised linear mixed model (GLMM), is relevant. This is because in each of the three estimators some terms are omitted with respect to the full model and these omitted terms act as a random effect. With respect to the full model for the direct estimator the confounder is omitted, for the standard IV estimator the confounder and the phenotype error term is omitted and for the adjusted IV estimator the phenotype error term is omitted.

In the following the subject specific, or conditional, parameter estimate is denoted by $\beta_c$ and the population averaged, or marginal, parameter estimate is denoted by $\beta_m$. The relationships between marginal and conditional parameter estimates for generalized linear models with a normally distributed random intercept were given by Zeger *et al.* (1988). These relationships can be described as follows; for models with an identity link function the marginal and conditional parameter estimates are the same. For models using a log link, such as Poisson regression, the marginal intercept is offset but the other coefficients in the linear predictor are identical. For models with probit and logit links all the parameters

in the linear predictor are attenuated towards the null effect. For the logit link, assuming the random effects are normally distributed, this relationship can only be approximated and is given by,

$$\beta_m \approx \beta_c \frac{1}{\sqrt{1 + c^2 V}}, \quad \text{with } c = \frac{16\sqrt{3}}{15\pi}, \tag{3.45}$$

where $V$ is the variance of the random effect, or for the three estimators described here the variance of the terms over which the subject specific data is averaged to produce the marginal estimates. The derivation of these relationships between marginal and conditional parameter estimates for GLMs with a random intercept are explained in more detail in Appendix B.

Therefore, to apply Equation 3.45 it is necessary to derive $V$ for each of the three estimators. First, the logistic regression for the association between the phenotype and disease is approximated as a linear regression of the log odds ratio, $\eta = \log(p/(1-p))$ on the covariates and confounders (Thomas *et al.*, 2007). If the terms included in the linear predictor of the logistic regression are denoted by $T$ then the remaining variance after allowing for these terms will be given by,

$$V = \text{var}(\eta|T) = \text{var}(\eta) - \frac{\text{cov}(\eta, T)^2}{\text{var}(T)} \tag{3.46}$$

since $\eta$ and $T$ can both be assumed to be normally distributed (Anderson, 1958). From Equation 3.34,

$$\eta_i = \beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 g_i + (\beta_1\alpha_2 + \beta_2)u_i + \beta_1\epsilon_i \tag{3.47}$$

and because $u$ is standardised, it follows that

$$\text{var}(\eta) = (\beta_1\alpha_1)^2\text{var}(g) + (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2 \tag{3.48}$$

and $\text{var}(g)$ is $2q(1-q)$ where $q$ is the minor allele frequency.

### 3.4.3 The direct estimator

The direct estimator performs a logistic regression of disease on the intermediate phenotype. In this case $T = x_i$ where,

$$x_i = \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i \tag{3.49}$$

so,

$$\text{var}(T) = \alpha_1^2 \text{var}(g) + \alpha_2^2 + \sigma_\epsilon^2. \tag{3.50}$$

The covariance between the log odds and the terms in the linear predictor is given by,

$$\begin{aligned}\text{cov}(\eta, T) &= \begin{bmatrix}\alpha_1 & \alpha_2 & 1\end{bmatrix} \cdot \begin{bmatrix}\text{var}(g) & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma_\epsilon^2\end{bmatrix} \cdot \begin{bmatrix}\beta_1\alpha_1 \\ \beta_1\alpha_2 + \beta_2 \\ \beta_1\end{bmatrix} \\ &= \alpha_1^2\beta_1 \text{var}(g) + \alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_\epsilon^2. \end{aligned} \tag{3.51}$$

Hence for the direct estimator $V_{\text{direct}}$ can be formed using Equations 3.48, 3.51 and 3.50.

### 3.4.4 The standard IV estimator

For the standard IV estimator the log odds are regressed on the fitted values from the linear regression of the phenotype on the genotype. Thus $T \approx \alpha_0 + \alpha_1 g$ and,

$$\text{var}(T) = \alpha_1^2 \text{var}(g) \tag{3.52}$$

$$\text{cov}(\eta, T) = \alpha_1^2 \beta_1 \text{var}(g). \tag{3.53}$$

Hence for the standard IV estimator $V$ is given by,

$$V_{\text{standard}} = (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2. \tag{3.54}$$

### 3.4.5 The adjusted IV estimator

The adjusted IV estimator makes use of the estimated residuals, $r$, from the regression of the phenotype on genotype to capture some of the variance explained by confounding variables not included in the standard IV estimator. Therefore the value of $V$ is reduced compared with the standard IV estimator. For the adjusted IV estimator $V$ is given by,

$$\begin{aligned} V &= \text{var}(\eta|T) - \frac{\text{cov}(\eta|T,r)^2}{\text{var}(r)} \\ &= \text{var}(\eta) - \frac{\text{cov}(\eta,T)^2}{\text{var}(T)} - \frac{\text{cov}(\eta|T,r)^2}{\text{var}(r)}. \end{aligned} \tag{3.55}$$

If the confounder $U$ is standardized the estimated residuals and their variance are given by,

$$r_i = \alpha_2 u_i + \epsilon_i \tag{3.56}$$

$$\text{var}(r_i) = \alpha_2^2 + \sigma_\epsilon^2 \tag{3.57}$$

The covariance between the log odds given the phenotype information and the estimated residuals is given by,

$$\text{cov}(\eta|T,r) = \begin{bmatrix} \beta_1\alpha_2 + \beta_2 & \beta_1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & \sigma_\epsilon^2 \end{bmatrix} \cdot \begin{bmatrix} \alpha_2 \\ 1 \end{bmatrix} \tag{3.58}$$

$$= \alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_\epsilon^2. \tag{3.59}$$

Since $\text{var}(\eta|T) = V_{\text{standard}}$ from the standard IV estimator above, for the adjusted IV estimator,

$$V_{\text{adjusted}} = (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2 - \frac{(\alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_\epsilon^2)^2}{\alpha_2^2 + \sigma_\epsilon^2}. \tag{3.60}$$

Hence it is possible to apply Equation 3.45 for all three estimators. These expressions are tested through simulation in the next chapter. The theory given here is similar to the theory for the `ivprobit` program in the Stata manual (Stata Corp, 2007). The Stata manual used a simplified form of the first stage regression which allows the $V$ term to be expressed as the correlation between the instrument and the phenotype.

## 3.5 Causal inference for the adjusted IV estimator

This section outlines rationale for the adjusted IV estimator from the causal inference and clinical trials literature.

The adjusted IV estimator uses the estimated residuals as well as the predicted values from the first stage regression of the genotype on the phenotype as covariates in the second stage logistic regression between the phenotype and the disease outcome. This estimator was introduced in the context of using instrumental variable analysis to correct for non-compliance in clinical trials (Nagelkerke *et al.*, 2000).

In clinical trials treatment randomization can be used as an instrumental variable to control for confounding in the intention-to-treat (ITT) analysis caused by non-compliance of subjects to their randomized treatment. The use of treatment randomization in this way is described as the estimation of 'treatment efficacy' or estimating the 'effects of the treatment received'. This estimate differs from both ITT and per-protocol analyses, and has been referred to as the adjusted treatment received (ATR) estimate or the IV(ATR) estimate (Dunn & Bentall, 2007; Nagelkerke *et al.*, 2000).

It should be noted that Dunn & Bentall (2007) attributed the adjusted IV estimator to

Hausman (1978). Specifically Hausman (1978, Equation 2.18) describes the use of the first stage residuals in the second stage regression but its implications are not discussed for non-continuous outcome measures and the paper is complex.

### 3.5.1 Back-door paths on DAGs and Pearl's back-door criterion

Graphs as used in causal inference, and in particular DAGs were defined in Section 1.5. Graphical assumptions are qualitative and non-parametric because they do not imply the specific functional form of the relationships between the variables. A marginal association between two variables in a graph requires that there is an unblocked path between them. In a DAG there are only two kinds of unblocked path, directed paths and back-door paths through a shared ancestor. Therefore a marginal association between two variables in a DAG requires that there is either a causal pathway from one to the other or that they share a common cause. Confounding can be shown on a DAG if there is an unblocked path between an phenotype and an outcome that is not direct. In econometric terms a variable is endogenous if it has an arrow into it otherwise it is an exogenous variable.

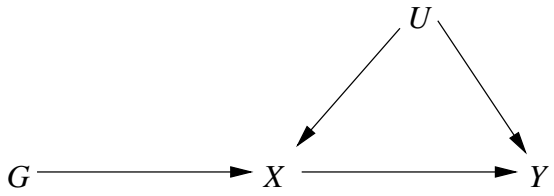



Figure 3.2: Typical DAG representing the use of genotype as an IV in a Mendelian randomization analysis, the genotype, phenotype, confounder and disease outcome variables are represented by $G, X, U$ and $Y$ respectively.

Figure 3.2 shows the DAG for a Mendelian randomization analysis using the genotype as an instrumental variable. This figure has appeared in Hernán & Robins (2006); Lawlor *et al.* (2008d); Thompson *et al.* (2003) and Didelez & Sheehan (2007b) and is the typical DAG used to represent an instrumental variable analysis. Valid instruments for the effect of $X$ on $Y$ can be used to test the null hypothesis that $X$ has no effect on $Y$.

The first stage residuals contain some information about the unmeasured confounder since they capture the variance in the phenotype that is not explained by the genotype. It has

been argued that the first stage residuals satisfy Pearl's back-door criterion (Nagelkerke *et al.*, 2000), which would mean that the adjusted IV estimate of $\beta_1$ would have a causal interpretation.

A back-door path on a DAG from $X$ to $Y$ is a path which begins at $X$ and whose first edge has an arrow pointing into $X$, the path should then end at $Y$ (Greenland & Brumback, 2002). In Figure 3.2 the path $X - U - Y$ is a back-door path from $X$ to $Y$. Following the notation of Nagelkerke *et al.* (2000), a variable $E$ satisfies the back-door criterion of Pearl (1995) relative to an ordered pair of variables $(X, Y)$ in a DAG if;

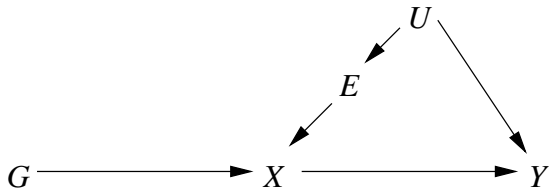
(i) no node in $E$ is a descendant of $X$, and

(ii) $E$ blocks every path between $X$ and $Y$ which contains an arrow into $X$.

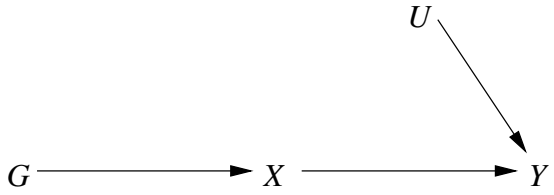
Hence an arrow between $E$ and $X$ must point into $X$ and in the case of Figure 3.2 $E$ must block the path between $X$ and $U$. Figure 3.3(a) shows the DAG for the adjusted IV estimator. It can be seen that the first stage residuals $E$ satisfy Pearl's back-door criterion since they are a parent of $X$ and block the path between $X$ and $U$. The first stage residuals only satisfy the back-door criterion if $G$ is a valid instrument since there must be no direct path between $G$ and $Y$ except through $X$. It is also important that $U$ is a true confounder of the $X - Y$ relationship and not for example an effect modifier of $Y$ (Nagelkerke *et al.*, 2000).

The DAG in Figure 3.3(a) has been reproduced by Dunn *et al.* (2005, Figure 2) and Keogh-Brown *et al.* (2007, Figure 34) from the original by Nagelkerke *et al.* (2000). Figure 3.3(b) shows the DAG after conditioning on the first stage residuals $E$. It shows that after conditioning there is only one edge connecting $X$ and $Y$ and hence the association between $X$ and $Y$ is no longer confounded.

In fact the DAG in Figure 3.3(a) and its corresponding moral graph have been drawn in the context of Mendelian randomization (Didelez & Sheehan, 2007a, Figure 1.3). These authors comment that in order to identify the average causal effect of $X$ on $Y$ that only one of $U$ or $E$ need be adjusted for in the analysis. This argument follows Dawid (2002),




(a) DAG for the adjusted IV estimator.




(b) DAG after conditioning on $E$.

Figure 3.3: DAGs for the adjusted IV estimator adapted from Figures 1 and 2 of Nagelkerke *et al.* (2000). $E$ represents the first stage residuals.

who argues it is only necessary to adjust for one of two related confounders.

It has been argued that the first stage residuals could be used in conjunction with many of the commonly used statistical models at the second stage of the analysis including linear, Poisson, logistic, probit and Cox regression models (Nagelkerke *et al.*, 2000). Dunn & Bentall (2007) also commented on the significance test mentioned by Wooldridge for the presence of confounding that is obtained from a test of the significance of parameter estimate from the first stage residuals. It is the case that if there are other known confounders they should be adjusted for at both stages of the standard IV and adjusted IV estimators, which is referred to in econometrics as including other exogenous variables at both stages of the analysis.

## 3.6 Discussion

This chapter has explained the econometric methods of two-stage least squares and non-linear two-stage least squares and that alternative methods are required for the analysis of case-control and cohort studies applying Mendelian randomization.

The adjusted IV estimator partially compensates for the unknown confounding factors by

exploiting information from the residuals of the regression of the intermediate phenotype on the genotype. The adjusted IV estimator is an alternative to other estimators termed direct and standard IV. The adjusted estimator is essentially equivalent to procedure 15.1 of Wooldridge (2002), and the theory used to explain the parameter estimates relates to the threshold model of Maddala (1983).

It is argued that for the direct estimator the confounder and additionally for the IV estimators the variability in the phenotype acts as a random effect and causes attenuation in the estimate of the phenotype-disease association $\beta_1$ that is analogous to the difference between marginal and conditional parameter estimates in generalized linear models with a random intercept (Breslow & Clayton, 1993; Zeger *et al.*, 1988). These formulae could be used to perform a sensitivity analysis under hypothesised levels of the unmeasured confounding. This point is analogous to the way the reliability ratio can be applied to parameter estimates from measurement error models whose parameter estimates are also attenuated (Carroll *et al.*, 2006, Chapter 3).

The adjusted IV estimator uses the first stage residuals to make the marginal likelihood of the model approximate the conditional likelihood of the underlying fully specified model. In this sense the adjusted IV estimator could be described as pseudo-conditional estimator.

The formulae for approximating this difference for the three estimators are assessed in the next chapter through simulations.

# Chapter 4

# An adjusted instrumental variable estimator: simulation study

## 4.1 Introduction

This chapter compares the properties of the direct, standard IV and adjusted IV estimators introduced in the previous chapter through a simulation study. Interest lies in assessing the bias and coverage of the estimators and the type I error of their respective significance tests.

### 4.1.1 Data simulation

In a cohort of 10,000 individuals, each individual was randomly assigned two alleles of a diallelic genetic variant (e.g. a SNP which has two alleles) in Hardy-Weinberg equilibrium. The allele frequency of the risk allele set to 30%. The confounding variable, $U$, was simulated to be normally distributed with mean 0 and variance equal to 1, $u_i \sim N(0, 1)$. The phenotype, $X$, was generated as a Normal random variable with mean equal to, $\alpha_0 + \alpha_1 g_i + \alpha_2 u_i$ following Equation 3.32, and the standard deviation of the phenotype error term, $\sigma_\epsilon$, was set to 1. Each subject's probability of disease was simulated, following

Equation 3.33 such that $\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i + \beta_2 u_i$.

The $\beta_0$ parameter was set to $\log(0.05/0.95)$. Different amounts of confounding were considered by changing the values of $\alpha_2$ and $\beta_2$. In particular four confounding scenarios were considered by setting the confounding effect on the phenotype, $\alpha_2$, to 0, 1, 2 and 3 whilst the confounding effect on the disease, $\beta_2$, was varied between 0 and 3 for each scenario. The other parameters were fixed as follows; $\alpha_0 = 0$, $\alpha_1 = 1$ and $\beta_1 = 1$. For each set of parameter values 10,000 simulations were performed. Statistical analysis was performed using the R software package (version 2.6.1) (R Development Core Team, 2008).

## 4.2 Simulation results for the logit link

The three estimators are assessed using the median parameter estimates, coverage probabilities and type I error of the Wald test of the phenotype-disease log odds ratio, $\beta_1$. The coverage probability of $\beta_1$ was calculated as the proportion of simulations whose confidence interval included the true value of $\beta_1$. A set of simulations was performed with $\beta_1$ equal to 0 to represent the situation in which there is no association between phenotype and disease. For those simulations, the proportion of statistically significant estimates from a significance test of $\beta_1$ is an estimate of the type I error of the test.

### 4.2.1 Bias

The value of $\beta_1$ was set to 1 in these simulations. Figure 4.1 shows the median of $\beta_1$ for the three estimators from the simulations, represented by the symbols, and the values of the estimators calculated from the formulae given in the previous chapter represented by the lines.

In Figure 4.1 the median values from the simulations are in good agreement with the theoretical predictions using Equation 3.45 from the previous chapter and the corresponding terms for $V$ for the different estimators. There is the same pattern to the estimates of $\beta_1$

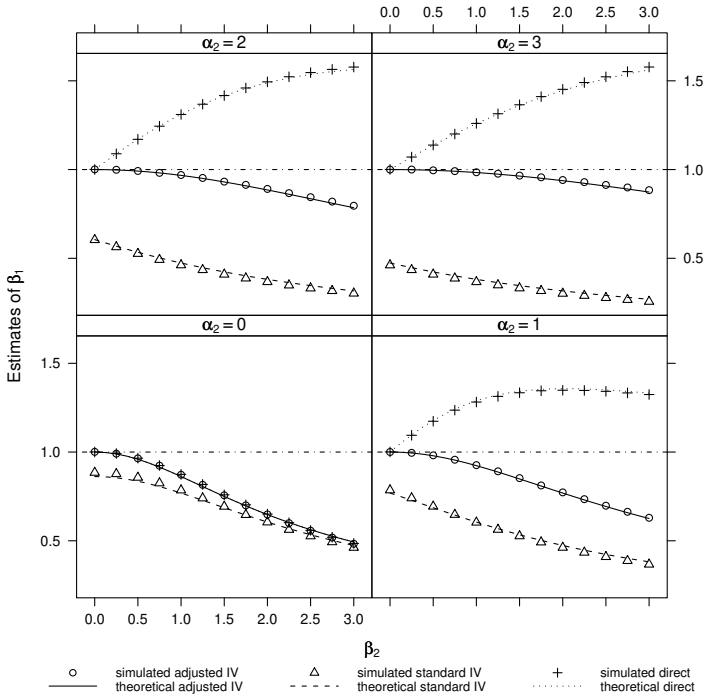



Figure 4.1: Simulated and theoretical estimates of $\beta_1$ for a true value of 1.

for the different values of $\alpha_2$ except when $\alpha_2$ is equal to zero. When $\alpha_2$ is equal to zero the direct and adjusted estimators are equivalent because in this instance the confounder no longer affects the phenotype and is therefore no longer a true confounder so the first stage residuals do not carry any information about the confounder. When $\alpha_2$ is non-zero, allowing the confounder to take effect, the direct estimate of $\beta_1$ is greater than the set value of 1. However, the effect the unmeasured confounding has on the standard IV estimates is to bias them towards zero, producing estimates that are always below the true value of 1. The values of the adjusted IV estimator are between the other two sets of estimates and have the smallest bias of the three estimators. For the adjusted IV estimates the bias in $\beta_1$ reduces with larger values of $\alpha_2$ because in these situations the estimated residuals are more informative since the confounder has a larger effect on the phenotype.

### 4.2.2 Coverage

Figure 4.2 shows the coverage probabilities of the three estimators. The nominal coverage level was 95% since 95% confidence intervals were derived for the parameter estimates. The direct estimator and the standard IV estimator demonstrate very low coverage for all four scenarios due to the bias in the estimates of $\beta_1$. The adjusted IV estimator demonstrates the best coverage properties with levels around 95% over the range of values of $\beta_2$ for which its estimate of $\beta_1$ was approximately equal to the set value of 1 in Figure 4.1. The coverage of the adjusted IV estimator improves as the value of $\alpha_2$ increases because the bias reduces.

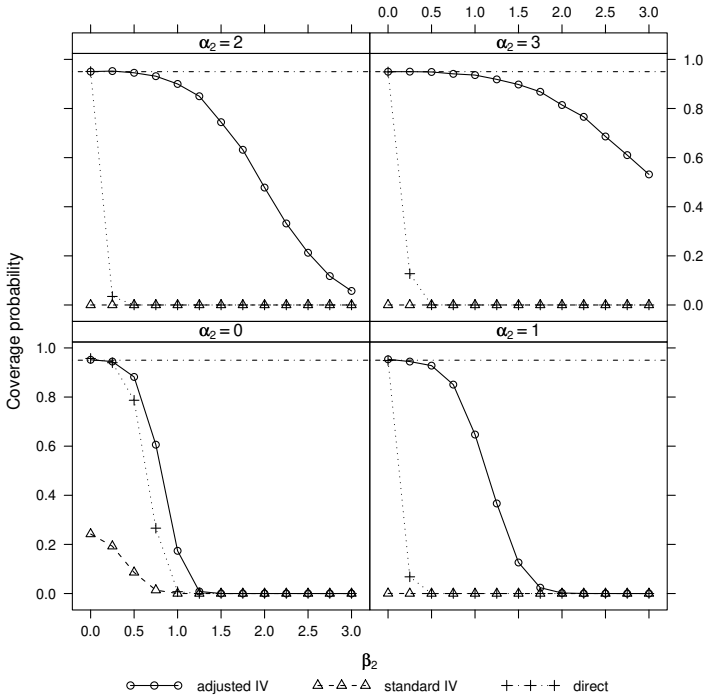



Figure 4.2: Coverage probabilities of the three estimators.

Figure 4.3 shows the coverage of the three estimators with respect to the marginal value of $\beta_m$ from the simulations. It shows that the direct estimator has the correct coverage with levels around 95% and the standard IV also had the expected or greater than expected levels of coverage. The adjusted IV estimator had the correct coverage levels when the confounder was not a true confounder and did not have an effect on the risk of disease.

However, when the confounder acted as a true confounder, in the three panels with $\alpha_2 \neq 0$ the coverage of the adjusted IV estimator was less than the expected level of 95% with respect to $\beta_m$.

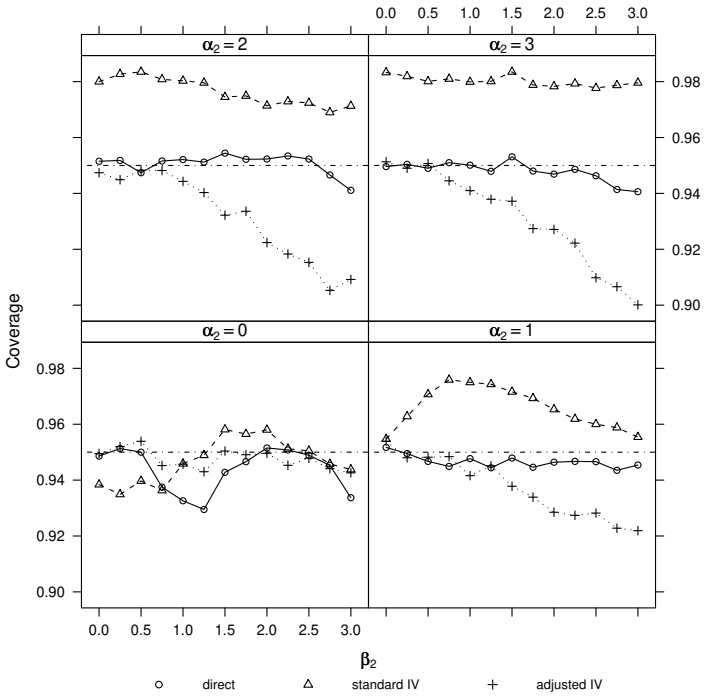



Figure 4.3: Coverage probabilities of the three estimators with respect to $\beta_m$.

### 4.2.3 Type I error

Figure 4.4 shows the type I error of the standard IV and adjusted IV estimators when the nominal rate is 5%. The type I error of the direct estimator is not shown on Figure 4.4 because the values were above 90% which was caused by the impact of the confounder. Under the three scenarios with non-zero values of $\alpha_2$ the adjusted IV estimator has a substantially higher type I error rate than the standard IV estimator because the inclusion of the estimated residuals in the adjusted IV estimator reduced its estimated standard error and hence produced larger $Z$ statistics in the Wald test.

That the adjusted IV estimator produces incorrect type I error rates is known in the econometrics literature. For example, Wooldridge (2002, p 474) states that when the

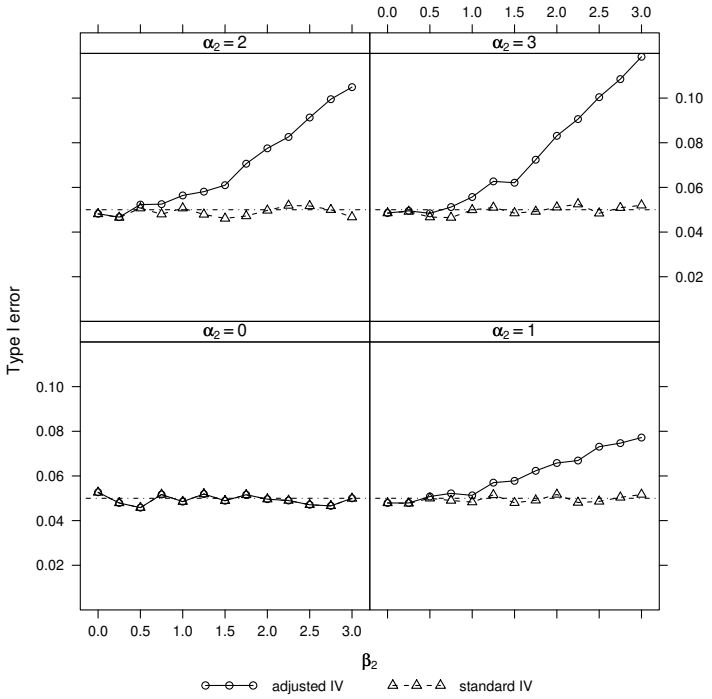



Figure 4.4: Type I error rate of the Wald test of $\beta_1$ for the IV estimators.

confounding is statistically significant then the resulting standard errors and test statistics for Wooldridge's equivalent of the adjusted IV estimator are not valid (as discussed in the previous chapter Wooldridge used a probit regression at the second stage), a point also noted by Nichols (2007a).

A possible explanation for the elevated type I error rate of the adjusted IV estimator could be that the Wald statistic for the test of $\beta_1 = 0$ may have a non-normal sampling distribution. To investigate this, the 2.5% and 97.5% quantiles of the $Z$-scores from the Wald test of $\beta_1$ are plotted for the standard and adjusted IV estimators in Figure 4.5.

In Figure 4.5 the observed quantiles for the standard estimator were at the expected values of $\pm 1.96$. The quantiles for the adjusted estimator were outside those for the standard estimator.

The Wald test is an approximation to the likelihood ratio test because it is based on a quadratic approximation to the log-likelihood (Pawitan, 2001). The type I error of the

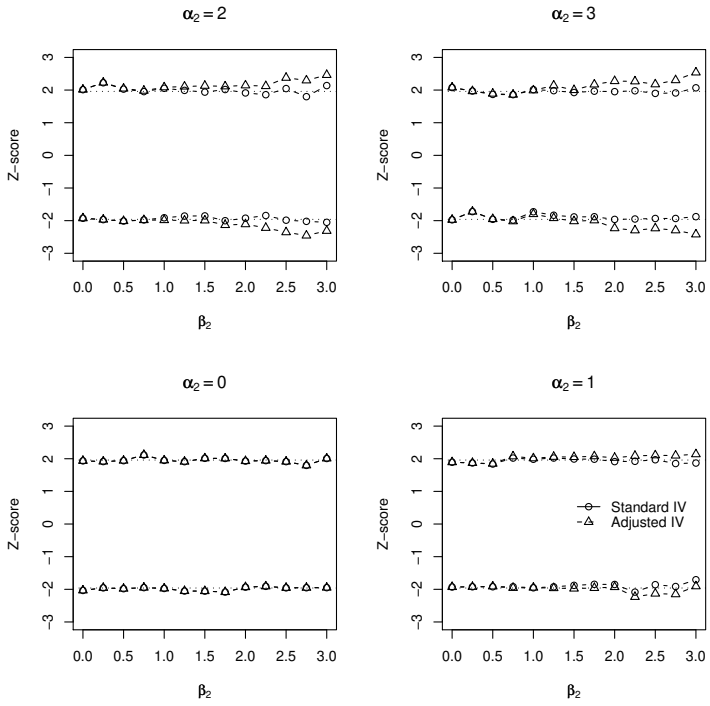



Figure 4.5: 2.5% and 97.5% quantiles of the Z-score for the standard and adjusted IV estimators.

likelihood ratio test for the estimators is shown in Figure 4.6. The likelihood ratio test of the standard IV estimator compared it to the null model. To perform the likelihood ratio test of $\beta_1$ for the adjusted IV estimator it was compared with a model containing only the first stage residuals in the second stage. The type I error of the likelihood ratio test of $\beta_1$ was also inflated for the adjusted IV estimator when the impact of the confounder was large as shown in Figure 4.6.

The use of a logistic regression model allowing for over-dispersion at the second stage was investigated to test whether it would reduce the type I error in the Wald test for the adjusted IV estimator. GLMs with over-dispersion can be fitted in R for binary outcomes by using the `quasibinomial` family option in the `glm` function. Figure 4.7 shows that using a logistic regression allowing for over-dispersion at the second stage did not reduce the type I error of the Wald test of $\beta_1$.

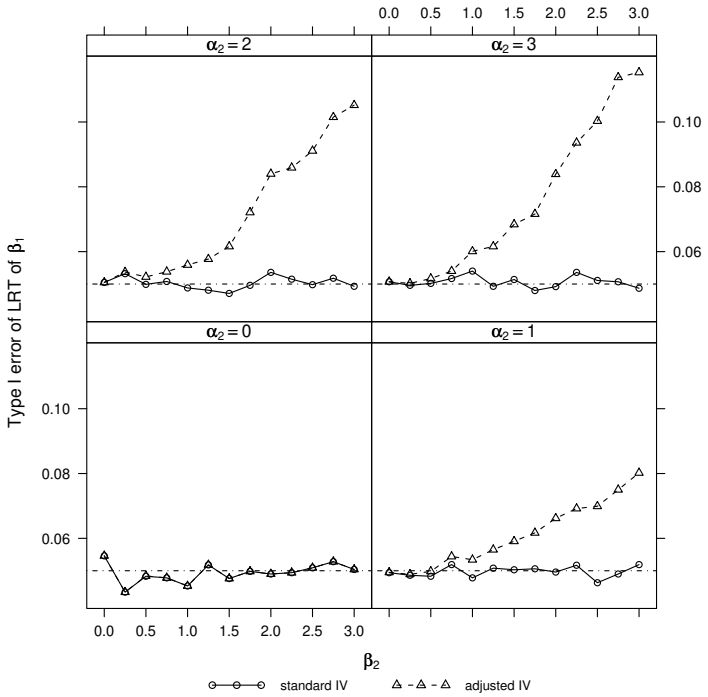



Figure 4.6: Type I error rate of the likelihood ratio test for two IV estimators of $\beta_1$.

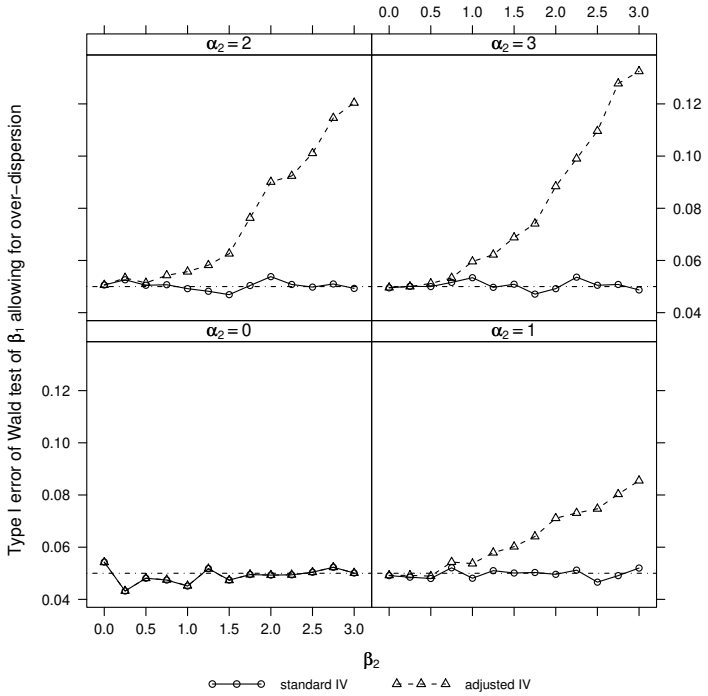



Figure 4.7: Type I error rate of the Wald test for the IV estimators of $\beta_1$ allowing for over-dispersion.

### 4.2.4 Intercept

The equations relating marginal and conditional parameter estimates also predict that the intercept in the second stage logistic regression will be shrunk towards the null. As explained in Appendix B this would also be the case using probit regression at the second stage. The median of the estimates of $\beta_0$ from the simulations are shown in Figure 4.8. The median of the simulated values of the intercept are in good agreement with the theoretical values.

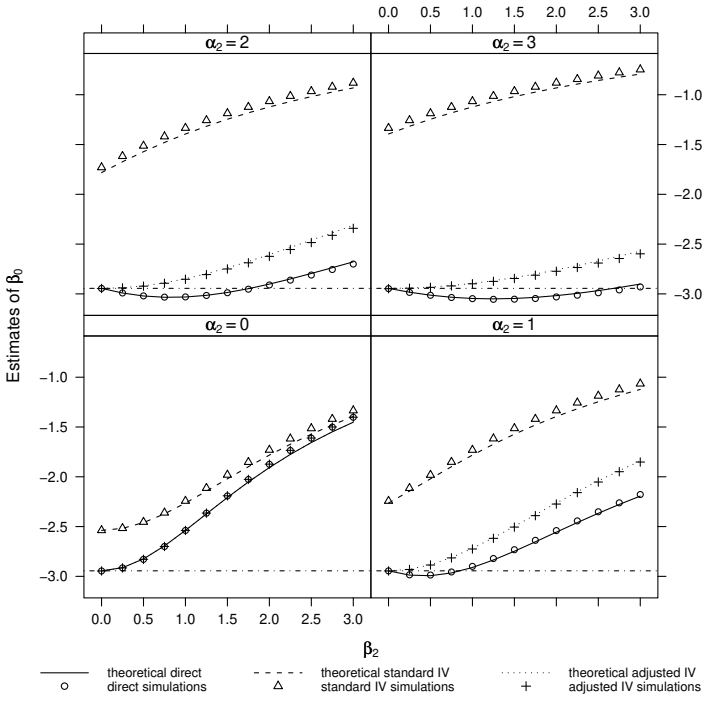



Figure 4.8: Theoretical and simulated estimates of $\beta_0$ with the logit link.

### 4.2.5 Proportion of variance due to the confounder

Figure 4.9 shows the proportion of variance explained by the confounder in the linear predictor of the first and second stage regressions. When $\alpha_2 = 0$, 1, 2 and 3, the confounder accounted for approximately 0, 45, 80 and 90 percent of the phenotype variance. For the log odds of disease the confounder accounted for between 0 and 90 percent of the variance

in the linear predictor when $\alpha_2 = 0$ and $\beta_2$ varied from 0 to 3, between 45 and 90 percent when $\alpha_2 = 1$, between 80 and 90 percent when $\alpha_2 = 2$ and between 85 and 95 percent when $\alpha_2 = 3$.

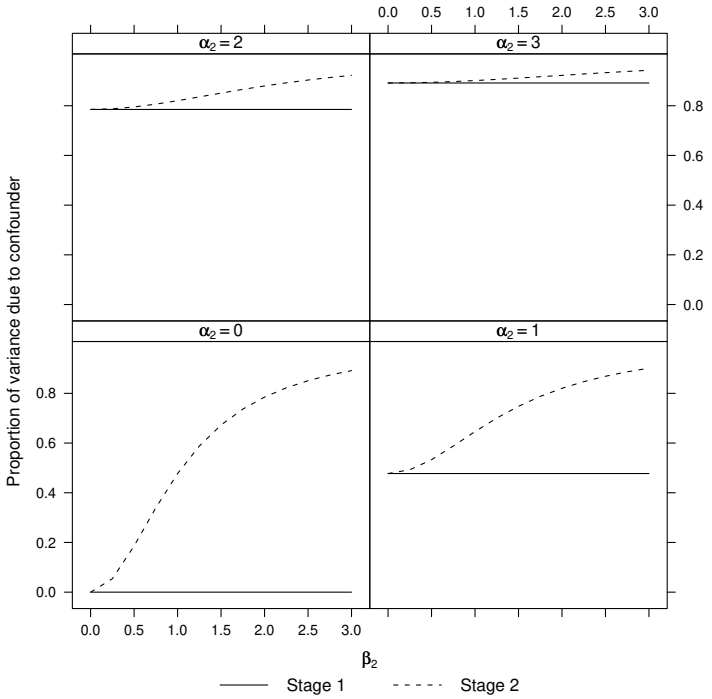



Figure 4.9: The proportion of variance due to the confounder in the stage 1 and 2 linear predictors.

### 4.2.6 Assessment of instrument strength through the first stage $R^2$

The genotype is categorical and in these simulations it is used to predict a continuous phenotype. It is therefore possible the first-stage regression may have low predictive power, since under the additive genetic model there are only three levels of the genotype with which to explain the variability in the phenotype. Under the recessive or dominant genetic models there are only be two levels of the genotype, so the first stage regression may explain less of the variability in the phenotype in this instance.

If an instrumental variable has a low $F$-statistic, typically less than 10, at the first-stage it is referred to as a weak instrument (Lawlor *et al.*, 2008d; Staiger & Stock, 1997). The

$F$-statistic from a linear regression is closely related to the coefficient of determination, $R^2$, which expresses the proportion of variability in the outcome variable explained by the regressors. More precisely the $F$ statistic is closely related to the ratio, $R^2/(1-R^2)$. If the genotype is a weak instrument it will not a explain a substantial amount of the variability in the phenotype and will hence report a low first stage $R^2$ and hence a low $F$-statistic.

Figure 4.10 shows plots of the first stage $R^2$ for the simulations. The value of $R^2$ reduces as the impact of the confounder, through $\alpha_2$, increases. As $\alpha_2$ increases, and the instrument becomes weaker, it would be expected that estimators of $\beta_1$ would demonstrate increased bias. However, the adjusted estimator does not follow this trend since the bias reduced as $\alpha_2$ increased in the simulations, which implies that the adjusted IV estimator may be more robust to the weak instrument problem.

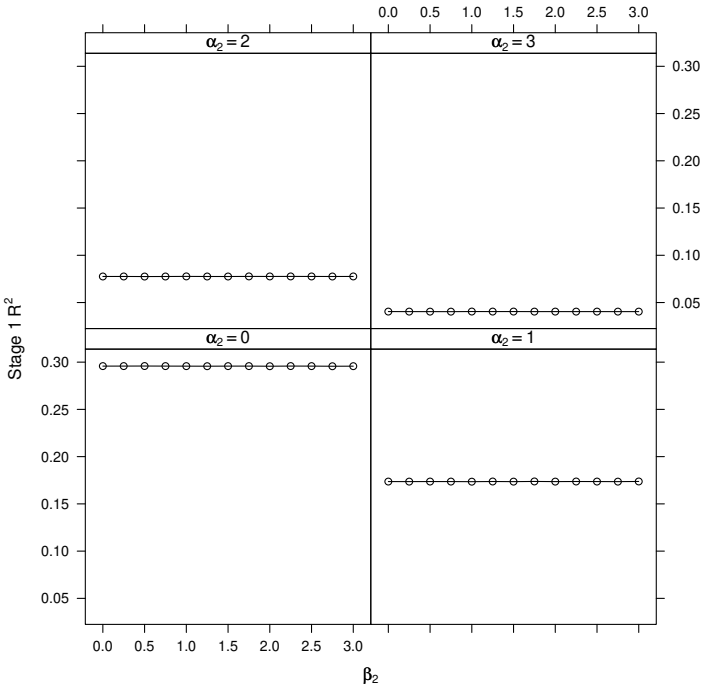



Figure 4.10: Values of the first stage $R^2$ from the simulations.

### 4.2.7 Comparison of the Neuhaus and Zeger equations

The expression given in Equation 3.45 based on that given by Zeger *et al.* (1988) can be compared with an expression given by Neuhaus *et al.* (1991) which is given in more detail in Appendix B.

The standard IV estimates under the Zeger approximation were compared with the Neuhaus approximation in Figure 4.11. The approximations are similar and follow the same trend. A disadvantage of the Neuhaus approximation is that the true probabilities of the outcome measure are required in the second stage regression. The true probabilities were used in Figure 4.11 because the data was simulated, however in practice they would not be known and predictions would have to be used.

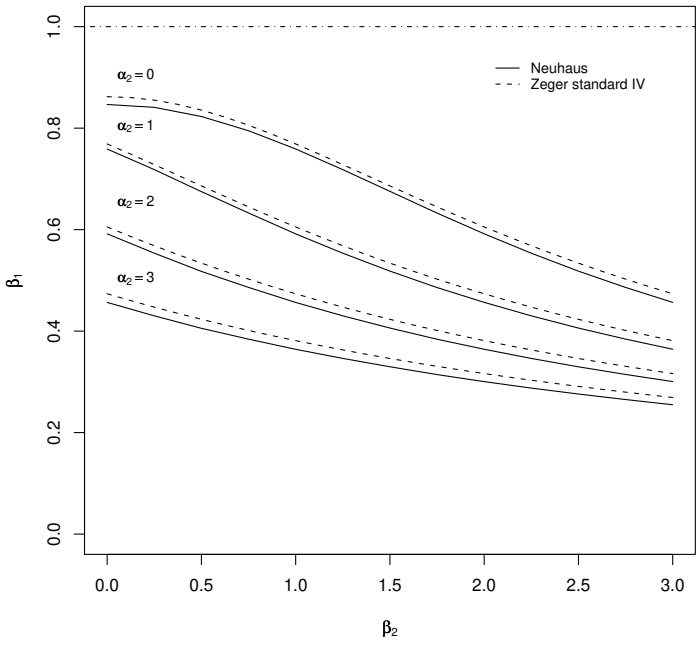



Figure 4.11: The Neuhaus approximation compared with the Zeger adjustment of the standard IV estimator using the logit link.

### 4.2.8 Results for the probit link

When studies are of case-control design there is a strong case to use logistic regression, however it is interesting to consider the implications for the analysis if a probit regression were used instead for the three estimators. Therefore, the simulations were repeated replacing the logistic regressions by probit regressions at the second stage of the estimation procedures and also in the data simulation. The relationship between marginal and conditional parameter estimates for probit regression with a random intercept is given in Appendix B.

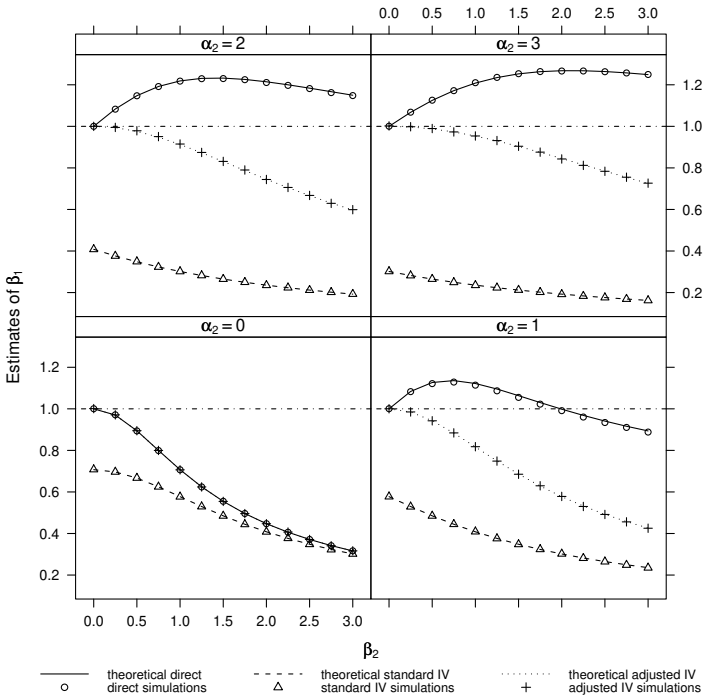



Figure 4.12: Theoretical and simulated estimates of $\beta_1$ with the probit link.

Figure 4.12 shows the median estimated parameter values from the simulations. Again the theoretical values using the Zeger approximation are in close agreement with the simulated values and the adjusted IV estimator falls between the other two. However, in this instance the adjusted IV estimator does not always have the smallest bias of three. For example in the panel with $\alpha_1$, the direct estimates have the smallest bias. This is perhaps a reason to prefer logistic over probit regression for Mendelian randomization analyses. The plots

of the intercept estimates, type I error and coverage all demonstrated the same patterns as for the logit link and are given in Appendix C.

## 4.3 Discussion

This simulation study has evaluated the properties of three estimators of the phenotype-disease log odds ratio, termed: direct, standard IV and adjusted IV for Mendelian randomization studies with binary outcomes. Under the logit link the adjusted IV estimator has the least bias of the three in terms of the difference between the estimated parameter value and the underlying true parameter.

In this modelling framework the effects of the unmeasured confounder and the random error about the phenotype act as a random effect in the logistic regression between the phenotype and disease. Hence the difference between the underlying and estimated parameter values can be explained in terms of the difference between marginal and conditional parameter estimates in generalized linear mixed models. Formulae for this difference were given in the previous chapter and more details about their derivation are given in Appendix B. These formulae could be used to perform a sensitivity analysis of the size of the phenotype-disease association after an analysis for different hypothesised amounts of confounding. The theory and further simulations show that the difference between marginal and conditional parameter estimates also exists for the estimators using probit regression at the second stage.

Despite being the least biased the adjusted IV estimator had high type I error when the effect of the unmeasured confounder was large. The inflated type I error for the adjusted IV estimator was observed under the Wald and likelihood ratio tests and the Wald test using standard errors from logistic regression allowing for over-dispersion. The high type I error of the adjusted IV estimator may be improved if a correction were applied to the standard errors of the form as for two-stage least squares and non-linear two stage least squares detailed in the previous chapter. The need to correct the standard errors of the adjusted estimator was also demonstrated by the coverage plots which showed low coverage with respect to the conditional and marginal values of $\beta_1$. Dunn & Bentall (2007) have used bootstrapping to obtain standard errors for the adjusted IV estimator in the case of

a continuous outcome measure and the identity link and Dunn *et al.* (2005) commented that bootstrapping may help give correct standard errors for non-linear IV models. The Wald test of the standard IV estimates had the correct type I error, and it may therefore be possible to use the standard IV estimator for hypothesis testing and the adjusted IV estimator for reporting parameter estimates [1].

The unmeasured confounding factors need to act as true confounders. If they act solely on the phenotype-disease association then the adjusted IV estimator is equivalent to the direct estimator. In these simulations an additive genetic model was assumed, because a linear regression was assumed between the phenotype and genotype. The general conclusions drawn here would also hold for the recessive and dominant genetic models.

The simulations investigated the performance of the estimators over a range of values of the confounder. Typically the gene used in a Mendelian randomization study will only explain a small percentage of the variance in the phenotype, perhaps less than 10 percent. It is therefore plausible that the impact of unmeasured confounding factors could be high, or indeed the error variance of the phenotype-disease association could be high or that the variance in the phenotype attributable to other polymorphisms could be high, causing bias in the parameter estimates of a Mendelian randomization analysis. In the analysis of a real study if there is information on other covariates which are possibly confounders they should be included in both stages of the analysis. This is because their inclusion will reduce the importance of the unmeasured confounders and help to reduce the bias in the parameter estimates. There could of course be many other determinants of the phenotype or the risk of disease that would not act as true confounders and which therefore could not be accounted for by the adjusted IV estimator.

Comparing Figures 4.1 and 4.10 the bias in the adjusted IV estimator reduces as the first stage $R^2$ decreases across the four values of $\alpha_2$. Therefore, the adjusted IV estimator may be robust to the weak instrument problem when the weak instrument is caused by a large effect of the unmeasured confounder.

[1] This suggestion was made by one of the anonymous reviewers of Palmer *et al.* (2008a)

A weakness of these simulations is that there were instances in which the predicted levels of the phenotype were perfect predictors for the disease outcome (Venables & Ripley, 2002, page 197). Fitted probabilities very close to zero or one in binomial GLMs can lead to a loss in power of the Wald test (Hauck Jr & Donner, 1977).

### 4.3.1 Practical implications of these simulations

Studies such as Timpson *et al.* (2005) that have reported Mendelian randomization analyses for continuous outcome measures using two-stage least squares or another of the equivalent methods are not affected by the results of these simulations. This is because the adjusted IV estimator with an identity link at the second stage produces equivalent parameter estimates to two-stage least squares as shown in Appendix C.

Qi *et al.* (2007) reported a study investigating the effect of plasma interleukin 6 (IL-6) levels on the risk of type II diabetes risk. The study reported an odds ratio of diabetes of 1.78 per unit change in log(IL-6) (95% CI: 1.49, 2.10) for the direct association and an odds ratio of 1.59 per unit change in log(IL-6) (95% CI: 0.45, 5.66) using the `qvf` command in Stata to perform instrumental variable analysis. Notably the direct association is statistically significant but the instrumental variable analysis is not. It is expected that the adjusted IV estimator would estimate an odds ratio between the direct and `qvf` estimates that is also not statistically significant because it is expected that for binary outcomes the `qvf` command gives estimates similar to the standard IV estimator in the simulations.

Another binary outcome study applying Mendelian randomization is Davey Smith *et al.* (2005a) investigating the association between CRP and hypertension using the 1059G/C polymorphism in the human CRP gene. The direct odds ratio of hypertension was reported to be 1.14 per quartile of CRP (95% CI: 1.09, 1.19) whereas the odds ratio using the `qvf` command was 1.03 per quartile of CRP (95% CI: 0.61, 1.73). Again it is notable that the direct estimate is statistically significant and that the instrumental variable estimate is not. Similarly to the previous example it is expected that the adjusted IV estimate of the odds ratio would be between the direct and `qvf` estimates and would also not be

statistically significant.

Hence, for both binary outcome examples it is expected that the use of the adjusted IV estimator would change the magnitude of the reported phenotype-disease association but not the overall conclusions of the instrumental variable analysis.

Table 4.1 shows an analysis on a subset of the data ($N$=3,597) from Davey Smith *et al.* (2005a). This was the same model as cited above and I cannot explain why the `qvf` point estimate is different to that cited in the original paper (OR=0.89 versus OR=1.03), although the confidence intervals are similar. Importantly, the output, including the $p$-values, from the standard IV estimator and the `qvf` command is similar. The adjusted IV odds ratio is also very slightly larger than that from the `qvf` command as predicted.

| Model | OR (95% CI) | P-value |
|---|---|---|
| Direct | 1.1464 (1.0976, 1.1974) | <0.001 |
| `qvf` | 0.8931 (0.5098, 1.5648) | 0.693 |
| Standard IV | 0.8932 (0.5143, 1.5514) | 0.689 |
| Adjusted IV | 0.8934 (0.5129, 1.5563) | NA |

Table 4.1: Comparing different estimators in a subset of Davey Smith *et al.* (2005a).

### 4.3.2 Further work

In terms of extending these simulations the standard deviation of the phenotype could also be varied. It is expected that as the phenotype standard deviation increases that the bias in the estimators would increase.

These simulations used cohort studies so there is an issue how the results would compare using case-control studies. Comparing the parameter estimates from a logistic regression of a cohort study and a case-control study only the intercept differs between the two under the rare disease assumption (Farewell, 1979). In a cohort study the intercept represents the baseline log odds of disease, whereas in a case-control study, for a rare disease, the intercept represents the baseline log odds of disease plus the logarithm of the ratio of the sampling fraction of the cases and controls. Hence it is expected that the results would

be same for the $\beta_1$ parameter using case-control studies in the simulations, whilst the estimates of the intercept would follow the same trend but have slightly different values.

It is important to investigate how to correct the standard error of the adjusted IV estimator so that its significance tests have the correct type I error. For example, the `qvf` command has the option to use Murphy-Topel standard errors (Hardin, 2002; Hardin & Carroll, 2003b; Murphy & Topel, 1985) which could be investigated for the adjusted IV estimator. Murphy-Topel standard errors were described in section 3.2.4.

### 4.3.3 Further work II: a scaling factor correction to the standard errors of the adjusted IV estimator

It is known that the standard IV estimator has the correct type I error, therefore it should be possible to use the $z$-statistic of its $\beta_1$ parameter to adjust the standard errors of the adjusted IV estimator. Denoting a scaling factor by $k$ and the estimated parameters from the adjusted and standard IV estimators by $\beta_A$ and $\beta_S$ and their standard errors by $s_A$ and $s_S$, equating $z$-statistics for $\beta_1$ we know that,

$$z_A k = z_S \tag{4.1}$$

$$\frac{\beta_A}{s_A} k = \frac{\beta_S}{s_S} \tag{4.2}$$

$$\Rightarrow k = \frac{\beta_S s_A}{s_S \beta_A}. \tag{4.3}$$

Hence $s_A$ should be multiplied by $1/k$ to give the correct type I error for the adjusted IV estimator, or equivalently the corrected standard errors are given by $s_S \beta_A / \beta_S$. When one of $\beta_A$ or $\beta_S$ is negative this will produce a negative corrected standard error, in order to avoid this the absolute value should be used.

Figure 4.13 shows the type I error of the adjusted IV estimator using the scaled standard error for the scenario with $\alpha_2 = 3$. There is an improvement in the type I error over Figure

4.4 which is now at the nominal value of 5%, the intended outcome of the correction.

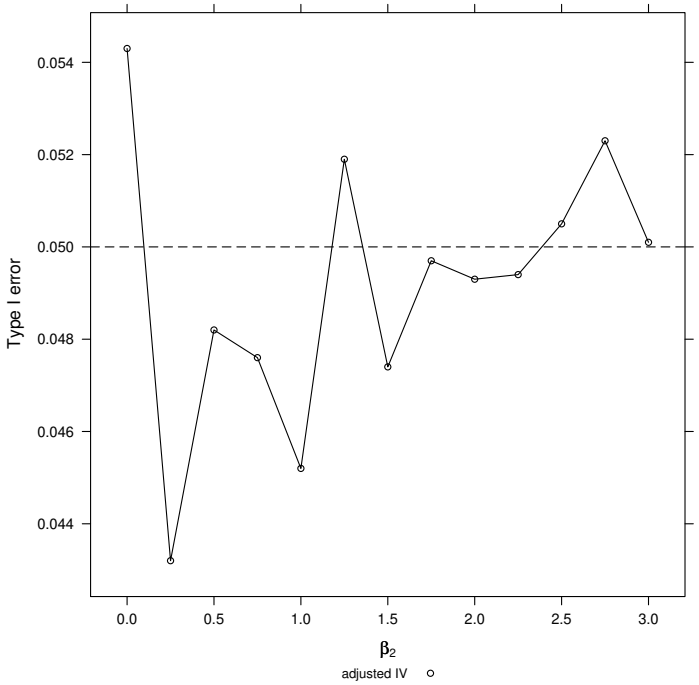



Figure 4.13: Type I error of the Wald test of the adjusted IV estimator using scaled standard errors for $\alpha_2 = 3$.

Figure 4.14 shows the coverage of the adjusted IV estimator using scaled standard errors with respect to $\beta_1 = 1$ (left) and $\beta_m$ (right). Both coverage plots improve upon their unscaled equivalents in Figures 4.2 and 4.3. However, many of the coverage levels are above the 95% level which suggests that the correction may have to be used with caution. A problem with this correction is that it will give corrected standard errors close to 0 when $\beta_A \approx 0$ and very large corrected standard errors when $\beta_S \approx 0$. Hence, this correction may be inappropriate when there is a null effect.

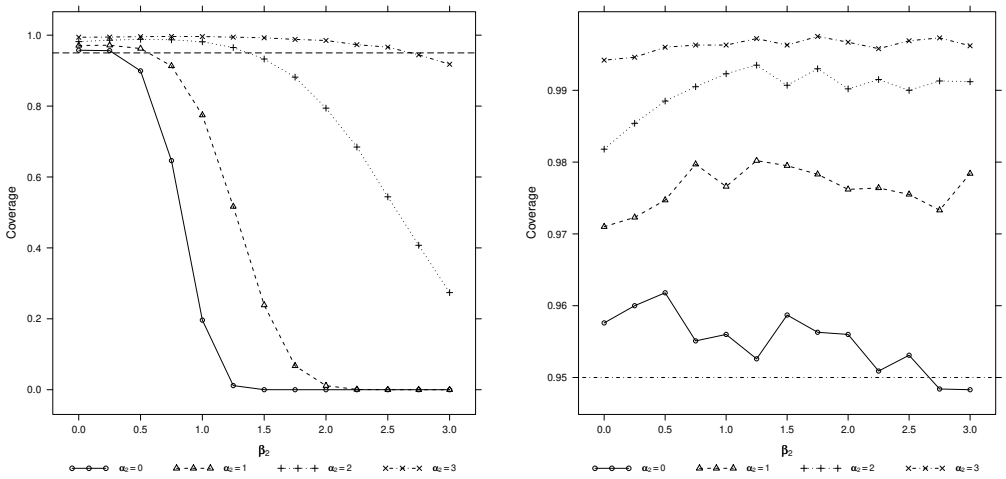



Figure 4.14: Coverage probability of the adjusted IV estimator using scaled standard errors with respect to $\beta_c$ (left) and $\beta_m$ (right).

# Chapter 5

# Meta-analysis models for Mendelian randomization studies

## 5.1 Introduction

Meta-analyses provide quantitative summaries of the evidence available on a particular research question. It has been argued that statistical power can be low in individual Mendelian randomization studies since large sample sizes are required to produce precise estimates of the phenotype-disease association (Davey Smith *et al.*, 2004; Lawlor *et al.*, 2008d). The CRP CHD Genetics Collaboration calculated sample size requirements for a Mendelian randomization analysis based on the expected effects of the genotype on the phenotype and the phenotype effects on the disease risk for the association between CRP and coronary heart disease. This study will require more than 10,000 cases to detect odds ratios less than 1.2. However, such a large sample size is not usually feasible in a single study, hence the CRP CHD Genetics Collaboration, (2008) is a collaboration of studies. So it is an advantage if the genotype-disease and genotype-phenotype estimates are derived from meta-analyses.

An issue in meta-analysis is whether the underlying effect is assumed to be the same in

all studies, which means whether a fixed or random effects analysis is more appropriate (DerSimonian & Laird, 1986; Whitehead & Whitehead, 1991). For a meta-analysis of Mendelian randomization studies it would be possible to include between study heterogeneity on the gene-phenotype, gene-disease and the phenotype-disease associations. There are methods to quantify the amount of heterogeneity in a meta-analysis such as the $I^2$ statistic which is effectively an intra-class correlation coefficient for a meta-analysis (Higgins & Thompson, 2002).

One approach to the meta-analysis of Mendelian randomization studies would be to perform univariate meta-analyses of the gene-phenotype and gene-disease associations. The the ratio of coefficients approach could then be used to derive the phenotype-disease association. There are also meta-analysis models for studies that report multiple outcome measures based upon the multivariate normal distribution (van Houwelingen *et al.*, 1993, 2002). These models can be specified as fixed or random effects models and have the advantage that they can accommodate the within and between study variances and correlations between the outcome measures. If each study in the meta-analysis is truly a 'Mendelian randomization' study then it will report gene-disease and gene-phenotype outcome measures. Multivariate meta-analysis models for Mendelian randomization studies have been proposed based on the multivariate normal distribution making use of the ratio of coefficients approach (Minelli *et al.*, 2003, 2004; Thompson *et al.*, 2005).

Some of the work in this chapter is published in Palmer *et al.* (2008c) which is included at the end of the thesis.

## 5.2 Meta-analysis methods

This section describes the information available from a case-control study and the estimation of the phenotype-disease association using Mendelian randomization. Methods are then proposed for the meta-analysis of Mendelian randomization studies incorporating all three genotypes by using two genotype comparisons. This idea is then related to the

genetic model-free approach of Minelli *et al.* (2005a,b).

### 5.2.1 The ratio of coefficients approach for case-control studies

Suppose that the genotype, phenotype and disease status information are collected in the same study. For a genetic polymorphism with two alleles, the common allele, $g$ and a minor allele, $G$, there are three possible genotypes; the common or wild-type homozygote ($gg$), the heterozygote ($Gg$), and the mutant or uncommon homozygote ($GG$). Table 5.1 summarises the genotype-disease and genotype-phenotype associations in a case-control study. In the table the counts of cases and controls are denoted $n_{dj}$, subscript $d$ indicates case or control status (1 or 0) and subscript $j$ denotes the genotype (1, 2 or 3 corresponding to $gg$, $Gg$ and $GG$).

It has been commented that the phenotype should be measured in the controls since reverse causation might affect the level of the phenotype in the cases (Thompson *et al.*, 2003). The observed mean phenotype levels in the controls are denoted by $\overline{x}_j$ which are estimates of the true mean phenotype levels denoted $\mu_j$. The observed standard deviations of the phenotype levels are denoted $sd_j$. The observed mean phenotype differences between either the heterozygotes or the rare homozygotes versus the common homozygotes are given by $\widehat{\delta}_j = \overline{x}_j - \overline{x}_1$, the subscript indicates the genotype with which the common homozygotes are compared. The true genotype-phenotype mean differences are given by $\delta_j = \mu_j - \mu_1$. The genotype-disease log odds ratios are denoted by $\theta_j$.

| | Genotypes | | |
|---|---|---|---|
| | $gg$ | $Gg$ | $GG$ |
| Number of controls | $n_{01}$ | $n_{02}$ | $n_{03}$ |
| Number of cases | $n_{11}$ | $n_{12}$ | $n_{13}$ |
| Mean phenotypes in controls (s.d.) | $\overline{x}_1$ ($sd_1$) | $\overline{x}_2$ ($sd_2$) | $\overline{x}_3$ ($sd_3$) |

Table 5.1: Data available from a Mendelian randomization case-control study

The usual estimates of the gene-disease log odds ratios and their variances, based on the

delta-method, and covariances are given by,

$$\widehat{\theta}_2 = \log\left(\frac{n_{12}/n_{02}}{n_{11}/n_{01}}\right) \tag{5.1}$$

$$\text{var}(\widehat{\theta}_2) = \sum_{d=0}^{1}\sum_{j=1}^{2}\frac{1}{n_{dj}} \tag{5.2}$$

$$\widehat{\theta}_3 = \log\left(\frac{n_{13}/n_{03}}{n_{11}/n_{01}}\right) \tag{5.3}$$

$$\text{var}(\widehat{\theta}_3) = \sum_{d=0}^{1}\sum_{j=1,3}\frac{1}{n_{dj}} \tag{5.4}$$

$$\text{cov}(\widehat{\theta}_2, \widehat{\theta}_3) = \sum_{d=0}^{1}\frac{1}{n_{d1}}. \tag{5.5}$$

The variances of the mean phenotype differences are given by,

$$\text{var}(\widehat{\delta}_2) = \text{var}(\overline{x}_2) + \text{var}(\overline{x}_1) \tag{5.6}$$

$$\text{var}(\widehat{\delta}_3) = \text{var}(\overline{x}_3) + \text{var}(\overline{x}_1) \tag{5.7}$$

$$\text{cov}(\widehat{\delta}_2, \widehat{\delta}_3) = \text{var}(\overline{x}_1). \tag{5.8}$$

In an individual study if the disease status variable were a continuous outcome measure then the application of instrumental variable methods would produce an unbiased estimate of the phenotype-disease association, assuming that the genotype met the core conditions to qualify as an instrumental variable (Didelez & Sheehan, 2007b; Greene, 1999). However, case-control studies typically rely on binary disease status variables. So the ratio of coefficients approach can be used to estimate the phenotype-disease association by using the gene-disease log odds ratios and gene-phenotype mean differences as continuous outcome measures (Thomas *et al.*, 2007; Thompson *et al.*, 2003). The phenotype-disease log odds ratio, denoted by $\eta$, is given by,

$$\widehat{\eta}_{[k]} \approx \frac{k\theta}{\delta} \quad \text{where } k \text{ is a constant.} \tag{5.9}$$

Sometimes a unit increase in the phenotype will be biologically implausible and so an arbitrary constant $k$ can be included in the ratio so that $\eta$ represents the log odds ratio associated with a $k$-unit change in the phenotype (Thompson *et al.*, 2003).

From the data available from a Mendelian randomization case-control study reporting all three genotypes two non-redundant estimates of the phenotype-disease log odds ratio are possible. One estimate of $\eta$ is based on the comparison of the common homozygotes with the heterozygotes, using $\theta_2$ and $\delta_2$. The other is based on the rare homozygotes compared with the common homozygotes, using $\theta_3$ and $\delta_3$. In many situations it will be sensible to assume that the two estimates relate to a common underlying log odds ratio. In the meta-analysis model these two estimates of $\eta$ can be combined into a single, more efficient, estimate.

### 5.2.2 Meta-analysis incorporating two genotype comparisons

The meta-analysis model incorporating two genotype comparisons builds on previous meta-analysis models for Mendelian randomization studies for a single genotype comparison (Minelli *et al.*, 2004; Thompson *et al.*, 2005). The model relates the pooled gene-disease log odds ratios and pooled gene-phenotype mean differences using the ratio of coefficients approach from Equation 5.9 through the mean vector of a multivariate normal distribution. The model follows multivariate meta-analysis methodology, such as van Houwelingen *et al.* (2002), through the specification of the marginal distribution of the study outcome measures by combining within and between study variance-covariance matrices. The approach is the multivariate analogue of the univariate random-effects meta-analysis model of DerSimonian and Laird (DerSimonian & Laird, 1986).

In the following notation subscript $i$ denotes a study. It is assumed that the observed mean phenotype differences are normally distributed such that $\widehat{\delta}_{ji} \sim N(\delta_{ji},\ \mathrm{var}(\widehat{\delta}_{ji}))$ and that the true study-specific mean differences are normally distributed such that $\delta_{ji} \sim N(\delta_j,\ \tau_j^2)$, where $\tau_j^2$ is the between study variance of the true study mean differences. Then the marginal distribution of the observed mean differences is given by $\widehat{\delta}_{ji} \sim N(\delta_j,\ \mathrm{var}(\widehat{\delta}_{ji}) +$

$\tau_j^2$). Denoting the correlation between the pooled mean phenotype differences by $\rho$, the multivariate Mendelian randomization meta-analysis model, referred to as the MVMR model, then takes the following form,

$$\begin{bmatrix} \widehat{\theta}_{2i} \\ \widehat{\delta}_{2i} \\ \widehat{\theta}_{3i} \\ \widehat{\delta}_{3i} \end{bmatrix} \sim MVN \left( \begin{bmatrix} \eta\delta_2 \\ \delta_2 \\ \eta\delta_3 \\ \delta_3 \end{bmatrix}, \; \boldsymbol{V}_i + \boldsymbol{B}_1 \right), \tag{5.10}$$

$$\boldsymbol{V}_i = \begin{bmatrix} \text{var}(\widehat{\theta}_{2i}) & 0 & \text{cov}(\widehat{\theta}_{2i}, \widehat{\theta}_{3i}) & 0 \\ 0 & \text{var}(\widehat{\delta}_{2i}) & 0 & \text{cov}(\widehat{\delta}_{2i}, \widehat{\delta}_{3i}) \\ \text{cov}(\widehat{\theta}_{3i}, \widehat{\theta}_{2i}) & 0 & \text{var}(\widehat{\theta}_{3i}) & 0 \\ 0 & \text{cov}(\widehat{\delta}_{3i}, \widehat{\delta}_{2i}) & 0 & \text{var}(\widehat{\delta}_{3i}) \end{bmatrix}, \tag{5.11}$$

$$\boldsymbol{B}_1 = \begin{bmatrix} \tau_2^2 & \tau_2\tau_3\rho \\ \tau_2\tau_3\rho & \tau_3^2 \end{bmatrix} \otimes \begin{bmatrix} \eta^2 & \eta \\ \eta & 1 \end{bmatrix} = \begin{bmatrix} \eta^2\tau_2^2 & \eta\tau_2^2 & \eta^2\tau_2\tau_3\rho & \eta\tau_2\tau_3\rho \\ \eta\tau_2^2 & \tau_2^2 & \eta\tau_2\tau_3\rho & \tau_2\tau_3\rho \\ \eta^2\tau_2\tau_3\rho & \eta\tau_2\tau_3\rho & \eta^2\tau_3^2 & \eta\tau_3^2 \\ \eta\tau_2\tau_3\rho & \tau_2\tau_3\rho & \eta\tau_3^2 & \tau_3^2 \end{bmatrix}. \tag{5.12}$$

The terms in the within-study covariance matrix, $\boldsymbol{V}_i$, are assumed known from the data reported by the studies and it is also assumed that there is no correlation between the gene-phenotype and gene-disease outcome measures as in Thompson *et al.* (2005). From the use of the Kronecker product it is apparent that $\boldsymbol{B}_1$ is singular, however $\boldsymbol{V}_i + \boldsymbol{B}_1$ is not, which allows the calculation of the likelihood.

The parameters of this model can be estimated by maximising the log-likelihood. For $i = 1 \ldots n$ studies $Y_i$ represents the $(4 \times 1)$ vector of outcome measures, $\beta$ represents the $(4 \times 1)$ mean vector of the multivariate normal distribution and $\boldsymbol{\Sigma}_i = \boldsymbol{V}_i + \boldsymbol{B}_1$. The log-likelihood of the multivariate normal distribution up to a constant is given by,

$$\sum_{i=1}^{n} -1/2 \left\{ \log(|\boldsymbol{\Sigma}_i|) + (Y_i - \beta)' \boldsymbol{\Sigma}_i^{-1} (Y_i - \beta) \right\}. \tag{5.13}$$

To improve the quadratic properties of the log-likelihood the log of $\tau_2^2$ and $\tau_3^2$ and the Fisher's $z$-transform of $\rho$ were used in the maximization which was performed using the optim function in R (version 2.7.0) (R Development Core Team, 2008).

### 5.2.3 Meta-analysis incorporating the genetic model-free approach

In the analysis of genetic association studies the mode of inheritance is usually unknown and so an assumption is made about the underlying genetic model. In contrast the genetic model-free approach estimates this underlying genetic model from the available data through a parameter $\lambda$ (Minelli *et al.*, 2005a,b). When $\lambda$ is equal to 0, 0.5 and 1 this represents recessive, additive and dominant models for the minor allele respectively.

The genetic model-free approach was devised in the context of a meta-analysis of two genotype comparisons for gene-disease outcome measures (Minelli *et al.*, 2005a,b). A consequence of assuming that the phenotype-disease association is constant across the comparison of the heterozygotes with the common homozygotes and the comparison of the rare homozygotes with the common homozygotes in Equation 5.10 is that the genetic model is assumed to be equal using either gene-disease or gene-phenotype outcomes, such that,

$$\frac{\theta_2}{\delta_2} = \frac{\theta_3}{\delta_3} \quad \text{implies} \quad \lambda = \frac{\theta_2}{\theta_3} = \frac{\delta_2}{\delta_3}. \tag{5.14}$$

The suggestion that the underlying genetic model can be inferred from either the gene-disease or gene-phenotype outcome measures was made by Thakkinstian *et al.* (2005, Section 2.4 and Table I), although they did not explicitly calculate the $\lambda$ statistic or note the relationship between genetic model-free and Mendelian randomization approaches.

The multivariate Mendelian randomization meta-analysis model incorporating the genetic

model-free approach, referred to as the MVMR-GMF model, is given by,

$$\begin{bmatrix} \widehat{\theta}_{2i} \\ \widehat{\delta}_{2i} \\ \widehat{\theta}_{3i} \\ \widehat{\delta}_{3i} \end{bmatrix} \sim MVN \left( \begin{bmatrix} \eta\lambda\delta_3 \\ \lambda\delta_3 \\ \eta\delta_3 \\ \delta_3 \end{bmatrix}, \; \boldsymbol{V}_i + \boldsymbol{B}_2 \right), \tag{5.15}$$

$$\boldsymbol{B}_2 = \begin{bmatrix} \lambda^2\tau_3^2 & \lambda\tau_3^2 \\ \lambda\tau_3^2 & \tau_3^2 \end{bmatrix} \otimes \begin{bmatrix} \eta^2 & \eta \\ \eta & 1 \end{bmatrix} = \begin{bmatrix} \eta^2\lambda^2\tau_3^2 & \eta\lambda^2\tau_3^2 & \eta^2\lambda\tau_3^2 & \eta\lambda\tau_3^2 \\ \eta\lambda^2\tau_3^2 & \lambda^2\tau_3^2 & \lambda\eta\tau_3^2 & \lambda\tau_3^2 \\ \eta^2\lambda\tau_3^2 & \lambda\eta\tau_3^2 & \eta^2\tau_3^2 & \eta\tau_3^2 \\ \eta\lambda\tau_3^2 & \lambda\tau_3^2 & \eta\tau_3^2 & \tau_3^2 \end{bmatrix}. \tag{5.16}$$

Similarly to the previous model $\boldsymbol{B}_2$ is singular but again $\boldsymbol{\Sigma}_i = \boldsymbol{V}_i + \boldsymbol{B}_2$ is not. The $z$-transform of $\lambda$ can be used in the maximization along with the other transformations previously described to help improve the quadratic properties of the log-likelihood. This model was also fitted by maximizing the log-likelihood in Equation 5.13.

It is also possible to estimate the parameters of this model using Bayesian methods. One Bayesian approach known as the Product Normal Formulation (PNF) expresses the multivariate normal distribution for each study's outcome measures as a series of univariate normal distributions linked by the relationships between the means (Spiegelhalter, 1998), such that,

$$\begin{aligned} \widehat{\theta}_{2i} &\sim N(\eta\lambda\delta_{3i}, \; \text{var}(\widehat{\theta}_{2i})), \\ \widehat{\delta}_{2i} &\sim N(\lambda\delta_{3i}, \; \text{var}(\widehat{\delta}_{2i})), \\ \widehat{\theta}_{3i} &\sim N(\eta\delta_{3i}, \; \text{var}(\widehat{\theta}_{3i})), \\ \widehat{\delta}_{3i} &\sim N(\delta_{3i}, \; \text{var}(\widehat{\delta}_{3i})), \\ \delta_{3i} &\sim N(\delta_3, \; \tau_3^2). \end{aligned} \tag{5.17}$$

It should be noted that the PNF relies upon the use of the Gibbs sampler (Geman &

Geman, 1984) to be estimated, since the correlations between the variables are induced by the sequential nature of parameter updating under this algorithm. The Gibbs sampler is implemented in WinBUGS which was used to fit this model (Lunn *et al.*, 2000). The following prior distributions were assumed for the parameters to be estimated,

$$\delta_3 \sim N(0,\ 1\times10^6),\ \tau_3^{-2} \sim Gamma(0.1,\ 0.1),\ \eta \sim N(0,\ 1\times10^6),\ \lambda \sim Beta(1,\ 1). \quad (5.18)$$

The normal prior distribution is approximately uniform over a broad range. The Beta prior distribution restricts $\lambda$ to lie between 0 and 1.

### 5.2.4 Missing outcomes

In a meta-analysis it is possible that some studies may not report all four outcomes. If studies are missing either gene-disease or gene-phenotype outcome measures these studies can be included in the model fitting using the appropriate bivariate log-likelihood derived by taking the appropriate rows and columns from equations (2), (3) and (4) or equations (7), (3) and (8). This requires the assumption that the missing outcomes are missing at random and not missing for a systematic reason.

### 5.2.5 Diagnostic plots

The results of a bivariate Mendelian randomization meta-analysis have been presented using a two column forest plot instead of two separate forest plots (Minelli *et al.*, 2004; Thompson *et al.*, 2005). For example, it can be difficult to detect consistent trends in the results of individual studies across multiple outcomes if separate single column forest plots are used, such as Lewis *et al.* (2006). For the models presented here using four outcomes the two column forest plot can be extended to a four column forest plot. To help compare the precision of the estimates the two columns of gene-disease log odds ratios should use the same scale as should the two columns of gene-phenotype mean differences. This plot

is shown in Figure 5.1.

In the meta-analysis models the assumption of the common phenotype-disease association in both genotype comparisons can be assessed by plotting the gene-disease outcome measures against the gene-phenotype measures (Minelli *et al.*, 2004). From the ratio of coefficients approach the phenotype-disease association can be expressed as the gradient of the line of best fit through the origin on this plot which is shown in Figure 5.2.

In the MVMR-GMF meta-analysis model the assumption that the genetic model is the same in the gene-disease and gene-phenotype outcomes can be assessed by plotting the Gg vs gg comparison against the GG vs gg comparison for each set of outcomes respectively (Minelli *et al.*, 2005b). From the genetic model free approach $\lambda$ is given by the gradient of the line of best fit through the origin on these plots which are shown in Figure 5.3.

## 5.3 Application to bone mineral density and osteoporotic fracture

An example meta-analysis which reported the four outcome variables required to demonstrate the models was a meta-analysis which investigated the relationship between a polymorphism in the COL1A1 gene and bone mineral density (BMD) and the risk of osteoporotic fracture (Mann *et al.*, 2001).

### 5.3.1 Description of the meta-analysis

The *COL1A1* gene codes for one of the main forms of collagen and the Sp1 polymorphism has been shown in epidemiological studies to be associated with both bone mineral density and the risk of fracture (Grant *et al.*, 1996; Uitterlinden *et al.*, 1998). This polymorphism is therefore a candidate for use as an instrumental variable in the estimation of the association between bone mineral density and fracture risk. The *COL1A1* study presented two meta-analyses based on a single nucleotide, $G$ to $T$, polymorphism affecting a binding site for

the transcription factor Sp1 in the *COL1A1* gene. One meta-analysis investigated studies into *COL1A1* and bone mineral density and the other meta-analysis investigated studies of *COL1A1* and osteoporotic fracture risk. It is therefore possible to apply Mendelian randomization meta-analysis to this example. The studies of the gene-phenotype and gene-disease associations should be free from confounding whereas studies of the association of BMD with fracture may be confounded by factors such as the subject's age or the amount of exercise they take, and there may also be unknown confounders which cannot be controlled for in the analysis.

The $G$ and $T$ alleles of the polymorphism in the *COL1A1* gene are sometimes labelled $S$ and $s$ for the common and minor alleles respectively, but for consistency with the methods section they are labelled $g$ and $G$. In estimating the phenotype-disease association using Mendelian randomization a one unit change in the phenotype can have a large impact on disease risk. In the example the standard deviation of the mean difference in BMD was 0.05 $g/cm^2$ between the homozygote genotypes and 0.03 $g/cm^2$ for comparison of the heterozygotes versus the common homozygotes. Therefore the scaling constant, $k$, was set to 0.05 in the analysis to ensure the pooled phenotype-disease odds ratio was estimated on an appropriate scale.

### 5.3.2 Results of the meta-analysis

Figure 5.1 shows a four column forest plot of the *COL1A1* meta-analyses. The first and second columns of the forest plot present the genotype-disease (G-D) and genotype-phenotype (G-P) outcomes for the $Gg$ versus $gg$ genotypes whilst the third and fourth columns show the outcomes for the $GG$ versus $gg$ genotypes. The forest plot shows that there is an increased risk of fracture in the $Gg$ over the $gg$ genotype and an increased risk again in the $GG$ genotype. The heterozygotes and the rare homozygotes had lower BMD than the common homozygotes. The forest plot shows that the comparison of the heterozygotes with the common homozygotes has more precise estimates because the confidence intervals around the point estimates are narrower and shows less between study

heterogeneity because the point estimates are more similar to one another.

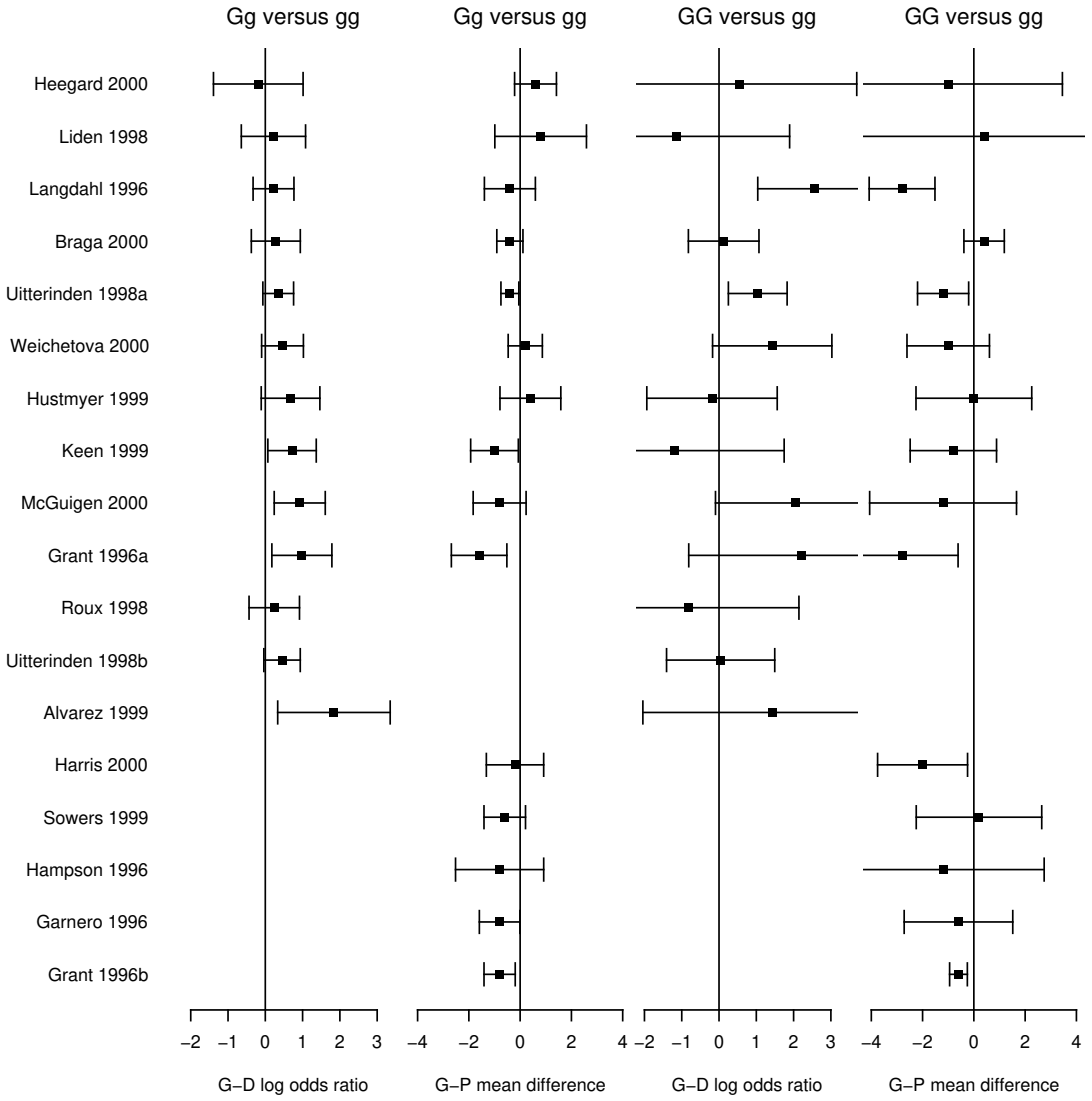



Figure 5.1: Four column forest plot of the *COL1A1* multivariate meta-analysis. The genotype-phenotype (G-P) columns are on a per $0.05g/cm^2$ scale.

The parameter estimates from the meta-analysis models incorporating all three genotypes are shown in Table 5.2. In the tables of parameter estimates, NA indicates a parameter that was not estimated in that particular model. The estimation of the PNF model was performed with a burn-in of 10,000 iterations followed by a chain of 50,000 iterations and MCMC convergence was assessed graphically. The estimates of $\eta$ were similar across the three models with odds ratios of osteoporotic fracture of 0.38 and 0.39 per $0.05g/cm^2$

increase in BMD. All three pooled odds ratios were statistically significant at the 5% level. The parameters in the PNF model had wider 95% credible intervals than the 95% confidence intervals in the MVMR-GMF model. The estimates of $\lambda$ in the MVMR-GMF and PNF models were close to 0.5 with both 95% intervals including 0.5 suggesting an additive model.

| Parameter | MVMR Est (95% CI) ($n = 18$) | MVMR-GMF Est (95% CI) ($n = 18$) | PNF Est (95% CrI) ($n = 18$) |
|---|---|---|---|
| $\eta$ | -0.96 (-1.39, -0.53) | -0.94 (-1.41, -0.47) | -0.97 (-1.53, -0.58) |
| $\exp(\eta)$ | 0.38 (0.25, 0.59) | 0.39 (0.24, 0.63) | 0.38 (0.22, 0.56) |
| $\lambda$ | NA | 0.43 (0.20, 0.61) | 0.47 (0.28, 0.74) |
| $\delta_2$ | -0.47 (-0.63, -0.30) | NA | NA |
| $\delta_3$ | -0.85 (-1.35, -0.35) | -0.94 (-1.34, -0.55) | -0.92 (-1.44, -0.49) |
| $\tau_2^2$ | 0.03 (0.001, 1.17) | NA | NA |
| $\tau_3^2$ | 0.53 (0.15, 1.91) | 0.35 (0.10, 1.28) | 0.43 (0.07, 1.35) |
| $\rho$ | 0.05 (-0.89, 0.91) | NA | NA |
| Log-likelihood | -6.42 | -11.24 | NA |

Table 5.2: Parameter estimates for meta-analysis models using studies with complete and incomplete outcomes.

As a comparison parameter estimates from bivariate meta-analysis models similar to those considered by Thompson *et al.* (2005) for the two genotype comparisons separately are given in Table 5.4. The pooled odds ratio of fracture was 0.34 (95% CI: 0.17, 0.68) per 0.05 $g/cm^2$ for the $Gg$ vs $gg$ comparison and 0.42 (95% CI: 0.25, 0.72) for the $GG$ vs $gg$ comparison and the three estimates from the models in Table 5.2 are between the two values. The estimates in Table 5.2 are also more precise, as shown by the narrower confidence intervals, because of the inclusion of data for both genotype comparisons.

| Parameter | PNF SA 1 Est (95% CrI) ($n = 18$) | PNF SA 2 Est (95% CrI) ($n = 18$) |
|---|---|---|
| $\eta$ | -1.01 (-1.61, -0.60) | -0.97 (-1.53, -0.58) |
| $\exp(\eta)$ | 0.38 (0.20, 0.55) | 0.39 (0.22, 0.56) |
| $\lambda$ | 0.49 (0.28, 0.78) | 0.46 (0.27, 0.75) |
| $\delta_3$ | -0.89 (-1.41, -0.48) | -0.93 (-1.49, -0.48) |
| $\tau_3^2$ | 0.35 (0.01, 1.23) | 0.21 (0.04, 1.62) |

Table 5.3: Sensitivity analyses for the PNF model.

In Table 5.3 two sensitivity analyses were performed for the product normal formulation model by varying the prior distribution on $\tau_3^2$. Sensitivity analysis 1 was performed with $1/\tau_3^2 \sim Gamma(0.001, 0.001)$ and sensitivity analysis 2 was performed with $1/\tau_3 \sim Uniform(0.1,\ 10)$. The values of the parameter estimates did not change substantially compared with the original model.

| Parameter | *Gg* vs *gg* Estimate (95% CI) ($n = 18$) | *GG* vs *gg* Estimate (95% CI) ($n = 18$) |
|---|---|---|
| $\eta$ | -1.08 (-1.76, -0.39) | -0.86 (-1.39, -0.33) |
| $\exp(\eta)$ | 0.34 (0.17, 0.68) | 0.42 (0.25, 0.72) |
| $\delta_2$ | -0.44 (-0.59, -0.28) | NA |
| $\delta_3$ | NA | -0.90 (-1.42, -0.38) |
| $\tau_2^2$ | 0.02 (0.001, 2.27) | NA |
| $\tau_3^2$ | NA | 0.56 (0.16, 1.96) |

Table 5.4: Parameter estimates from bivariate Mendelian randomization meta-analysis models using studies with complete and incomplete outcomes.

Parameter estimates from the bivariate meta-analysis models incorporating the genetic model-free approach using the gene-disease and gene-phenotype associations separately as in Minelli *et al.* (2005b) are given in Table 5.5. The maximization of the gene-disease model failed to converge and so the between study variance, $\tau_{\theta_3}^2$, was held constant. The fixed value of $\tau_{\theta_3}^2$ of 0.31 was taken from the univariate random effects meta-analysis of the *GG* vs *gg* gene-disease log odds ratios. The estimate of $\lambda$ was 0.44 (95% CI: 0.19, 0.64) from the gene-disease log odds ratios and 0.42 (95% CI: 0.08, 0.67) from the gene-phenotype mean differences and the estimate of $\lambda$ from the MVMR-GMF model is between these two values with increased precision. Two sensitivity analyses (SA) for the gene-disease genetic model free meta-analysis model were performed with values of $\tau_{\theta_3}^2$ greater and less than 0.31. The parameter estimates from these sensitivity analyses were qualitatively similar to the original model.

Figure 5.2 shows the diagnostic plot to assess the pooled estimate of $\eta$ with the gene-phenotype outcome measures on the $x$-axis and the gene-disease outcome measures on the $y$-axis. Given that two genotype comparisons are assessed, each study can contribute

| Parameter | Gene-disease Estimate (95% CI) ($n = 13$) | GD SA 1 Est (95% CI) ($n = 13$) | GD SA 2 Est (95% CI) ($n = 13$) | Gene-phenotype Estimate (95% CI) ($n = 15$) |
|---|---|---|---|---|
| $\lambda$ | 0.44 (0.19, 0.64) | 0.61 (0.10, 0.86) | 0.41 (0.18, 0.60) | 0.42 (0.08, 0.67) |
| $\theta_3$ | 0.96 (0.50, 1.43) | 0.78 (0.51, 1.04) | 1.01 (0.47, 1.54) | NA |
| $\exp(\theta_3)$ | 2.62 (1.65, 4.16) | 2.17 (1.67, 2.82) | 2.74 (1.60, 4.69) | NA |
| $\tau^2_{\theta_3}$ | fixed at 0.31 | fixed at 0 | fixed at 0.5 | NA |
| $\delta_3$ | NA | NA | NA | -0.88 (-1.40, -0.37) |
| $\tau^2_3$ | NA | NA | NA | 0.48 (0.10, 2.31) |

Table 5.5: Parameter estimates from bivariate genetic model-free meta-analysis models.

two points to the plot. A line with gradient equal to the pooled estimate of $\eta$ is drawn on the plot to help assess the fit of the model. Only one point did not lie within one standard deviation of the fitted line. Figure 5.2 also shows that the point estimates from the *GG* versus *gg* comparison have greater between study heterogeneity because the point estimates are spread over a wider range, and they are less precise than the point estimates from the *Gg* versus *gg* comparison.

Figures 5.3(a) and 5.3(b) assess the estimated genetic model from the MVMR-GMF meta-analysis model. On both figures lines have been plotted with gradients equal to $\widehat{\lambda}$ from the MVMR-GMF model and 0.5 to represent the additive genetic model. For this meta-analysis these figures are sensitive to the fact that not all studies reported both sets of outcome measures and so not all studies could be shown on each plot.

## 5.4 Discussion and conclusions

In observational epidemiology estimates from a Mendelian randomization analysis can provide improved estimates of the association between a biological phenotype and a disease compared with direct estimates of this association. The proposed meta-analysis models extend previous literature by incorporating both genotype comparisons for a given genetic polymorphism into the same model. The MVMR-GMF and PNF meta-analysis models also incorporate the estimation of the underlying genetic model for the risk allele in a

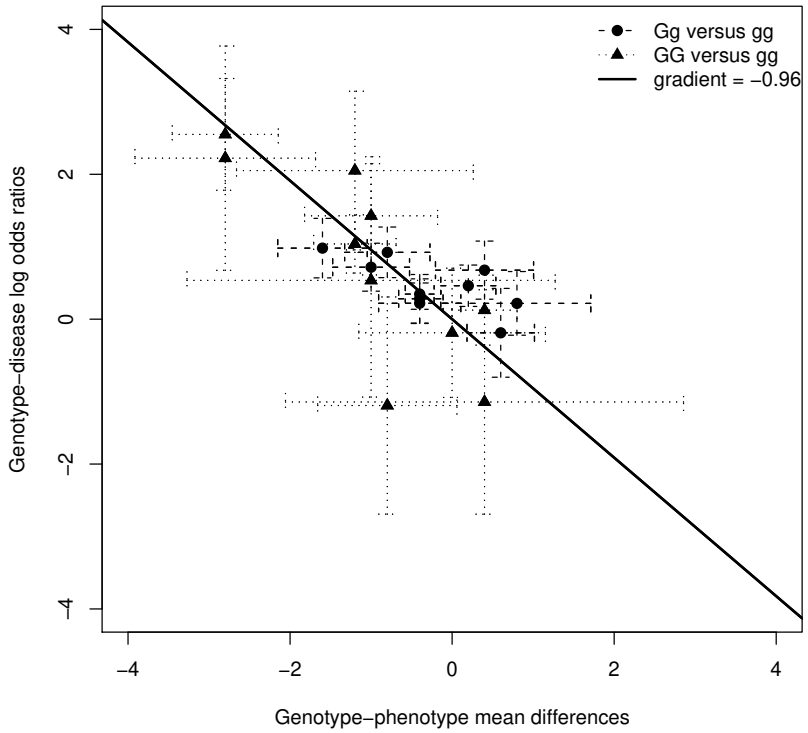
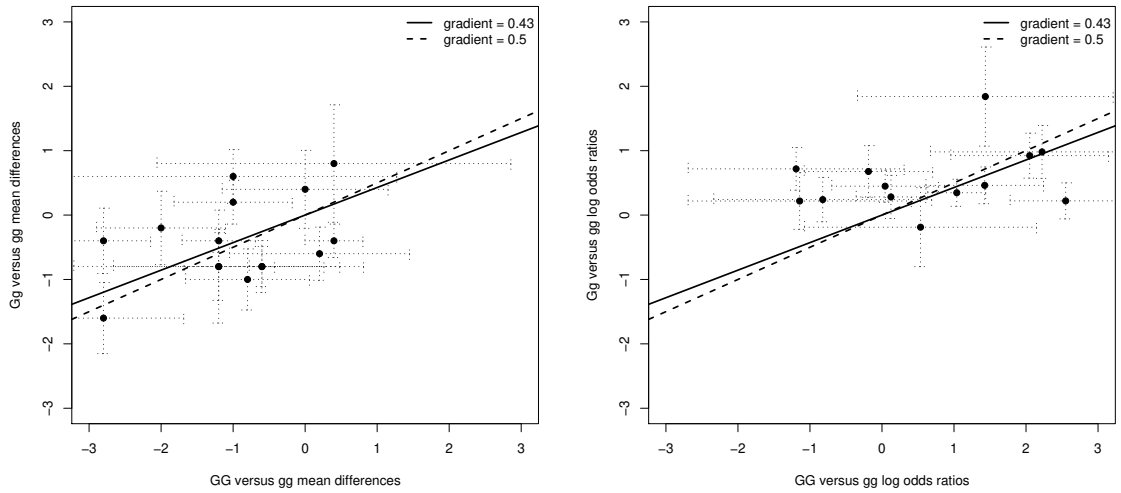



Figure 5.2: Gene-disease log odds ratios versus gene-phenotype mean differences (per $0.05g/cm^2$) plotted with 1 standard deviation error bars. The gradient of the line is given by $\widehat{\eta}$ from the MVMR meta-analysis model.

Mendelian randomization analysis.

The proposed meta-analysis models rely on two important assumptions, namely; that the phenotype-disease association is the same in the *Gg* versus *gg* and the *GG* versus *gg* genotype comparisons and that the underlying genetic model is the same in the gene-phenotype and gene-disease associations. These assumptions are assessed in Figures 5.2 and 5.3. The modelling approach could be extended to allow the phenotype-disease log odds ratio, $\eta$, to vary across studies. This would most easily be implemented using Bayesian methodology. Figure 5.1 shows a four column forest plot for a Mendelian randomization meta-analysis across two genotype comparisons. From the plot the relative precision of the estimates from the two genotype comparisons and the patterns in the estimates of individual studies can be assessed.




(a) Genotype-phenotype information per $0.05g/cm^2$ (b) Genotype-disease information

Figure 5.3: Graphical assessment of the estimated genetic model. The gradient of the bold lines is $\widehat{\lambda}$ from the MVMR-GMF model. A dashed line with gradient 0.5 representing the additive genetic model is also shown, a lines with gradients 0 and 1 would represent the recessive and dominant genetic models respectively.

Incorporating multiple genotype comparisons into a Mendelian randomization analysis is advantageous because the comparison of the heterozygotes with the common homozygotes has the larger sample size, whilst the comparison of the rare homozygotes with the common homozygotes has the larger difference in disease risk. Therefore the pooled estimate of the phenotype-disease association from the MVMR, MVMR-GMF and PNF models in Table 5.2 were between the estimates for the two separate bivariate meta-analysis models using single genotype comparisons in Table 5.4. The pooled estimate of the phenotype-disease association in the MVMR and MVMR-GMF models also showed increased precision over the single genotype comparison models because they included more information. Another advantage of incorporating all three genotypes is that if some of the studies omit to report either genotype-phenotype or genotype-disease outcome measures then they can be accommodated in the meta-analysis model using the appropriate bivariate normal likelihood. This requires the additional assumption that the missing outcomes were missing at random and not missing for a systematic reason such as reporting bias.

The estimation of the underlying genetic model for the risk allele, known as the genetic model-free approach, can also be incorporated within this meta-analysis framework. The proposed approach extends previous literature through the joint synthesis of the genotype-disease and genotype-phenotype information to estimate the genetic model. This means that no strong assumptions about the genetic model are required prior to the analysis. In the example meta-analysis the genetic model was estimated close to the additive genetic model. Apart from random variation, an explanation for estimates of $\lambda$ not at one of the genetic models is that in some studies there may have been a recessive effect and in other studies an additive effect and hence the value of $\lambda$ represents the average of these. Another explanation is that a gene's mode of action in complex diseases may differ from that found in Mendelian traits since the genotype is only one of many factors acting in a complex causal cascade leading to the disease (Minelli *et al.*, 2005b). Allowing heterogeneity within the estimation of $\lambda$ could be investigated using Bayesian methods.

The estimation of bivariate meta-analysis models has been shown to be problematic when correlation parameters are near $\pm 1$ (Riley *et al.*, 2007a,b, 2008; van Houwelingen *et al.*, 2002). To overcome this problem an alternative form of the marginal distribution for a multivariate meta-analysis model has been proposed which assumes a common correlation term both within and between studies (see model A in Thompson *et al.* (2005) or Riley *et al.* (2008)). The advantage of this alternative covariance structure is that only study outcome measures and their respective variances are required to fit the multivariate meta-analysis model. The same information is required to perform the univariate meta-analyses for each outcome measure separately. A further discussion of how the relative magnitudes of the within and between study covariance matrices can affect parameter estimates in multivariate meta-analysis models is provided by Ishak *et al.* (2008). To fit multivariate meta-analysis models the restricted log-likelihood could be used in the maximization as an alternative to the log-likelihood (Riley *et al.*, 2008).

It would be possible to use these and the previously proposed bivariate meta-analysis models for Mendelian randomization studies reporting continuous disease outcome measures

since the models assume that the log odds ratios are continuous and normally distributed. For case-control studies it would be possible to achieve similar pooled estimates of the phenotype-disease log odds ratio across two genotype comparisons using either a retrospective or a prospective likelihood for the genotype-disease outcome measures, which has previously been demonstrated for the genetic model-free approach (Minelli *et al.*, 2005a). Meta-analysis models have been used to estimate other parameters of interest from genetic data. For example, meta-regression has been used to investigate deviations from Hardy-Weinberg equilibrium (Salanti *et al.*, 2007) and merged genotype comparisons have been used to assess Hardy-Weinberg equilibrium and estimate the genetic model-free approach (Salanti & Higgins, 2008). The work presented here also has parallels with modelling baseline risk in meta-analyses (Thompson *et al.*, 1997; van Houwelingen & Senn, 1999).

The limitations that apply to the analysis of a single study using Mendelian randomization also apply to each of the studies in the meta-analysis. Therefore, it is important to assess that the selected genotype fulfills the conditions of an instrumental variable (Didelez & Sheehan, 2007b) and whether any of the factors which could potentially affect Mendelian randomization analyses such as pleiotropy or canalization are present (Davey Smith & Ebrahim, 2004).

With respect to the example a large study, with a sample size of approximately 20,000, has subsequently been published investigating the *COL1A1* Sp1 polymorphism and its effects on osteoporosis outcomes (Ralston *et al.*, 2006). The study observed that the polymorphism is associated with reduced bone mineral density in women and could predispose to incident vertebral fractures, although the observed associations were modest.

With respect to the plot of the gene-disease estimates versus the gene-phenotype estimates, in Figure 5.2, Freathy *et al.* (2008) used meta-analysis estimates instead of study estimates on an identical plot. Using estimates from meta-analyses in this way is referred to as meta-epidemiology (Egger *et al.*, 2002; Naylor, 1997). As noted by Egger *et al.* (2003) the aims of the first meta-epidemiological studies, such as Schulz *et al.* (1995), was to assess possible bias in the pooled effects reported in meta-analyses but some authors have started to pool

the results of meta-analyses, for example Sterne *et al.* (2002, Figures 1 & 2) and Wood *et al.* (2008).

### 5.4.1 Conclusion

In conclusion, estimating the phenotype-disease association using separate genotype comparisons is often limited in that the comparison of the homozygote genotypes has a smaller sample size, whereas the comparison of the heterozygotes with the common homozygotes involves a smaller difference in disease risk. Pooling the phenotype-disease association across these comparisons produces an estimate that is a weighted average of the two but with increased precision. This meta-analysis framework can incorporate the estimation of the genetic model-free approach so that no strong prior assumptions about the underlying genetic model are required.

# Chapter 6

# The ratio of coefficients approach

## 6.1 Introduction

The aim of this chapter is to investigate the properties of the ratio of coefficients approach for estimating the phenotype-disease log odds ratio in Mendelian randomization studies reporting binary outcomes. This includes the application of the ratio of coefficients approach within a single cohort and within the multivariate meta-analysis models proposed in Chapter 5.

The motivation for the work in this chapter is that Thompson *et al.* (2003) investigated the ratio of coefficients approach for Mendelian randomization studies reporting binary outcomes using a numerical approximation to estimate the gene-disease log odds ratio (the numerator of the ratio). The report concluded that the variance of the denominator, the genotype-phenotype association, is important in determining the accuracy of the ratio of coefficients estimate of the phenotype-disease association. This report also discussed some of the problems associated with deriving a confidence interval for the ratio of coefficients estimate using Fieller's Theorem. This chapter investigates the ratio of coefficients approach using an alternative approximation, specifically a Taylor series expansion, which is also used to derive a confidence interval for ratio estimate.

Additionally, in Chapter 5 two estimates of the phenotype-disease log odds ratio, $\eta_2$ and $\eta_3$, were defined using the two genotype comparisons of the Gg (heterozygotes) genotype and the GG (rare homozygotes) genotype with the gg (common homozygotes) genotype respectively. It was hypothesised that $\eta_2$ and $\eta_3$ should be equal. Therefore, this chapter investigates whether these estimates have similar properties within a single cohort and also within the multivariate MVMR meta-analysis model from the previous chapter.

## 6.2 Taylor series expansion

An alternative approach to investigating the ratio of coefficients estimate, to that used by Thompson *et al.* (2003), is to use a Taylor series expansion of the expectation of the ratio which has been discussed by Thomas *et al.* (2007). The Taylor series can also be used to derive an approximation of the variance of the ratio which in turn allows a confidence interval to be derived for the ratio estimate. The phenotype-disease log odds ratio, the genotype-disease log odds ratio and the genotype-phenotype association are denoted by $\eta$, $\theta$ and $\delta$; and $\theta$ and $\delta$ are random variables with means $\overline{\theta}$ and $\overline{\delta}$.

The standard formula for a Taylor series expansion, upto the second order, of a function of two variables is given below (Spiegel, 1971, Equation 16),

$$f(\theta,\delta) \approx f(\overline{\theta},\overline{\delta}) + \frac{\partial f(\overline{\theta},\overline{\delta})}{\partial\theta}(\theta-\overline{\theta}) + \frac{\partial f(\overline{\theta},\overline{\delta})}{\partial\delta}(\delta-\overline{\delta}) + \frac{1}{2!}\left[\frac{\partial^2 f(\overline{\theta},\overline{\delta})}{\partial\theta^2}(\theta-\overline{\theta})^2 + 2\frac{\partial^2 f(\overline{\theta},\overline{\delta})}{\partial\theta\partial\delta}(\theta-\overline{\theta})(\delta-\overline{\delta}) + \frac{\partial^2 f(\overline{\theta},\overline{\delta})}{\partial\delta^2}(\delta-\overline{\delta})^2\right]. \quad (6.1)$$

Therefore, the Taylor series expansion of $f(\theta,\delta) = \eta = \theta/\delta$ about $\overline{\theta}$ and $\overline{\delta}$ is given by,

$$\frac{\theta}{\delta} \approx \frac{\overline{\theta}}{\overline{\delta}} + \frac{1}{\overline{\delta}}(\theta-\overline{\theta}) + \frac{-\overline{\theta}}{\overline{\delta}^2}(\delta-\overline{\delta}) + \frac{1}{2}\left[0(\theta-\overline{\theta})^2 + 2\frac{-1}{\overline{\delta}^2}(\theta-\overline{\theta})(\delta-\overline{\delta}) + 2\frac{\overline{\theta}}{\overline{\delta}^3}(\delta-\overline{\delta})^2\right]. \quad (6.2)$$

Interest is in the expected value of the ratio, $E(\theta/\delta)$, so it is necessary to take the ex-

pectation of the previous expression. Taking the expectation of the previous expression removes the first order terms since $E(\theta - \overline{\theta}) = 0$ and $E(\delta - \overline{\delta}) = 0$, and hence,

$$\begin{aligned} E\left(\frac{\theta}{\delta}\right) &\approx \frac{\overline{\theta}}{\overline{\delta}} - \frac{E((\theta - \overline{\theta})(\delta - \overline{\delta}))}{\overline{\delta}^2} + \frac{\overline{\theta}E((\delta - \overline{\delta})^2)}{\overline{\delta}^3} \\ &= \frac{\overline{\theta}}{\overline{\delta}} - \frac{\text{cov}(\theta, \delta)}{\overline{\delta}^2} + \frac{\overline{\theta}\text{var}(\delta)}{\overline{\delta}^3}. \end{aligned} \tag{6.3}$$

The above expression for the Taylor series expansion of the ratio of two means is well known in the statistics literature and has been given by Hayya *et al.* (1975, Equation 7) and Thomas *et al.* (2007) amongst others. Rearranging Equation 6.3 for $\overline{\theta}/\overline{\delta}$ shows that the accuracy of the ratio of coefficients approach will primarily depend the upon the variance of $\delta$ and the covariance between the gene-disease and gene-phenotype associations.

The expression for the variance of the ratio using the Taylor series expansion is also well known in the statistics literature, for example it has been given by Kendall & Stuart (1977, Equation 10.17) and Wolter (2003, Equation 6.8.1). The expression for the variance takes the form,

$$\text{var}\left(\frac{\theta}{\delta}\right) \approx \frac{\overline{\theta}^2\text{var}(\delta)}{\overline{\delta}^4} + \frac{\text{var}(\theta)}{\overline{\delta}^2} - \frac{2\overline{\theta}\text{cov}(\theta, \delta)}{\overline{\delta}^3}. \tag{6.4}$$

It is then possible to derive a confidence interval for $E(\theta/\delta)$ under the assumption it is normally distributed with mean and variance given by Equations 6.3 and 6.4 respectively (Hayya *et al.*, 1975).

The Taylor series approximations are investigated by substituting in some hypothetical parameter values. It is simplest to assume that there is no correlation between $\theta$ and $\delta$ and so the covariance terms are dropped from the expressions. Thompson *et al.* (2003) considered an example using the MTHFR gene, levels of homocysteine as the phenotype

and coronary heart disease, the parameter values are given below,

$$\overline{\theta} = 0.15, \quad \text{var}(\overline{\theta}) = 0.05^2 = 0.0025,$$
$$\overline{\delta} = 1.5, \quad \text{var}(\overline{\delta}) = 0.2^2 = 0.04.$$
$$\overline{\theta}/\overline{\delta} = 0.1000$$
$$E(\overline{\theta}/\overline{\delta}) \approx 0.1018 \text{ (95\% CI: } 0.0314, 0.1721).$$

For this example the ratio of the means is close to the Taylor series approximation because the variance of the gene-phenotype association is small. The 95% confidence interval for phenotype-disease log odds ratio shows that it is statistically significant from zero at the 5% level.

A 95% confidence interval for the ratio estimate using Fieller's Theorem, as given by Thompson *et al.* (2003), can be derived using the expression below,

$$\frac{\overline{\theta}/\overline{\delta}}{1 - 1.96^2\text{var}(\overline{\delta})/\overline{\delta}^2}\left[1 \pm 1.96\sqrt{\frac{\text{var}(\overline{\theta})}{\overline{\theta}^2} + \frac{\text{var}(\overline{\delta})}{\overline{\delta}^2} - 1.96\frac{\text{var}(\overline{\theta})}{\overline{\theta}^2}\frac{\text{var}(\overline{\delta})}{\overline{\delta}^2}}\right]. \tag{6.5}$$

A confidence interval derived using Fieller's Theorem is not necessarily symmetric. The 95% confidence interval using Fieller's Theorem for the example is (0.0330, 0.1817) which is in good agreement with the Taylor series confidence interval.

## 6.3 The ratio of coefficients estimates of $\eta_2$ and $\eta_3$

In the previous chapter $\eta_2$ and $\eta_3$ were defined as the phenotype-disease log odds ratios when the Gg and GG genotypes were compared with the gg genotype respectively. Therefore, each of $\eta_2$ and $\eta_3$ is the ratio of the respective gene-disease log odds ratios to the difference in mean phenotypes. Where $p_j$ represents the probability of disease for genotype

$j$, $\eta_2$ and $\eta_3$ are given by,

$$\eta_2 = \frac{\log(p_2/(1-p_2)) - \log(p_1/(1-p_1))}{\mu_2 - \mu_1} = \frac{\theta_2}{\delta_2} \tag{6.6}$$

$$\eta_3 = \frac{\log(p_3/(1-p_3)) - \log(p_1/(1-p_1))}{\mu_3 - \mu_1} = \frac{\theta_3}{\delta_3}. \tag{6.7}$$

The estimates of these parameters for a single study were described in the previous chapter in the section including Equations 5.1 and 5.3.

### 6.3.1 Simulation algorithm

Cohorts were simulated using the approach described in Chapter 4 in Section 4.1.1. The genotype variable, $G$, was generated in accordance with Hardy-Weinberg equilibrium by setting the minor allele frequency, $q$. The phenotype variable, $X$, was then simulated from a normal distribution conditional on the genotype. The phenotype variable was in turn used to generate the logit of the probability of disease and the probability of disease was then calculated through back transformation. The disease status variable was assigned if the probability of disease for a subject exceeded a random number generated between 0 and 1. The confounder, $U$, was simulated from a normal distribution.

$$u_i \sim N(0, \sigma_u^2) \tag{6.8}$$

$$x_i = \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i, \quad \epsilon \sim N(0, \sigma_\epsilon^2) \tag{6.9}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 y_i + \beta_2 u_i \tag{6.10}$$

In these simulations the phenotype-disease log odds ratio was set to be smaller than the value of 1 set in Chapter 4. In these simulations $\beta_1$ was set to $\log(1.25) = 0.2231436$. The aim of these simulations is to determine whether both $\eta_2$ and $\eta_3$ recover this value.

### 6.3.2 Cohort size

With a minor allele frequency, $q$, of 30%, under Hardy-Weinberg equilibrium, as explained in Section 1.3.2, 49% of subjects are expected to have the gg genotype, 42% are expected to have the Gg genotype and 9% are expected to have the GG genotype. A cohort study was simulated with the following parameter values; $N = 3 \times 10^5,\ q = 0.3,\ \alpha_0 = 0,\ \alpha_1 = \alpha_2 = 1,\ \sigma^2_\epsilon = 1,\ \beta_0 = \log(0.05/0.95),\ \beta_1 = \log(1.25),\ \beta_2 = 1,\ \sigma^2_u = 1$. The cohort was found to have the cell probabilities shown in Table 6.1.

| | *gg* | *Gg* | *GG* | Sum |
|---|---|---|---|---|
| Controls | 0.4452 | 0.3708 | 0.0784 | 0.8944 |
| Cases | 0.0454 | 0.0475 | 0.0128 | 0.1056 |
| Sum | 0.4906 | 0.4183 | 0.0912 | |

Table 6.1: Expected cell probabilities for the simulated cohorts.

The column sums in Table 6.1 show that the simulation algorithm followed Hardy-Weinberg equilibrium and the row sums show that the risk of disease in the cohort was about 11%. In this cohort the cases with the GG genotype have the smallest probability of being observed. For this cohort it would require a sample size of 79 subjects in order to expect to a single disease case in the GG genotype. Under Hardy-Weinberg equilibrium it is interesting to note that the expected probability in the heterozygotes exceeds that of the common homozygotes at a risk allele frequency of $q > 1/3$.

When there are no cases in the GG genotype then gene-disease odds ratios involving the GG group, such as $\exp(\theta_3)$, will be equal to 0 and the corresponding log odds ratio will be undefined, or more precisely $\widehat{\theta}_3 \to -\infty$. It is not possible to include these undefined values of $\widehat{\theta}_3$ in the meta-analysis models, hence it is desirable to simulate cohorts in which they don't occur.

The probability of observing zero cases in the GG genotype is given by the Poisson distribution using an event rate of the expected cell probability from Table 6.1 multiplied by

the sample size, hence,

$$P(0 \text{ GG cases}) = \frac{\exp(-0.0128N)(0.0128N)^0}{0!}. \tag{6.11}$$

For $N = 100, 200, 300, 500, 1000$ then $P = 0.28, 0.08, 0.02, 0.002, 0.000035$ respectively. Hence, using a larger sample size in the simulations implies a smaller probability of generating zero cases in the GG genotype and the use of a sample size of around 1,000 subjects is sufficient to ensure that this has a very small probability of occurrence.

### 6.3.3 Single cohort simulations

The simulations were performed for several scenarios which used the same parameter values as for the cohort in Table 6.1 except for the following changes:

- Scenario 1: N=3,000, $\alpha_2, \beta_2 = 0$, $\sigma^2_\epsilon = 1$.
- Scenario 2: N=3,000, $\alpha_2, \beta_2 = 1$, $\sigma^2_\epsilon = 1$.
- Scenario 3: N=3,000, $\alpha_2, \beta_2 = 0$, $\sigma^2_\epsilon = 0.1^2 = 0.01$.
- Scenario 4: N=3,000, $\alpha_2, \beta_2 = 1$, $\sigma^2_\epsilon = 0.1^2 = 0.01$.

Hence, scenarios 1 and 3 have no confounder effect whilst the confounder is present in scenarios 2 and 4. Scenarios 1 and 2 have a relatively large phenotype error variance whilst this is smaller in scenarios 3 and 4. Each scenario was performed for 10,000 iterations. Tables 6.2 and 6.3 show the results of the simulations for the Gg versus gg and GG versus gg genotype comparisons respectively. In the tables $\widehat{\eta}_j$ represents the ratio of coefficients estimate averaged over the simulations; the $\widehat{\theta}_j$'s and $\widehat{\delta}_j$'s also represent the average of these estimates over the simulations. Also in the tables a 95% confidence interval is given for the estimates of $\eta$; the bias column relates to the bias in estimate of $\eta$ with respect to the set value; the mean squared error column is the average of the squared bias over the simulations; and the $\widehat{\eta}_{j,TS}$ column gives the Taylor series estimate of $\eta$ for the simulations as per Equation 6.3.

| | $\widehat{\theta}_2$ | $\widehat{\delta}_2$ | $\widehat{\eta}_2$ (95% CI) | Bias | MSE | $\widehat{\eta}_{2,TS}$ (95% TS CI) |
|---|---|---|---|---|---|---|
| 1 | 0.2215 | 0.9977 | 0.2224 (0.2191, 0.2258) | -0.0007 | 0.0289 | 0.2220 (0.2187, 0.2254) |
| 2 | 0.2182 | 0.9936 | 0.2201 (0.2176, 0.2226) | -0.0030 | 0.0163 | 0.2196 (0.2171, 0.2221) |
| 3 | 0.2221 | 1.0000 | 0.2221 (0.2187, 0.2254) | -0.0011 | 0.0287 | 0.2221 (0.2187, 0.2254) |
| 4 | 0.2191 | 0.9976 | 0.2197 (0.2172, 0.2222) | -0.0034 | 0.0162 | 0.2197 (0.2172, 0.2222) |

Table 6.2: Simulation results for the ratio of coefficients approach in a single cohort for genotypes Gg versus gg.

The general trend in the ratio of coefficients estimates of $\eta_2$, the $\widehat{\eta}_2$ column, in Table 6.2 is slight attenuation towards the null. However, the attenuation is not conclusive for scenarios 1 and 3 because their respective 95% confidence intervals for $\widehat{\eta}_2$ contain the set value of $\log(1.25) = 0.2231$.

From Table 6.2 it can be seen that the Taylor series approximation is in agreement with $\widehat{\eta}_2$ for scenarios 3 and 4, since the phenotype standard deviation was small in these scenarios. It can also be noted that due to the large number of iterations the Taylor series confidence interval was identical to the Fieller's Theorem confidence interval for all scenarios.

| | $\widehat{\theta}_3$ | $\widehat{\delta}_3$ | $\widehat{\eta}_3$ (95% CI) | Bias | MSE | $\widehat{\eta}_{3,TS}$ (95% TS CI) |
|---|---|---|---|---|---|---|
| 1 | 0.4262 | 1.9940 | 0.2140 (0.2114, 0.2166) | -0.0091 | 0.0180 | 0.2137 (0.2112, 0.2164) |
| 2 | 0.4285 | 1.9850 | 0.2161 (0.2141, 0.2181) | -0.0070 | 0.0105 | 0.2158 (0.2138, 0.2178) |
| 3 | 0.4275 | 2.0000 | 0.2137 (0.2111, 0.2163) | -0.0094 | 0.0177 | 0.2137 (0.2111, 0.2163) |
| 4 | 0.4304 | 1.9950 | 0.2157 (0.2137, 0.2177) | -0.0074 | 0.0103 | 0.2157 (0.2137, 0.2177) |

Table 6.3: Simulation results for the ratio of coefficients approach in a single cohort for genotypes GG versus gg.

In Table 6.3 $\theta_3$ and $\delta_3$ are twice as large as $\theta_2$ and $\delta_2$ from the previous table because an additive genetic model was used in the simulations. The ratio of coefficients estimates of $\eta_3$, $\widehat{\eta}_3$, also showed an attenuation towards the null. The attenuation for $\widehat{\eta}_3$ is larger than for $\widehat{\eta}_2$ and none of the confidence intervals for $\widehat{\eta}_3$ contain the set value of 0.2231. Again $\widehat{\eta}_{3,TS}$ coincided with $\widehat{\eta}_3$ for scenarios 3 and 4 due to the small value of the phenotype error variance.

The next section considers the ratio of coefficients approach in the MVMR meta-analysis model proposed in the previous chapter.

### 6.3.4 Meta-analysis simulations

Simulations using meta-analyses are performed to investigate whether the finite sample bias in the ratio estimate affects the pooled phenotype-disease log odds ratio from the multivariate MVMR meta-analysis model from the previous chapter. In these simulations ten cohort studies were considered to form a meta-analysis. For each meta-analysis the MVMR multivariate meta-analysis model was fit using maximum likelihood estimation as described in Section 5.2. Again $\eta$ was set to 0.2231 and in this instance each scenario was performed for 100 iterations. The results of these meta-analysis simulations are shown in Table 6.4.

| | $\widehat{\eta}_{\text{MVMR}}$ (95% CI) | Bias | MSE |
|---|---|---|---|
| 1 | 0.2416 (0.2343, 0.2488) | 0.0184 | 0.0017 |
| 2 | 0.2331 (0.2268, 0.2393) | 0.0099 | 0.0011 |
| 3 | 0.2309 (0.2241, 0.2377) | 0.0078 | 0.0013 |
| 4 | 0.2253 (0.2199, 0.2307) | 0.0022 | 0.0008 |

Table 6.4: Simulation results for $\widehat{\eta}$ from the MVMR model.

Table 6.4 shows that there is a positive bias in the pooled estimate of $\eta$ from the meta-analysis model and that for scenarios 1, 2 and 3 this was statistically significant from the set value 0.2231 at the 5% level. The positive bias conflicts with the attenuation towards the null found in the previous simulations. One possible explanation for the slight positive bias is that ten studies per meta-analysis may have been too few for the maximum likelihood algorithm to converge to the correct point on the likelihood surface. For example, the optimization algorithm may have found a local rather than a global maximum.

The simulations were rerun for a larger cohort size of 300,000 and increasing the number of studies per meta-analysis to 30, the results are shown in Table 6.5. The bias is reduced and in this instance and now the true value of $\beta_1$ is included in the confidence intervals of $\eta$ for the first three scenarios and was very close to the fourth. The estimate of $\eta$ was more precise in scenarios 3 and 4 with the smaller phenotype variance.

| | $\widehat{\eta}_{\text{MVMR}}$ (95% CI) | Bias | MSE |
|---|---|---|---|
| 1 | 0.2326 (0.2205, 0.2432) | 0.0010 | 0.0030 |
| 2 | 0.2253 (0.2173, 0.2332) | 0.0021 | 0.0016 |
| 3 | 0.2232 (0.2275, 0.2237) | 0.0001 | 0.000001 |
| 4 | 0.2219 (0.2209, 0.2229) | -0.0012 | 0.00003 |

Table 6.5: Simulation results for $\widehat{\eta}$ from the MVMR model using a cohort of 300,00 and 30 studies per meta-analysis.

## 6.4 Discussion

It is important to investigate the properties of the estimators of the phenotype-disease association for a Mendelian randomization analysis. For example, if the IV estimators are more biased than the direct estimators of the association then the application of the IV estimators would be redundant. This chapter investigated the ratio of coefficients approach using a complementary approximation to that used by Thompson *et al.* (2003), specifically a Taylor series approximation. Simulations were also carried out to investigate the properties of the ratio estimate in a single cohort and in several cohorts combined in a meta-analysis. The two estimates of the phenotype-disease association were investigated from the two genotype comparisons of the heterozygotes and rare homozygotes with common homozygotes.

A Taylor series approximation was used for the expectation and variance of the ratio of the means of two random variables. This meant a confidence interval could be derived for the ratio estimate as a comparison to that using Fieller's Theorem. The Taylor series and Fieller's Theorem confidence intervals will be similar unless either the gene-disease log odds ratios or gene-phenotype mean differences have a skewed distribution. The Taylor series approximation showed that the accuracy of the ratio estimate depends upon the variance of the genotype-phenotype association, which concurs with the findings of Thompson *et al.* (2003). Thomas *et al.* (2007) comment that this helps to explain why IV estimates are frequently more unstable than conventional ones. Additionally, under the Taylor series approximation the variance of the ratio estimate will be smaller if the gene-disease and gene-phenotype estimates are correlated because the covariance term has a negative sign.

The Taylor series approximations to the mean and variance of the ratio of two means are well known in the statistics literature. For example, confidence intervals for the ratio of means has been discussed in the health economics literature on the subject of cost-effectiveness ratios (Briggs & Fenn, 1998; Chaudhary & Stearns, 1996; O'Brien *et al.*, 1994). These studies considered a number of different methods for deriving confidence intervals for a ratio means and concluded that Fieller's method was preferable. They reached this conclusion because often in health economics either the numerator or denominator of the ratio follows a skewed distribution. It was found that Fieller's method is better at accommodating this skewness because it is not reliant upon the assumption of asymptotic normality. In a Mendelian randomization analysis the distributions of $\theta$ or $\delta$ may be more likely to be skewed in the presence of publication or reporting bias and hence Fieller's Theorem confidence intervals would be more appropriate in this instance. However, Fieller's method can be complicated to calculate so the Taylor series method is likely to provide an acceptable approximation in most circumstances with reasonable sample sizes. It can be noted that Walter *et al.* (2008) have recently proposed an equivalent method to Fieller's method for the confidence interval for the mean of the ratio of two normally distributed random variables based on a geometric argument.

The ratio of coefficients estimates of $\eta_2$ and $\eta_3$ were investigated using simulations for a single cohort. Some small attenuation bias was observed in these estimates which was more pronounced when the confounder was present. The attenuation in the estimate was also larger when the phenotype error term had a larger variance and was more pronounced for the comparison of the two homozygote genotypes, using $\eta_3$. However, the attenuation in the estimates of $\eta_2$ and $\eta_3$ is small and is unlikely to affect the conclusions of an analysis. Confidence intervals for $\eta_2$ and $\eta_3$ derived using the Taylor series were identical to the confidence intervals using Fieller's Theorem.

Simulations were performed to investigate the multivariate MVMR meta-analysis model incorporating all three genotypes from the previous chapter. In this instance a small positive bias was found in the pooled estimate of the phenotype-disease log odds ratio. The

MVMR model, which is a multi-variate normal model of dimension 4, was estimated using maximum likelihood. Riley *et al.* (2008) comment that multivariate normal meta-analysis models can encounter problems converging when the correlation between the outcomes is close to $\pm 1$. The difficulty in the convergence of the maximum likelihood algorithm occurs because as the correlation between the outcome measures approaches $\pm 1$ the probability density function of the multivariate normal distribution has develops a ridge instead of a pronounced global maximum. Therefore, it is possible that the maximum likelihood algorithm may converge to a local maximum instead of a global maximum.

In Section 6.3.2 it was explained that the probability of the occurrence of zero cases in the rare homozygotes was minimised by using a large cohort size. This was done in order to avoid generating an odds ratio of zero in this genotype group. In an applied meta-analysis it may be the case that studies with zero cases in the GG genotype may need to be included. Therefore, the discussion of continuity corrections in binary outcome studies is of relevance for Mendelian randomization meta-analyses.

Sweeting *et al.* (2004) compared several different methods for dealing with zero outcomes in binary outcome trials combined in a meta-analysis. The methods compared included the Mantel-Haenszel odds ratio (Mantel & Haenszel, 1959), the Peto method (Yusuf *et al.*, 1985) and the use of a constant continuity correction to avoid a zero odds ratio. The authors found that the common practice of applying the constant continuity correction in the context of a standard inverse variance weighted meta-analysis performed less well than the Mantel-Haenszel odds ratio and the Peto method with respect to the bias and coverage of the meta-analysis pooled estimate. Bradburn *et al.* (2007) also found that the Peto method was preferable for event rates less than 1% and that the Mantel-Haenszel odds ratio, without zero cell correction, was preferable for event rates greater than 1%.

These simulations have only investigated a limited set of factors. For a meta-analysis there are several other factors which could be investigated including publication bias and between study heterogeneity. Both of these issues are further complicated for multivariate meta-analyses, for example, the regression based tests for publication bias, such as Eg-

ger's test (Egger *et al.*, 1997), have not been generalised for multivariate outcomes. For a univariate meta-analysis one measure of between study heterogeneity is the $I^2$ statistic (Higgins & Thompson, 2002). The $I^2$ statistic is effectively an intra-class correlation coefficient for a meta-analysis since it expresses the magnitude of the between study variance of the pooled effect estimate with respect to the sum of the within and between study variances. The $I^2$ statistic is now commonly reported for meta-analyses (Ioannidis *et al.*, 2007) and like the tests for publication bias has also not been generalised for multivariate outcomes.

In conclusion, the ratio estimate of the phenotype-disease log odds ratio generally has minimal bias in the meta-analysis of binary outcome Mendelian randomization studies. However, the ratio estimate may be attenuated towards the null effect when the variance of the genotype-phenotype association is relatively large.

# Chapter 7

# Discussion & conclusions

## 7.1 Discussion

The work in this thesis has reviewed and developed statistical methods for the application of the Mendelian randomization approach within epidemiology. The work has taken the form of a review of the published literature, an investigation of estimators for binary outcome studies based on logistic regression, methods for meta-analysis and an investigation of the ratio of coefficients approach. In this discussion each of these areas is considered in turn followed by suggestions for further research and conclusions.

### 7.1.1 Literature review

The literature review considered the initiation of the idea behind the Mendelian randomization approach and its subsequent development within epidemiology. The rationale for the approach is based on Mendel's second law, which provides the basis for genotypes to be used as instrumental variables to infer the association between a phenotype and a disease. The approach can be traced back to Katan's letter to the Lancet (Katan, 1986) and a key milestone in its development was the recognition that it represented the use of genotypes as instrumental variables (Davey Smith & Ebrahim, 2003).

The literature review highlighted that when performing a Mendelian randomization analysis it is important to assess whether the genotype fulfills the necessary conditions to be an instrumental variable. Brunner *et al.* (2008, Tables S1 and S2) provides an example of an assessment of whether a genotype fulfills the conditions of an instrumental variable, for example, the distribution of the genotypes should be independent of the measured confounders.

The review identified that in epidemiology there is interest in estimating a causal association between a phenotype and a disease because a causal association implies that as well as knowing the magnitude of the association it is possible to know how a modification of the phenotype should reduce the risk of disease (Didelez & Sheehan, 2007b). An example of a modifiable factor that can reduce the risk of disease is periconceptual maternal folate supplementation which can reduce the risk of neural tube defects, such as spina bifida, in the foetus (Czeizel & Dudàs, 1992; PHOEBE, 2007; Wald & Sneddon, 1991).

The literature review also identified that there are a number of issues to consider when applying the Mendelian randomization approach. These issues include whether the selected gene is in linkage disequilibrium with other important genes, whether the risk of disease may be different for subgroups within the sample (population stratification), whether canalization (which is sometimes referred to as developmental compensation) could occur for the selected disease and whether the gene of interest has pleiotropic effects. Adequate sample size and statistical power are also important issues for studies applying instrumental variable analysis methods.

Instrumental variable analysis was commonly used in econometrics, causal inference and biostatistics prior to the development of the Mendelian randomization approach. The literature review identified that there is an array of statistical methods for various different types of analysis, the most well known being the method of two-stage least squares. Interestingly in the case where all variables are continuous the methods of two-stage least squares, the ratio of coefficients approach and the control function approach all produce equivalent parameter estimates. However, the review highlighted that instrumental vari-

able methods for studies reporting binary outcomes are not well developed which is problematic for epidemiological studies such as case-control and cohort studies. It was therefore identified that additional methods are required in this area.

### 7.1.2 The adjusted instrumental variable estimator

In Chapters 3 and 4 three estimators for the analysis of a binary outcome study are considered based on logistic regression. The direct estimator is the direct logistic regression of the disease status on the phenotype. The standard IV estimator is the logistic regression of the disease status on the predicted levels of the phenotype given the level of the genotype, which is what might be implemented following the principles underlying two-stage least squares estimation. The adjusted IV estimator includes the residuals as well as and the predicted levels of the phenotype (or the observed levels of the phenotype as was discussed in Section 3.3.2) from the regression of the phenotype on the genotype in the second stage logistic regression.

It was found that in the presence of an unmeasured confounder the direct estimate from the logistic regression of the disease status on the phenotype was positively biased, the standard IV estimator was found to be attenuated towards the null and the adjusted IV estimator was found to fall be between the two. However, it was also found that significance tests for the adjusted IV estimator had inflated type I error whereas the type I error values were at the nominal level for the standard IV estimator. Hence, the standard estimator could be used for significance testing and the adjusted estimator for detecting the magnitude of the phenotype-disease association.

The adjusted IV estimator has less bias because the residuals from the first stage regression capture some of the information about the confounding variable. Due to the construction of the simulations the bias in the direct, standard and adjusted estimators followed the relationships given by Zeger *et al.* (1988). These relationships explain the difference between marginal and conditional parameter estimates in generalised linear mixed models. In theory, it would be possible to use these theoretical expressions in the same way as the

reliability ratio, from measurement error models, to back-transform the marginal estimate to the conditional estimate. Realistically, it would be more convenient to compare the values of the three estimates in order to assess the likely magnitude of the effect of the unmeasured confounder.

In econometrics models of the form of the adjusted IV estimator are known as control function approaches (Nichols, 2006). It was found that the idea behind the adjusted estimator has been suggested previously using a probit regression by Rivers & Vuong (1988) which has then been described by Wooldridge (2002, procedure 15.1) and also by Nitsch *et al.* (2006). This modelling approach was also used by Nagelkerke *et al.* (2000) using binary variables for their equivalent of the gene, phenotype and disease variables. Nagelkerke *et al.* (2000) suggested that models of this form have reduced bias because the first stage residuals fulfill Pearl's back-door criterion (Pearl, 2000). Pearl's back-door criterion is closely related to the argument of Dawid (2002) about adjusting for one of two related confounders (Didelez & Sheehan, 2007a). If Pearl's back-door criterion does hold for the adjusted IV estimator then estimates of the phenotype-disease association should be assigned a causal interpretation.

It has also been suggested that in the form of the adjusted IV estimator proposed by Rivers & Vuong (1988), which uses the observed values of the phenotype with the first stage residuals as opposed to the predicted values, that the significance of the coefficient of the first stage residuals is a test for the presence of confounding (Wooldridge, 2002). In the form of the adjusted IV estimator in Chapter 3 this becomes a test of the equality between the coefficients for the first stage residuals and predicted values. Although, it can be noted that from a causal perspective Pearl (1998) is sceptical about the validity of tests for the presence of unmeasured confounding variables.

It was found that the adjusted IV estimator had inflated type I error, which means that the model is more likely to reject the null hypothesis when the null is true than the nominal rate of 5%. The inflation of the type I error increased as the magnitude of the unmeasured confounder increased. Therefore, an adjustment is required to increase the

standard errors of the parameter estimates from the adjusted IV estimator to reduce the type I error. The correction applied to the standard errors after the second stage of two-stage least squares was discussed but it is not trivial to apply the correction to logistic regression. Alternatively, the standard errors of the adjusted estimator could be derived using a method such as bootstrapping until the appropriate correction is known.

### 7.1.3 A note about Pearl's discussion of confounding

Pearl (1998) is sceptical about statistical tests for confounding because he argues that the statistical/associational definition of confounding is not strict enough within a causal framework. Pearl introduces a definition of confounding based on his *do()* operator and argues that the associational definition based on the association criterion does not ensure unbiased effect estimates and does not follow from the requirements of unbiasedness.

The associational definition of confounding can be expressed as; if $T$ is the set of variables that are not affected by $X$ then $X$ and $Y$ are not confounded by $T$ if for every member $Z$ of $T$:

(i) $Z$ is not associated with $X$, $P(x|z) = P(x)$, or

(ii) $Z$ is not associated with $Y$ within strata of $X$, $P(y|z,x) = P(y|x)$.

Specifically, Pearl presents four reasons why the associational criterion fails; permissiveness due to individuation, permissiveness due to small world assumptions, restriction due to barren proxies and restriction due to incidental cancellations. Summarising each in turn, permissiveness due to individuation says that although two variables $Z_1$ and $Z_2$ may separately not confound $X$ and $Y$ they may still jointly confound $X$ and $Y$. Permissiveness due to small-world assumptions says that it is impractical to use the associational criterion to test for the absence of confounding since an investigator can never know that they have data on every potential confounder. Restriction due to barren proxies says that the associational criterion of confounding is not able to exclude what are termed barren proxies. A barren proxy is a variable which has no influence on $X$ or $Y$ but is a proxy

for factors that do. Pearl then provides a modified form of the associational criterion to exclude barren proxies by splitting the set $T$ into two subsets $T_1$ and $T_2$. Then $X$ and $Y$ are said to be unconfounded if:

1. $T_1$ is not associated with $X$, and

2. $T_2$ is not associated with $Y$ given $X$ and $T_1$.

However, this definition is still not strong enough because it cannot overcome the fourth criticism of restriction due to incidental cancellation. This says that the association criterion fails to identify incidental cancellations. An incidental cancellation is an association that might falsely classify some unconfounded situations as confounded and adjusting for such a false confounder could bias the effect estimate.

Pearl's causal definition of confounding says that if $P(y|do(x))$ is the probability of the response $Y = y$ under an intervention $X = x$ calculated according to the causal model of the data generating process, then $X$ and $Y$ are not confounded if and only if,

$$P(y|do(x)) = P(y|x). \tag{7.1}$$

Pearl's causal definition of confounding is similar the comparability view of confounding of Greenland *et al.* (1989).

Pearl (1998) then defines what he terms stable unbiasedness which is robust to changes in the size of the causal effect and remains intact as long as the causal model is unchanged. Pearl effectively argues that whilst it is not possible to test for confounding it is possible to test for stable unbiasedness which is a closely related idea. Pearl proposes the use of his back-door criterion for identifying conditions of unbiasedness. The subtlety is that the back-door criterion can guarantee unbiasedness for models of the form of Figure 7.1(a), since $X - E - Z - A - Y$ is a back-door path, but not of the form of Figure 7.1(b) because there is no back-door path from $X$ to $Y$ since these variables remain correlated after conditioning on $Z$. Although it is possible that a special choice of the parameter

values in Figure 7.1(b) may coincidentally result in an unbiased effect estimate.

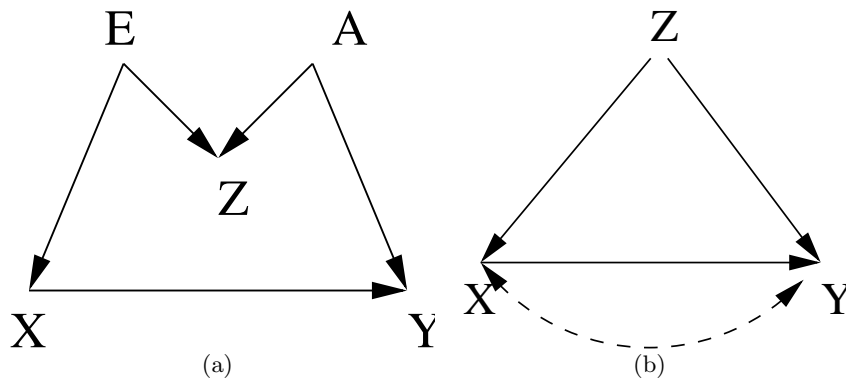



Figure 7.1: DAGs demonstrating models for which stable unbiasedness can and cannot be proved, taken from Pearl (1998, Figures 1 & 2).

Pearl argues that it is possible to disqualify a pair of variables $X$ and $Y$ as stably unconfounded using his operational test for stable nonconfounding. Variables $X$ and $Y$ are said to be stably uncofounded if and only if they have no common ancestor in a causal diagram. Pearl's operational test states that given a variable $Z$ which is independent of $X$ and possibly associated with $Y$ then $X$ and $Y$ are not stably unconfounded if either of the following criteria are violated:

(i) $P(x|z) = P(x)$, or,

(ii) $P(y|z, x) = P(y|x)$.

Confusingly, these two conditions are almost identical to the two conditions of the association criterion. Therefore, if just one variable if found that violates either of these conditions proves that $X$ and $Y$ are not stably unconfounded.

Pearl also links the operational test of stable confounding to the concept of collapsibility (see Section 1.3.1) because a violation of collapsibility will violate stable unbiasedness. This detracts from the possibly that the adjusted IV estimator might have a causal interpretation when used with the logit link despite the fact that it fulfills Pearl's back-door criterion and supports the case that the adjusted IV estimator could have a causal interpretation for collapsible link functions such as the identity and log links.

### 7.1.4 Multivariate meta-analysis

Chapter 5 considered multivariate meta-analysis models for Mendelian randomization analyses. Bivariate meta-analysis models have been proposed for Mendelian randomization analyses based on the ratio of coefficients approach in which the gene-disease and gene-phenotype effect estimates from each study are synthesised in the same model (Minelli *et al.*, 2004; Thompson *et al.*, 2005). This modelling approach assumes that the gene-disease log odds ratio can be treated as a continuous variable, an assumption which is commonly used in univariate meta-analyses of binary outcome studies.

Existing meta-analysis models are typically based on the comparison of the heterozygotes or rare homozygotes, or both, with the common homozygotes. The bivariate models were extended to incorporate both genotype comparisons and hence all three genotypes. The motivation for this is that it is important to be able to maximize the information that can be included in an analysis in order to increase its statistical power. It was hypothesised that the phenotype-disease association is common across both genotype comparisons and as a consequence the estimated genetic model, as proposed in the genetic model-free approach (Minelli *et al.*, 2005a,b), should be equivalent whether estimated using gene-disease or gene-phenotype associations. In an example meta-analysis reporting the necessary outcome measures no evidence was found to contradict these assumptions but further applied examples are required to test these assumptions.

The presentation of the results of multivariate meta-analysis was also discussed. A four column forest plot was presented which makes it easier to assess trends in each study's results across the four outcome measures. For Mendelian randomization meta-analyses it is possible to plot the study-level gene-disease estimates versus the study-level gene-phenotype estimates. A plot of this type helps to assess the pooled phenotype-disease association by comparing whether the points fall along the line with gradient equal to $\widehat{\eta}$. For meta-analyses employing the genetic model free approach, and considering two genotype comparisons, it is possible to assess the pooled estimate of the genetic model-free parameter $\lambda$ by plotting either the pair gene-disease or the pair of gene-phenotype

outcome measures against one another. Similarly to the previous plot, the points on the plot should fall along the line with gradient equal to $\widehat{\lambda}$.

Additionally, in a meta-analysis applying the Mendelian randomization approach it is possible that not all studies will report both gene-disease and gene-phenotype outcome measures. Under the assumption that these outcome measures are missing at random it is possible to include all the available information in the meta-analysis.

### 7.1.5 The ratio of coefficients approach

In Chapter 6 the ratio of coefficients approach was investigated using a Taylor series approximation. The Taylor series approximation showed that there can be a small attenuation in the ratio estimate for the phenotype-disease association, however this attenuation should not be large enough to alter the conclusions of an analysis as long as the variance of the genotype-phenotype association is small. The small attenuation in the ratio estimate was demonstrated in the single cohort simulations.

The application of the Taylor series also provided an estimate of the variance of the ratio estimate. For large sample sizes the confidence interval for the ratio estimate should be similar using either the Taylor series or Fieller's Theorem methods. A confidence interval derived using Fieller's Theorem is considered more appropriate if either or both of the gene-disease or gene-phenotype associations have skewed distributions.

## 7.2 Topics for further research

The Mendelian randomization approach is relatively new and there is potential for developments in both applied and methodological research. This section discusses some possible ares for further research using the approach.

### 7.2.1 Applied research

The Mendelian randomization approach has gained popularity due to the recent increased availability of genetic data in epidemiological studies. In particular the collection of genotype, phenotype and disease status information within the same study is becoming increasingly common. This is demonstrated by the creation of large-scale Biobanks such as the UK Biobank (Palmer, 2007) and large-scale collaborative genetic epidemiological studies such as the Wellcome Trust case control consortium (The Wellcome Trust Case Control Consortium, 2007).

In econometrics the use of multiple instrumental variables in an analysis is common. Brunner *et al.* (2008) have applied this idea using three tagging SNPs in the gene for C-reactive protein as instrumental variables. Another example of the use of multiple IVs is given by Kivimäki *et al.* (2008, Appendices 1 & 2) who compared analyses using one IV to analyses using multiple IVs. The authors found that because the multiple SNPs, known as a haplotype, were in linkage disequilibrium that the analyses using either one or multiple SNPs were very similar. However, there is scope for further analyses using multiple genetic instrumental variables.

A related idea to including multiple polymorphisms in an analysis termed 'factorial Mendelian randomization' was proposed by Davey Smith & Ebrahim (2003). The idea was developed in response to a criticism that the rare homozygotes may not influence the disease to a great extent. A factorial Mendelian randomization analysis would involve considering several polymorphisms at more than one locus which influence an intermediate phenotype, then combinations of polymorphisms at different loci could be found that produce differences in the intermediate phenotype that are substantial enough to generate detectable effects on disease risk. If the loci are not in linkage disequilibrium interest would be in the groups in which the combination of polymorphisms produce the most extreme difference in the phenotype. However, this idea has not yet been implemented.

With the increasing availability of genetic data there is scope to apply Mendelian random-

ization analyses to a wide range of polymorphisms, phenotypes and diseases in medical research as long as there is sufficient biological knowledge about the underlying causal pathway of disease.

### 7.2.2 Methodological research

There is considerable scope for investigating statistical models for Mendelian randomization analyses. Simulations similar to those performed for the adjusted IV estimator could be performed investigating other methods. For example, for the case of a continuous outcome measure, using an identity link function, Dunn *et al.* (2005) and Dunn & Bentall (2007) compared the adjusted IV estimator and a generalised method of moments approach and found that the methods gave similar but not identical parameter estimates. Hence the adjusted IV estimator should be investigated using different generalized linear models at the second stage and steps towards this are taken in Appendix C.

Some of the methods for instrumental variable methods that could be investigated for Mendelian randomization analyses include the generalized method of moments approach for estimating relative risks (Windmeijer & Santos Silva, 1997), the marginal structural mean modelling approach (Robins *et al.*, 2000) and the logistic structural mean modelling approach (Vansteelandt & Goetghebeur, 2003). It can be noted that a detailed discussion of instrumental variable theory for binary outcomes has been given from econometric perspective by Chesher (2007).

A drawback of the logistic structural mean modelling approach of Vansteelandt & Goetghebeur (2003) is that a fully saturated model is first of all fitted to the data. This means that all the main effects, of the phenotype and measured confounding variables, and their interaction terms with one another must be included in the model. However, this becomes impractical if there are more than a small number of measured confounders. Additionally, the estimation algorithm for fitting the logistic structural mean model relies on a grid search to find the value of the phenotype-disease log odds ratio which minimises the covariance between the genotype and the log odds of disease. Hence the fitting algorithm

does not automatically return a standard error for the estimated phenotype-disease log odds ratio. A confidence interval for the causal effect has to be derived using another method.

In this thesis the scenario of a continuous phenotype with a binary disease outcome has been considered, however, other scenarios of Mendelian randomization analyses could be considered such as a binary phenotype and a binary outcome measure as used by Nagelkerke *et al.* (2000).

There is scope to apply the theory used in this thesis to the discussion of Mendelian randomization methodology of Didelez & Sheehan (2007b). In particular, these authors discuss three estimators based on causal inference for use with a binary disease outcome. These estimators are the average causal effect (ACE), causal relative risk (CRR) and causal odds ratio (COR). These estimators are all based on an Equation 8 in the paper which the authors state cannot be evaluated because the distribution of the confounder is unknown. Equation 8 in the paper is the same as Equation B.11 without the assumption that the confounder is normally distributed in this thesis. The work in this thesis has shown that if the confounder can be assumed to be normally distributed then the approximation which results in Equation 3.45 can be used. Alternatively, the Neuhaus approximation, as given in Equation B.16, which relies on fewer assumptions, could also be used to approximate their Equation 8. In either case it should be possible to derive an algebraic approximation for Equation 8 of Didelez & Sheehan (2007b) and hence derive an algebraic approximation for each of the ACE, CRR and COR parameters. The approximations could then be compared with the authors' approach of using numerical integration to evaluate the three estimators (Meng, 2008).

### 7.2.3 Meta-analysis

Instrumental variable analysis originated in econometrics and causal inference in which meta-analysis is not as common as in biostatistics and due to its relatively early stage of development a number of meta-analysis topics have not been discussed specifically for

Mendelian randomization analyses.

The issue of publication bias has not been investigated in this thesis for Mendelian randomization meta-analysis estimates. Publication bias is usually investigated using a funnel plot of effect estimate variability plotted against the effect estimate. There are a number of statistical tests to investigate publication bias including the regression tests of Egger *et al.* (1997), Harbord *et al.* (2006) and Peters *et al.* (2006) and non-parametric methods such as the trim and fill method (Duval & Tweedie, 2000a,b). A useful addition to a funnel plot to help assess the possible presence of publication bias is to add contours of statistical significance (Palmer *et al.*, 2008b; Peters *et al.*, 2008).

The product normal formulation (Spiegelhalter, 1998) was used in the PNF model in Chapter 5, however, this method has not been extensively investigated in the Bayesian literature. In particular, the product normal formulation is reliant upon the sequential parameter updating under Gibbs sampling. As a result, it may be possible to induce an informative prior distribution on the correlation between the outcomes when only vague prior distributions are assumed for the other parameters in the model.

Also an individual patient data meta-analysis has not been performed implementing the Mendelian randomization approach. The panel-data versions of the instrumental variable models proposed in econometrics, such as Windmeijer & Santos Silva (1997), could be investigated for individual patient data meta-analysis models.

## 7.3 Conclusion

The aim of the work in this thesis has been to extend statistical modelling approaches for Mendelian randomization analyses. For the analysis of a single study the adjusted instrumental variable estimator has reduced bias compared to a standard estimator based on the principles of two-stage least squares. However, the standard estimator is preferable for testing the null hypothesis. In further work the adjusted IV estimator could be compared with other modelling approaches from biostatistics, causal inference and econometrics.

The meta-analysis methods investigated in this thesis combine and extend previous meta-analysis. Meta-analysis is particularly relevant for the Mendelian randomization approach since instrumental variable analyses require large sample sizes to detect small effect sizes compared with standard analysis methods.

The application of genetic polymorphisms as instrumental variables within epidemiology under the banner of the Mendelian randomization approach has recently received considerable discussion. Whilst methods for instrumental variable analysis using continuous outcome measures are well known, instrumental variable methods for discrete or categorical outcome measures are not straightforward and are less well developed. The work in this thesis is designed to contribute to the research into these methods for a single study and for the meta-analysis of several studies.

# Appendix A

# Glossary

Some genetic terminology is listed below, the entries have been compiled from Lawlor *et al.* (2008d, Table 1), Elston *et al.* (2002) and Balding *et al.* (2007).

Alleles: variant forms of a genetic polymorphism.

Canalization: the process by which potentially disruptive influences on normal development from genetic and environmental variations are damped by compensatory developmental processes, i.e. a phenotype is kept within narrow boundaries in the presence of disturbing environments or mutations.

Chromosome: an organized structure of DNA and protein that is found in cells. A chromosome typically contains genes, regulatory elements and other nucleotide sequences. Humans have 22 pairs of autosomal chromosomes and 1 pair of sex chromosomes.

Deoxyribonucleic acid (DNA): a molecule containing genetic instructions used in the development and functioning of living organisms. DNA contains the instructions to construct the components of cells, including proteins and ribonucleic acid (RNA) molecules. DNA has four nucleotide bases; Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). The two strands in the double-helix are complementary (sense and anti-sense) such that A combines with T, and C with G.

Diploid: a cell with two versions of each chromosome, one from the father one from the mother.

Gamete: a sex cell, sperm in males, egg in females. Two haploid gametes fuse to form a diploid zygote.

Gene: comprises a sequence of DNA made up of introns, exons and regulatory regions, related to transcription of a given Ribonucleic acid (RNA).

Genotype: two alleles inherited at a specific locus, if they are the same the genotype is homozygous, if different heterozygous.

Hardy-Weinberg equilibrium: describes the distribution in a population of the genotypes for a genetic locus given the frequencies of the common and rare alleles. In theory genotype frequencies in a population are in equilibrium unless specific disturbing influences are introduced. It is based on the assumptions that there is random mating, a large population, and no migration, mutation or selection.

Haploid: a cell with a single version of each chromosome.

Haplotype: describes the combination of alleles from linked loci found on a single chromosome.

Heterozygote: a single locus genotype consisting of two different alleles.

Homozygote: a single locus genotype consisting of two versions of the same allele.

Linkage: occurs when particular genetic loci are inherited jointly. For example, genetic loci on the same chromosome tend to segregate together during meiosis.

Linkage disequilibrium (LD): the correlation between allelic states at different loci within the population. LD describes a state that represents a departure from the hypothetical situation in which all loci exhibit complete independence known as linkage equilibrium. Note that LD is different from linkage.

Locus: the position in a DNA sequence, it can refer to different scales such as a SNP, a large region of DNA sequence or even a whole gene.

Meiosis: the process by which haploid gametes are formed from diploid somatic cells.

Mitosis: the process by which a somatic cell is replaced by two daughter somatic cells.

Mutation: a process that changes an allele.

Panmixia: describes random mating, it involves the mating of individuals regardless of any physical, genetic, or social preference.

Phenotype: the observed characteristic under study, it could be measured as either a quantitative, binary or categorical variable.

Pleiotropy: polymorphisms that have multiple phenotypic effects.

Single-nucleotide polymorphism (SNP): genetic variations in which one base in the DNA is altered.

Zygote: an egg cell that has been fertilized by a sperm cell.

# Appendix B

# Estimates from GLMs with a random intercept

This appendix gives the derivation of relationships between marginal and conditional parameter estimates in generalized linear mixed models as the derivations were not provided by Zeger *et al.* (1988). The relationships given by Neuhaus *et al.* (1991) are also discussed.

## B.1 The Zeger equations

The derivation of the equations was not provided by Zeger *et al.* (1988) although an outline of their derivation has been given by Hardin & Hilbe (2003, p96–97).

The modelling framework for GLMs with a random intercept is given by,

$$g(E(Y_i|u_i, X_i)) = X'\beta + u_i, \quad \text{where } u_i|X_i \overset{iid}{\sim} N(0, \sigma^2), \tag{B.1}$$

where $g(\cdot)$ is the link function of the GLM. Denoting the population averaged parameters $\beta_m$ and the subject-specific parameters $\beta_c$ the marginal and conditional likelihoods are

given by,

$$L_m = P(Y|X, \beta_m) \tag{B.2}$$

$$L_c = \int P(Y|u, \beta_c)P(u|X)du. \tag{B.3}$$

### B.1.1 Identity link

For the identity link,

$$\begin{aligned}
\mu_m &= E(y) \\
&= \int g^{-1}(X\beta_c + u)F(u)du \\
&= \int (X\beta + u)\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{u^2}{2\sigma^2}\right) du \\
&= X\beta + \int u\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{u^2}{2\sigma^2}\right) du \\
&= X\beta + E(u) \\
&= X\beta,
\end{aligned} \tag{B.4}$$

since $E(U) = 0$. Hence the marginal and conditional estimates under the identity link are identical.

### B.1.2 Log link

For the log link,

$$\begin{aligned}
\mu_m &= E(y) \\
&= \int g^{-1}(X\beta + u)F(u)du \\
&= \int \exp(X\beta + u)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-u^2}{2\sigma^2}\right)du \\
&= \exp(X\beta)\int \exp(u)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-u^2}{2\sigma^2}\right)du \\
&= \exp(X\beta)\exp(\sigma^2/2) \\
&= \exp\left(X\beta + \frac{\sigma^2}{2}\right). \qquad \text{(B.5)}
\end{aligned}$$

The step from the fourth to the fifth lines is possible since $U \sim N(0, \sigma^2)$ then $V = \exp(U)$ has a lognormal distribution, with expectation, $E(V) = \exp(\sigma^2/2)$ (Johnson & Kotz, 1994, p211). Equivalently, it can be noted that this expression is the form of the moment generating function of the normally distributed random variable $U$ (Richardson *et al.*, 1987). Hence with a log link the marginal intercept is different from the conditional estimate but the other parameters are the same.

### B.1.3 Probit link

The derivation of the relationship for the probit link was given by Smith & Diggle (1998),

$$\begin{aligned}
\mu_m &= E(y) \\
&= \int g^{-1}(X\beta + u)\phi(u)du \\
&= \int \Phi(X\beta + u)\phi(u)du. \qquad \text{(B.6)}
\end{aligned}$$

Then it can be noted that where $Z \sim N(0,1)$,

$$\begin{aligned}\Phi(X\beta + u) &= P(Z \leqslant X\beta + u) \\ &= P\left(Z - u \leqslant \frac{X\beta}{\sqrt{1+\sigma^2}}\right) \\ &= \Phi\left(\frac{X\beta}{\sqrt{1+\sigma^2}}\right). \end{aligned} \tag{B.7}$$

This was given by Owen *et al.* (1964) and hence it is possible to continue,

$$\begin{aligned}\mu_m &= \int \Phi\left(\frac{X\beta}{\sqrt{1+\sigma^2}}\right)\phi(u)du \\ &= \Phi\left(\frac{X\beta}{\sqrt{1+\sigma^2}}\right). \end{aligned} \tag{B.8}$$

This expression implies that for the probit link function the conditional parameter estimate is multiplied by the generalized form of the $s$ parameter of Gilmour *et al.* (1985), which is equal to $\sqrt{1-\text{ICC}}$ (ICC: intra-class correlation coefficient).

### B.1.4 Logit link

The relationship between marginal and conditional parameter estimates under the logit link function simply includes an extra constant, $c$, compared with the equation for the probit link. This is because it is based on using a probit approximation to the standardised logistic cumulative distribution function. As such an identical argument holds for the derivation of the relationship for the logit link to the one given for the probit link.

The value of the constant $c$ is derived from a comparison of the cumulative distribution functions of the normal and logistic distributions. In particular, the variance of the standardized logistic distribution is $\pi^2/3$. The standardised forms of the cumulative dis-

tribution functions of the normal and logistic distributions are given by $F_1$ and $F_2$ below,

$$F_1(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-1/2u^2} du \tag{B.9}$$

$$F_2(x) = \frac{1}{1 + \exp(-\pi x/\sqrt{3})}. \tag{B.10}$$

It is known that $F_1(16x/15) \approx F_2(x)$ (Johnson & Kotz, 1970) although a proof of this was not given apart from the comparison of the functions shown in Figure B.1, which are in good agreement.

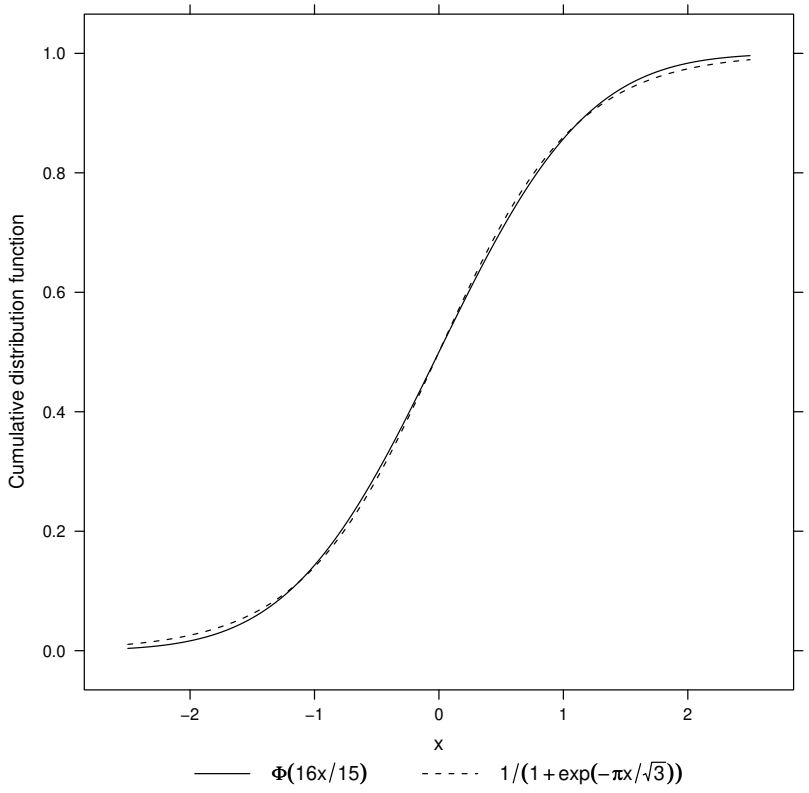



Figure B.1: Comparison of the Probit approximation to the standardized logistic cdf, adapted from Carroll *et al.* (1995, Figure 3.5).

Therefore, in Equation 3.45 $c = \frac{16\sqrt{3}}{15\pi}$ (0.5881 to 4 d.p.). It can be noted that other very similar values for $c$ have been proposed such as $\frac{\sqrt{3.41}}{\pi}$ (0.5878 to 4 d.p.) and 1/1.7017456 (0.5876 to 4 d.p.) (Johnson & Kotz, 1970).

Therefore, the derivation of the approximate relationship between marginal and condi-

tional parameter estimates for the logit link is given by,

$$\begin{aligned}
\mu_m &= E(y) \\
&= \int g^{-1}(X\beta + u)\phi(u)du \\
&= \int \frac{\exp(X\beta + u)}{1 + \exp(X\beta + u)}\phi(u)du \qquad \text{(B.11)} \\
&\approx \int \Phi\left(c(X\beta + u)\right)\phi(u)du \\
&= \Phi\left(\frac{X\beta}{\sqrt{1 + c^2\sigma^2}}\right). \qquad \text{(B.12)}
\end{aligned}$$

Line B.11 is Equation 8 of Didelez & Sheehan (2007b) assuming that $U \sim N(0, \sigma^2)$.

## B.2 The Neuhaus equations

In a series of papers including; Neuhaus *et al.* (1991), Neuhaus (1993), Neuhaus & Jewell (1993) and Neuhaus (1998), Neuhaus proposed alternative forms for the relationships between marginal and conditional parameter estimates for generalised linear mixed models. These are given below as per Neuhaus & Jewell (1993, Table 3), where $\mu$ denotes the linear predictor of the GLM,

$$\text{Identity}: \quad \beta_m \approx \beta_c \qquad \text{(B.13)}$$

$$\text{Log}: \quad \beta_m \approx \beta_c \qquad \text{(B.14)}$$

$$\text{Probit}: \quad \beta_m \approx \beta_c \cdot \frac{E[\phi\Phi^{-1}(\mu)]}{\phi\Phi^{-1}(E(\mu))} \qquad \text{(B.15)}$$

$$\text{Logit}: \quad \beta_m \approx \beta_c \cdot \left\{1 - \frac{\text{var}(\mu)}{E(\mu)E(1-\mu)}\right\}. \qquad \text{(B.16)}$$

The expression for the logit link function was also given by Gail (1988).

## B.3 Discussion

Other formulae based on the Zeger and Neuhaus approaches have been given by several authors. For example, the relationship between the marginal and conditional parameter estimates for models using the complementary log-log link has been given in the form of Zeger by Wang & Louis (2003) and in the form of Neuhaus by Jewell & Shiboski (1990). The relationship between marginal and conditional parameter estimates when the random effects are exponentially and double-exponentially distributed instead of normally distributed have been given in the form of Zeger by Ritz & Spiegelman (2004). For a probit link function Chao *et al.* (1997) demonstrated the circumstances under which the Zeger and Neuhaus approximations are equivalent.

In particular, the relationship between subject-specific and population averaged or conditional and marginal parameter estimates for logistic regression has been used widely in statistical research. The following are a few examples; (Carroll *et al.*, 1984), Stefanski (1985), Liang & Liu (1991), Hu *et al.* (1998), Carroll & Stefanski (1994), Carroll *et al.* (1995, Section 3.9.2), Ten Have *et al.* (1996), (Diggle *et al.*, 2002) and Ten Have *et al.* (2003).

# Appendix C

# An adjusted instrumental variable estimator: results for other link functions

## C.1 Introduction

This chapter gives results of further simulations like those in Chapter 4 for different generalized linear models at the second stage of the analysis, namely; probit, linear and Poisson regression models under the inverse normal cumulative distribution, identity and log link functions. The motivation for this work is to compare the standard and adjusted IV estimators for these different generalized linear models and relates to theory given in Appendix B.

## C.2 Probit link

The simulations performed to investigate the use of probit regression at the second stage were identical to the simulations in Chapter 4 which used logistic regression at the second

stage apart from that the inverse normal cumulative distribution function was used to generate the probabilities of disease instead of the inverse of the logit function,

$$p_i = \Phi^{-1}(\beta_0 + \beta_1 x_i + \beta_2 u_i). \tag{C.1}$$

Considering the simulations in Chapter 4 and the closely related theory for the probit link in Appendix B it is expected that the two IV estimators will exhibit attenuation bias and that in the presence of the confounding variable the direct estimator will be positively biased. The results of these simulations for the estimates of $\beta_1$ are given in Figure 4.12 in Chapter 4.

As under the logit link the coverage of the 95% Wald confidence intervals for $\beta_1$ were investigated, as shown in Figure C.1, and the adjusted IV estimator again has better coverage than the standard estimator because of its smaller bias.

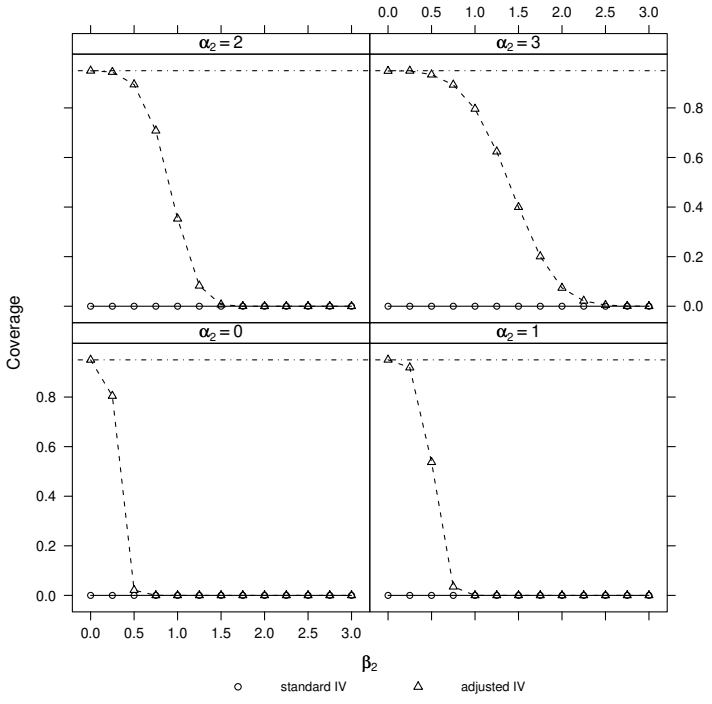



Figure C.1: Coverage of the Wald test for $\beta_1$ with the probit link.

The type I error of the Wald tests for the estimates of $\beta_1$ are shown in Figure C.2.

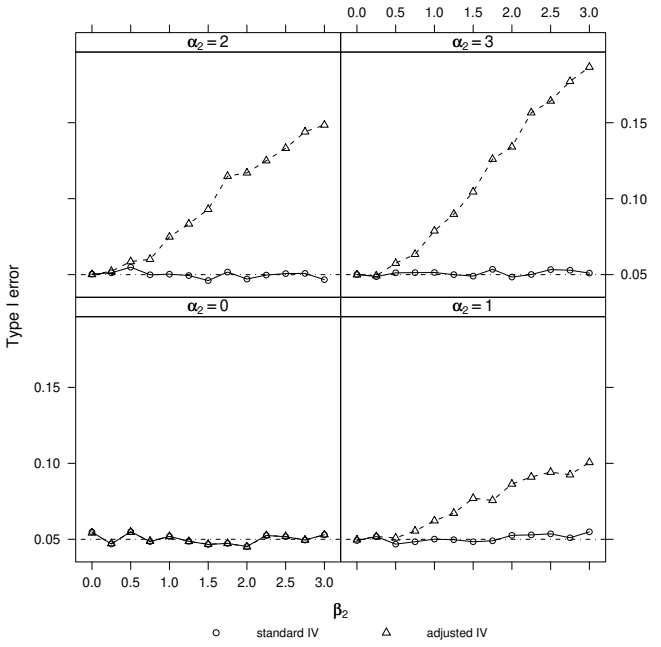



Figure C.2: Type I error of the Wald test for $\beta_1$ with the probit link.

As with the logit link simulations the adjusted estimator had inflated type I error with respect to the nominal level of 5% whereas the standard estimator was at the nominal level. For completeness the estimates of $\beta_0$ from the three estimators are shown in Figure C.3.

As with the logit link under the probit link the estimates of $\beta_0$ also follow the Zeger relationship between the marginal and conditional estimates.

In Appendix B the formulae of Zeger and Neuhaus relating marginal and conditional parameter estimates in generalized linear mixed models were given. Figure C.4 shows the comparison of the attenuation in the standard IV estimator, which is in line with the Zeger formula, and the attenuation in $\beta_1$ as given by the Neuhaus formula. It can be seen that the Neuhaus and Zeger approximations follow the same form although they are not identical they become asymptotic as the value of the confounder coefficient $\beta_2$ increases.

The agreement between the Zeger and Neuhaus methods is better for the probit link than for the logit link as shown in Figure 4.11.

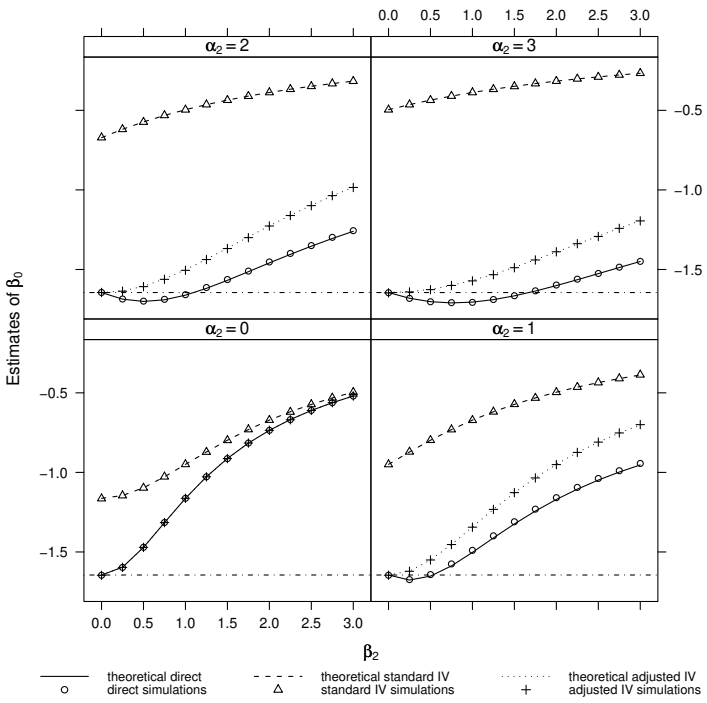



Figure C.3: Theoretical and simulated estimates of $\beta_0$ with the probit link.

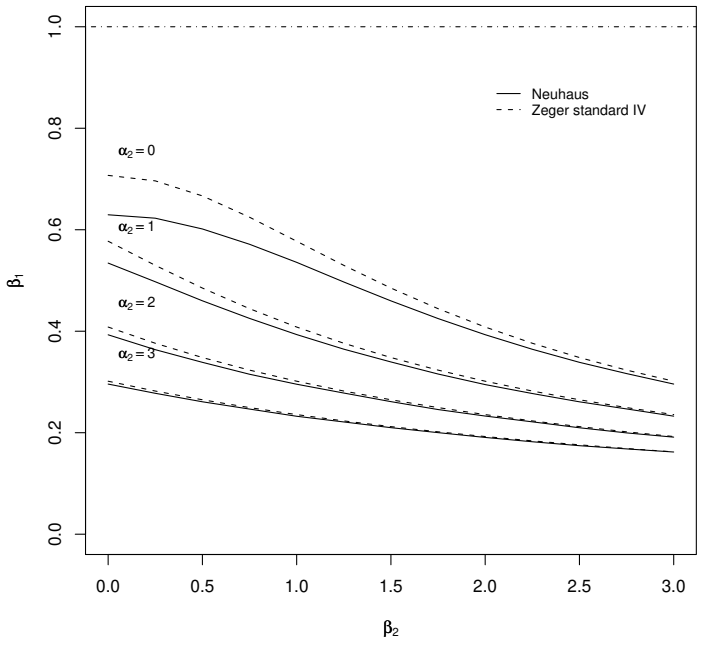



Figure C.4: Comparison of the Zeger and Neuhaus formulae for the standard IV estimator with a probit link function.

## C.3 Identity link

For the simulations using the identity link a continuous variable was simulated to represent disease status instead of a binary outcome disease variable,

$$y_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 u_i,\ \sigma_y^2) \tag{C.2}$$

From the theory in Appendix B it is expected that the standard and adjusted estimators will not be attenuated because there is no attenuation of marginal with respect to conditional estimates under the identity link. It is also expected that the standard and adjusted estimators of $\beta_1$ will be at the set value of $\beta_1$ because large sample sizes were used in the cohorts and two-stage least squares should provide a consistent estimate of $\beta_1$. Figure C.5 shows that these expectations were found to hold.

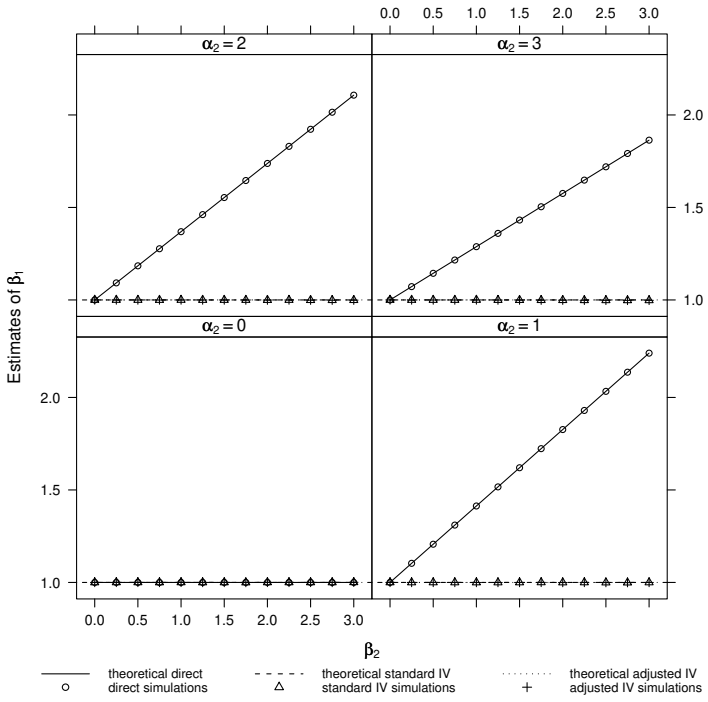



Figure C.5: Theoretical and simulated estimates of $\beta_1$ with the identity link.

Figure C.6 shows the theoretical and simulated values of $\beta_0$ under the identity link. Under the identity link $\beta_0$ was set to 0.05.

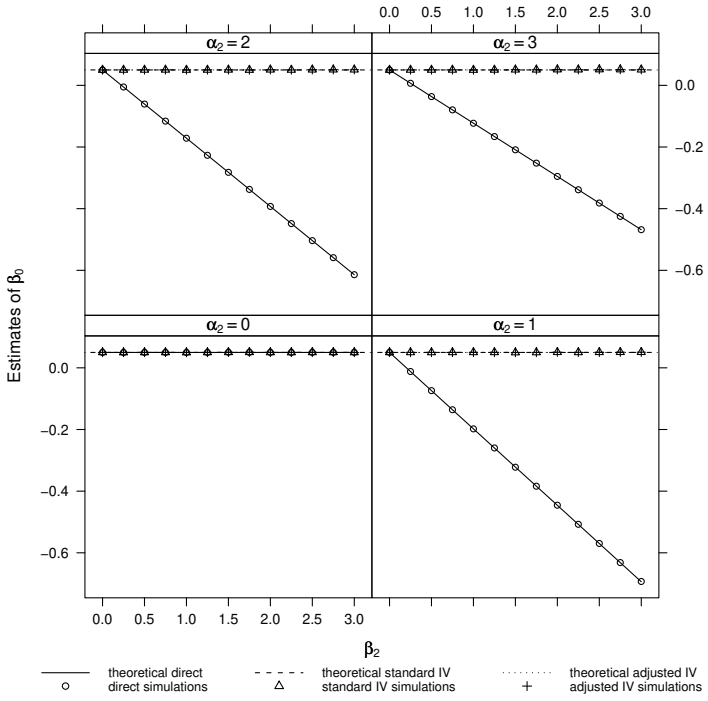



Figure C.6: Theoretical and simulated estimates of $\beta_0$ with the identity link.

The coverage of the Wald test of the standard and direct estimators of $\beta_1$ under the identity link is shown in Figure C.7. In this instance since the standard and adjusted estimators were unbiased the coverage will not be determined by the bias as was the case for the logit and probit links. For the panel with $\alpha_2 = 0$ the coverage of the standard estimator was greater than the nominal 95% whereas the adjusted estimator was exactly at 95%. For the other three panels with $\alpha_2 = 1, 2, 3$ the coverage of the standard estimator was approximately 100% whilst the coverage of the adjusted estimator started at 95% but decreased as the value of $\beta_2$ increased.

The type I error of the Wald test of the estimators of $\beta_1$ under the identity link is shown in Figure C.8. The type I error of the Wald test of the standard estimator is at the nominal level of 5%. The type I error of the adjusted estimate is inflated when the magnitude of the confounder is large.

Figure C.9 is for the same set of simulations, it shows that the type I error of the standard IV estimator and the type I error of two-stage least squares, using corrected standard

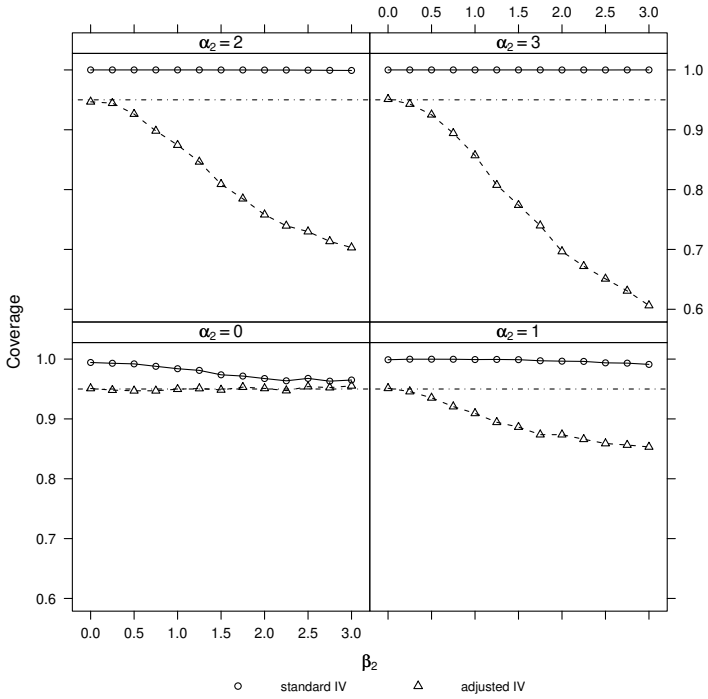



Figure C.7: Coverage the Wald test of $\beta_1$ under the identity link.

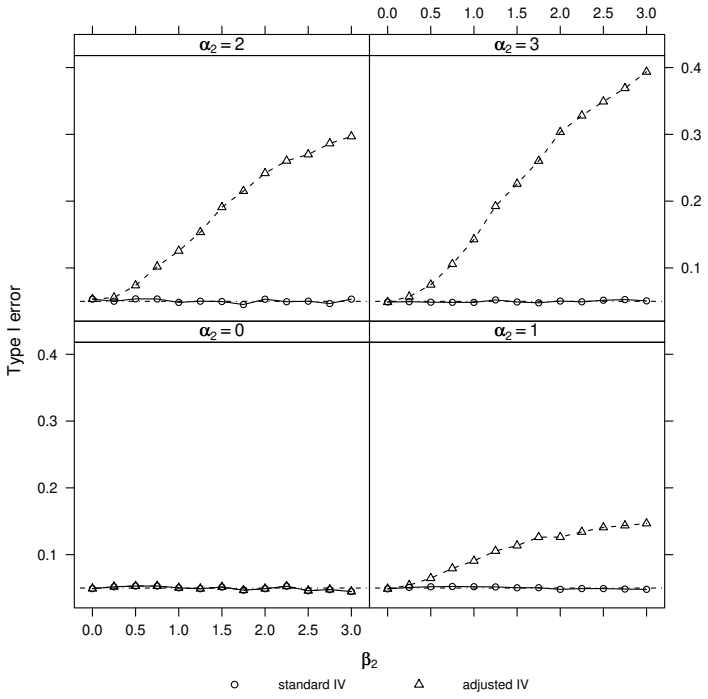



Figure C.8: Type I error of the Wald test of $\beta_1$ under the identity link.

errors are both at the nominal level in a large sample.

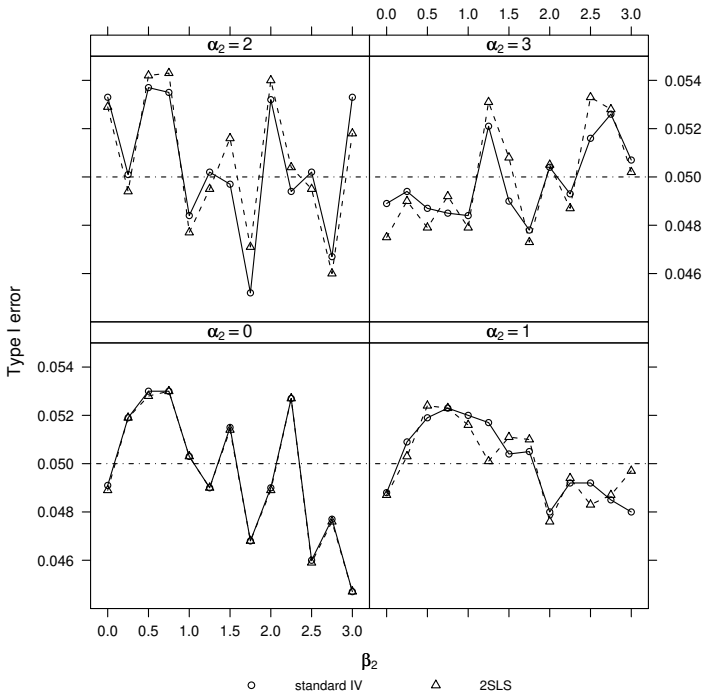



Figure C.9: Comparing the type I error of the Wald test of $\beta_1$ for the standard IV & two-stage least squares.

In Chapter 1 it was described that a test of the strength of an instrumental variable was given by the $F$ statistic from the first stage linear regression. In Chapter 4 Figure 4.10 investigated a related measure of strength the $R^2$ value from the first stage regression. Figure C.10 examines an idea to assess the magnitude of the unmeasured confounder by considering the correlation between the first and second stage residuals from the two-stages of two-stage least squares estimation.

In the panel with $\alpha_2 = 0$ the value of the correlation between the first and second stage residuals is not informative. However, in the subsequent panels when the confounder acts truly as a confounder then the value of the correlation increased as the magnitude of the confounder increased.

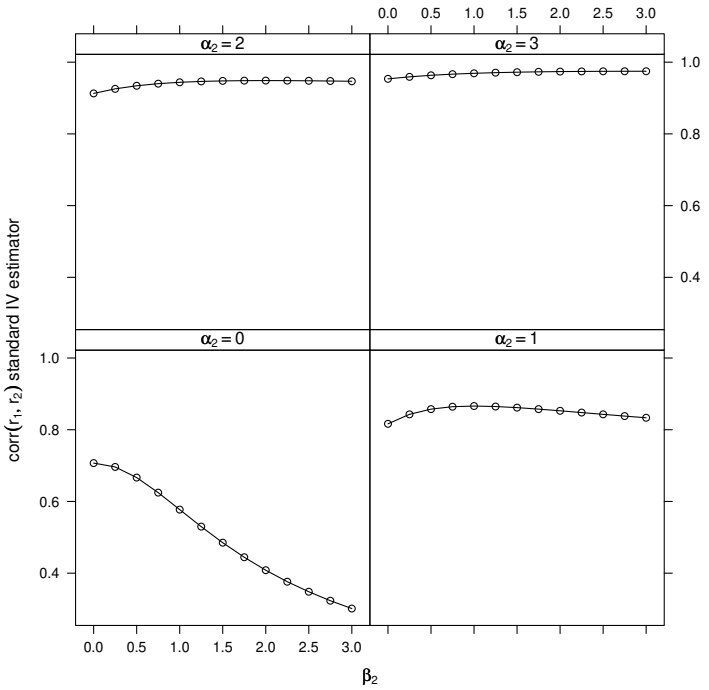



Figure C.10: The correlation between the first and second stage residuals for the standard IV estimator.

## C.4 Log link

Poisson regression uses a log link function in the terminology of generalised linear models and can be used to estimate a relative risks. The theory from Appendix B says in the second stage the estimate of $\beta_1$ should not be affected whilst the estimate of $\beta_0$ will be biased in proportion to the variance of the second stage error terms.

Figure C.11 shows the estimates of $\beta_1$ for the log link. The three estimates were equivalent in the first panel for $\alpha_2 = 0$. In the remaining panels as expected the direct estimate was positively biased by the confounder, however, surprisingly there was a difference between the standard and adjusted estimators.

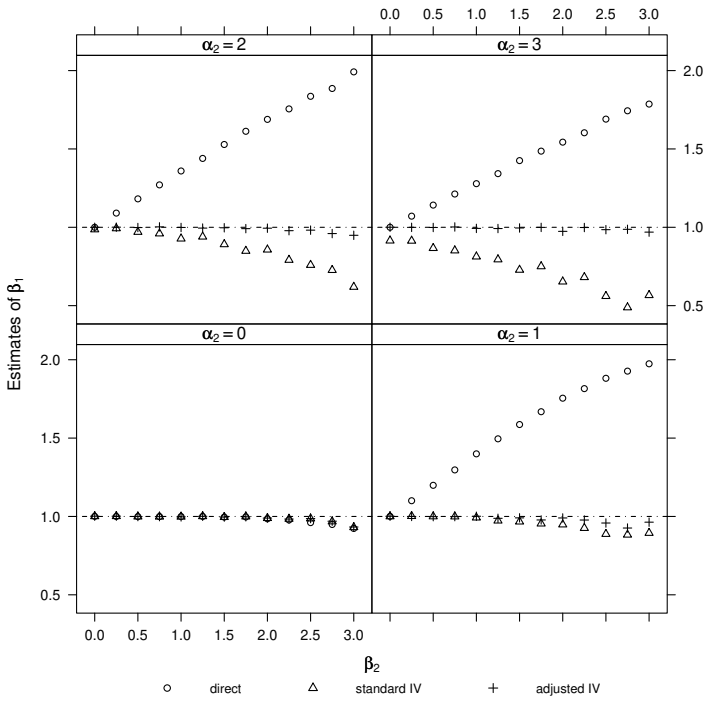



Figure C.11: The three estimators of $\beta_1$ under the log link.

Figure C.12 shows the estimates of the intercept in the linear predictor $\beta_0$, the baseline log disease risk, when the log link is used with the three estimators. The plots show that the baseline risk for the standard IV estimator is substantially increased, whereas the adjusted IV and direct estimates are much closer to the set value of $\log(0.05) = -2.995$.

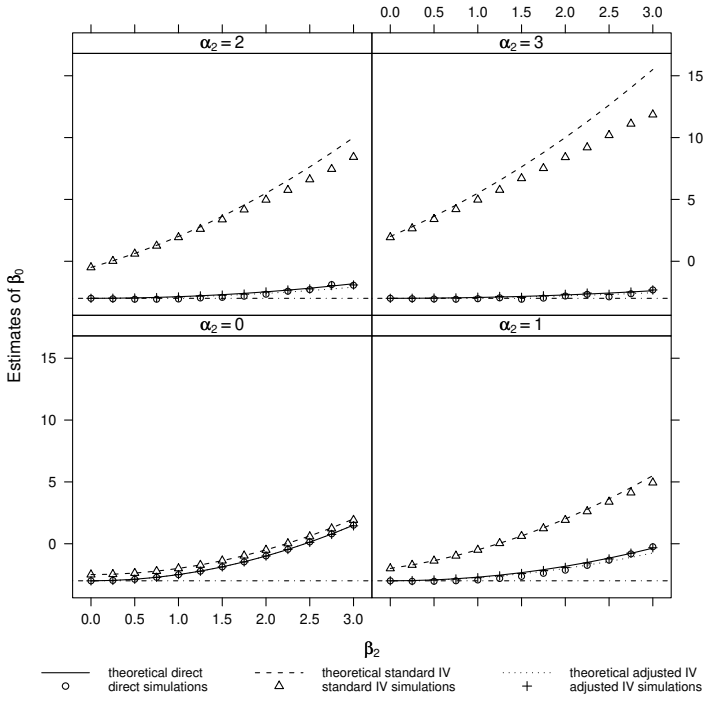



Figure C.12: Simulation and theoretical estimates of $\beta_0$ under the log link.

Figure C.13 shows simulation results using a rarer baseline probability of disease controlled through the $\beta_0$ parameter, $\beta_0 = \log(0.0005)$, and a larger cohort size of 30,000. This ensured that no more than 5% of subjects had a baseline probability of disease of more than 5%. The standard and adjusted parameters are now equivalent as expected.

In conclusion, the results of the simulations presented for the identity, log and probit link functions follow the relationship between marginal and conditional parameter estimates in generalised linear models with a normally distributed random intercept.

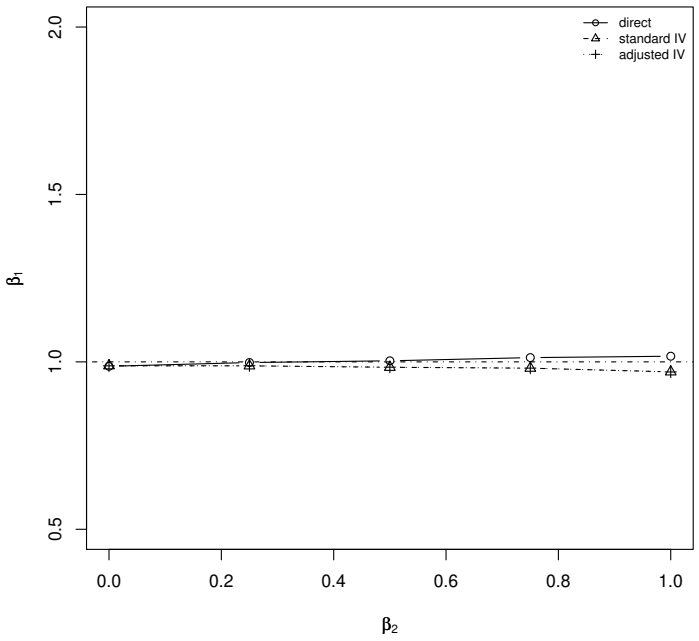



Figure C.13: The three estimators of $\beta_1$ under the log link when $\alpha_2 = 1$ with a smaller baseline risk of disease.

# Appendix D

# R and Stata code

## D.1 R and Stata programs for instrumental variable analysis

The following is a list of commands for performing instrumental variable analysis in Stata and R.

Stata programs include: `ivreg` or `ivregress` in version 10, `ivprobit`, `ivtobit` and `xtivreg`. User written Stata programs include: `ivreg2` (Baum *et al.*, 2003, 2007), `qvf` (Hardin & Carroll, 2003b), `cdsimeq` (Keshk, 2003), `xtivreg2` (Schaffer, 2005) and `cmp` (Roodman, 2008).

R packages include: the `sem` (Fox, 2008) package includes the `tsls` function, the `systemfit` package (Henningsen & Hamann, 2007) can perform varieties of two-stage least squares analyses and the `AER` package also contains the `ivreg` function for two-stage least squares regression.

## D.2 R code for the simulations in Chapter 4

```
setwd("/export/home/other/tmp8/logit")

# function for mr cohort study
mrsimstudy.logit <- function(q,obs,psd,alpha0,alpha1,beta0,beta1,alpha2=0,beta2=0,usd=1){
# usd = confounder standard deviation

# preliminaries
genotype <- seq(0,0,len=obs)
disease <- seq(0,0,len=obs)
rand.unif.g <- runif(n=obs)
rand.unif.d <- runif(n=obs)
confounder <- rnorm(n=obs,mean=0,sd=usd)

# genotype
genotype.1 <- rbinom(n=obs,size=1,prob=q)
genotype.2 <- rbinom(n=obs,size=1,prob=q)
genotype <- genotype.1 + genotype.2

# phenotype
phenotype.lp <- alpha0 + alpha1*genotype + alpha2*confounder
phenotype <- rnorm(n=obs,mean=phenotype.lp,sd=psd)

# disease status
p.disease <- plogis(beta0 + beta1*phenotype + beta2*confounder)
disease[p.disease >= rand.unif.d] <- 1

data <- data.frame(genotype=genotype, phenotype=phenotype,
            disease=disease, confounder=confounder)

# return results
return(list(genotype=genotype, phenotype=phenotype, disease=disease,
        confounder=confounder, p.disease=p.disease, data=data))
}

# function for the 3 estimators
mranalysis.logit.extra <- function(study){
# table: disease status by genotype
tab1 <- table(study$disease,study$genotype)
# table: proportion in each genotype by disease status
tab2 <- prop.table(tab1,1)

# direct model
direct <- glm(study$disease~study$phenotype, family=binomial)
direct.ci <- direct$coef + cbind(c(-1,-1),c(1,1))*qnorm(.975)*sqrt(diag(vcov(direct)))
direct.sum <- summary(direct)

# stage 1
stage1 <- lm(study$phenotype~study$genotype)
stage1.ci <- confint(stage1)
stage1.sum <- summary(stage1)

# naive model
naive <- glm(study$disease~stage1$fitted, family=binomial)
naive.ci <- naive$coef + cbind(c(-1,-1),c(1,1))*qnorm(.975)*sqrt(diag(vcov(naive)))
naive.sum <- summary(naive)

# adjusted model
adjusted <- glm(study$disease~stage1$fitted + stage1$residuals, family=binomial)
adjusted.ci <- adjusted$coef + cbind(c(-1,-1,-1),c(1,1,1))*qnorm(.975)*sqrt(diag(vcov(adjusted)))
adjusted.sum <- summary(adjusted)

# analysis to check the simulation
check <- glm(study$disease~study$phenotype + study$confounder, family=binomial)
check.ci <- check$coef + cbind(c(-1,-1,-1),c(1,1,1))*qnorm(.975)*sqrt(diag(vcov(check)))
```

```
check.sum <- summary(check)

# correlation between stage 1 residuals and stage 2 anscombe residuals
naive.corr <- cor(stage1$residuals, naive.sum$deviance.resid)
adjusted.corr <- cor(stage1$residuals, adjusted.sum$deviance.resid)

# return results
return(list(tab1=tab1,tab2=tab2,
        direct=direct, direct.ci=direct.ci, direct.sum=direct.sum,
        stage1=stage1, stage1.ci=stage1.ci, stage1.sum=stage1.sum,
        naive=naive, naive.ci=naive.ci, naive.sum=naive.sum,
        adjusted=adjusted, adjusted.ci=adjusted.ci, adjusted.sum=adjusted.sum,
        check=check, check.ci=check.ci, check.sum=check.sum,
        naive.corr=naive.corr, adjusted.corr=adjusted.corr))
}

# simulations
set.seed(1234)

q <- 0.3
obs <- 10000
alpha0 <- 0
alpha1 <- 1
alpha2 <- 0:3
p <- 0.05
beta0 <- qlogis(p)
beta1 <- 1
beta2 <- seq(0, 3, by=0.25)
psd <- 1
usd <- 1
its <- 10000
sig <- 0.05

resnames <- c("dir.b0","dir.b0.se","dir.b0.z","dir.b0.p",
            "dir.b1","dir.b1.se","dir.b1.z","dir.b1.p",
            "nai.b0","nai.b0.se","nai.b0.z","nai.b0.p",
            "nai.b1","nai.b1.se","nai.b1.z","nai.b1.p",
            "adj.b0","adj.b0.se","adj.b0.z","adj.b0.p",
            "adj.b1","adj.b1.se","adj.b1.z","adj.b1.p",
            "s1.r2","naive.corr","adjusted.corr",
            "dir.b1.cov","nai.b1.cov","adj.b1.cov",
            "dir.b1.pow","nai.b1.pow","adj.b1.pow",
            "dir.b0.sim.se","nai.b0.sim.se","adj.b0.sim.se",
            "dir.b1.sim.se","nai.b1.sim.se","adj.b1.sim.se",
            "s1.r2.sim.se","nai.corr.sim.se","adj.corr.sim.se",
            "dir.b1.cov.sim.se","nai.b1.cov.sim.se","adj.b1.cov.sim.se",
            "dir.b1.pow.sim.se","nai.b1.pow.sim.se","adj.b1.pow.sim.se",
            "b0","b1","b2",
            "a0","a1","a2",
            "psd","usd","q","p")

res <- matrix(nrow=length(alpha2)*length(beta2), ncol=length(resnames))
colnames(res) <- resnames

m <- 1
for(j in 1:length(alpha2)){
for(k in 1:length(beta2)){

print(Sys.time())

itresnames <- c("dir.b0","dir.b0.se","dir.b0.z","dir.b0.p",
            "dir.b1","dir.b1.se","dir.b1.z","dir.b1.p",
            "nai.b0","nai.b0.se","nai.b0.z","nai.b0.p",
            "nai.b1","nai.b1.se","nai.b1.z","nai.b1.p",
            "adj.b0","adj.b0.se","adj.b0.z","adj.b0.p",
            "adj.b1","adj.b1.se","adj.b1.z","adj.b1.p",
            "s1.r2","naive.corr","adjusted.corr",
```

```
            "dir.b1.cov","nai.b1.cov","adj.b1.cov",
            "dir.b1.pow","nai.b1.pow","adj.b1.pow")

itres <- matrix(nrow=its, ncol=length(itresnames))
colnames(itres) <- itresnames
print(m)

for(i in 1:its){
study <- mrsimstudy.logit(q=q, obs=obs,
    alpha0=alpha0, alpha1=alpha1, alpha2=alpha2[j],
    beta0=beta0, beta1=beta1, beta2=beta2[k],
    psd=1, usd=1)

analysis <- mranalysis.logit.extra(study)

dir.ci <- analysis$direct.sum$coefficients[2,1] + c(-1,1)*qnorm(.975)*analysis$direct.sum$coefficients[2,2]
nai.ci <- analysis$naive.sum$coefficients[2,1] + c(-1,1)*qnorm(.975)*analysis$naive.sum$coefficients[2,2]
adj.ci <- analysis$adjusted.sum$coefficients[2,1] + c(-1,1)*qnorm(.975)*analysis$adjusted.sum$coefficients[2,2]

itres[i,1:4] <- analysis$direct.sum$coefficients[1,]
itres[i,5:8] <- analysis$direct.sum$coefficients[2,]
itres[i,9:12] <- analysis$naive.sum$coefficients[1,]
itres[i,13:16] <- analysis$naive.sum$coefficients[2,]
itres[i,17:20] <- analysis$adjusted.sum$coefficients[1,]
itres[i,21:24] <- analysis$adjusted.sum$coefficients[2,]
itres[i,25:27] <- c(analysis$stage1.sum$r.squared, analysis$naive.corr, analysis$adjusted.corr)
itres[i,"dir.b1.cov"] <- as.numeric(dir.ci[1] <= beta1 & dir.ci[2] >= beta1)
itres[i,"nai.b1.cov"] <- as.numeric(nai.ci[1] <= beta1 & nai.ci[2] >= beta1)
itres[i,"adj.b1.cov"] <- as.numeric(adj.ci[1] <= beta1 & adj.ci[2] >= beta1)
itres[i,"dir.b1.pow"] <- as.numeric(analysis$direct.sum$coefficients[2,4] < sig)
itres[i,"nai.b1.pow"] <- as.numeric(analysis$naive.sum$coefficients[2,4] < sig)
itres[i,"adj.b1.pow"] <- as.numeric(analysis$adjusted.sum$coefficients[2,4] < sig)

study <- analysis <- NULL

if(i%%100 == 0){cat(".\n")}else{cat(".")}
}

res[m,1:length(itresnames)] <- colMeans(itres, na.rm=TRUE)
res[m,"dir.b0.sim.se"] <- sqrt(var(itres[,"dir.b0"])/its)
res[m,"nai.b0.sim.se"] <- sqrt(var(itres[,"nai.b0"])/its)
res[m,"adj.b0.sim.se"] <- sqrt(var(itres[,"adj.b0"])/its)
res[m,"dir.b1.sim.se"] <- sqrt(var(itres[,"dir.b1"])/its)
res[m,"nai.b1.sim.se"] <- sqrt(var(itres[,"nai.b1"])/its)
res[m,"adj.b1.sim.se"] <- sqrt(var(itres[,"adj.b1"])/its)
res[m,"s1.r2.sim.se"] <- sqrt(var(itres[,"s1.r2"])/its)
res[m,"nai.corr.sim.se"] <- sqrt(var(itres[,"naive.corr"])/its)
res[m,"adj.corr.sim.se"] <- sqrt(var(itres[,"adjusted.corr"])/its)
res[m,"dir.b1.cov.sim.se"] <- sqrt(res[m,"dir.b1.cov"]*(1 - res[m,"dir.b1.cov"])/its)
res[m,"nai.b1.cov.sim.se"] <- sqrt(res[m,"nai.b1.cov"]*(1 - res[m,"nai.b1.cov"])/its)
res[m,"adj.b1.cov.sim.se"] <- sqrt(res[m,"adj.b1.cov"]*(1 - res[m,"adj.b1.cov"])/its)
res[m,"dir.b1.pow.sim.se"] <- sqrt(res[m,"dir.b1.pow"]*(1 - res[m,"dir.b1.pow"])/its)
res[m,"nai.b1.pow.sim.se"] <- sqrt(res[m,"nai.b1.pow"]*(1 - res[m,"nai.b1.pow"])/its)
res[m,"adj.b1.pow.sim.se"] <- sqrt(res[m,"adj.b1.pow"]*(1 - res[m,"adj.b1.pow"])/its)
res[m,"b0"] <- beta0
res[m,"b1"] <- beta1
res[m,"b2"] <- beta2[k]
res[m,"a0"] <- alpha0
res[m,"a1"] <- alpha1
res[m,"a2"] <- alpha2[j]
res[m,"psd"] <- psd
res[m,"usd"] <- usd
res[m,"q"] <- q
res[m,"p"] <- p

save(res, file="/export/data/hs/tmp8/logit/mr.logit.extra.Rdata")
m <- m + 1
```

```
cat("\n")
}
}
```

## D.3 Code for Chapter 5

### D.3.1 R code for the maximum likelihood estimation of the MVMR model

The following code is for studies with complete data.

```
library(Matrix) # for kronecker()
library(foreign)

setwd("Z:/Mendelian.Randomization/Example/R/data")

#
# data edit
#

data <- read.dta("mann.dta")
data <- data[-19,]
data[,"year"] <- c(1999,2000,1996,1998,2000,1998,1996,1999,2000,
    2000,1998,1999,1998,1999,1996,2000,1996,1996)
zerotest <- data[1:13,4:15]==0
data[1:13,4:15][zerotest] <- 0.5

attach(data)

# log odds ratios
logor1 <- log((Ss.case*SS.control)/(Ss.control*SS.case))
vlogor1 <- 1/Ss.case + 1/SS.control + 1/Ss.control + 1/SS.case
plogor1 <- 1/vlogor1
logor2 <- log((ss.case*SS.control)/(ss.control*SS.case))
vlogor2 <- 1/ss.case + 1/SS.case + 1/ss.control + 1/SS.control
plogor2 <- 1/vlogor2
covlogor12 <- 1/SS.control + 1/SS.case
logor1low <- logor1 - qnorm(.975)*sqrt(vlogor1)
logor1upp <- logor1 + qnorm(.975)*sqrt(vlogor1)
logor2low <- logor2 - qnorm(.975)*sqrt(vlogor2)
logor2upp <- logor2 + qnorm(.975)*sqrt(vlogor2)
selogor1 <- sqrt(vlogor1)
selogor2 <- sqrt(vlogor2)

# differences in means
d1 <- Ss.m - SS.m
vd1 <- Ss.se^2 + SS.se^2
pd1 <- 1/vd1
d2 <- ss.m - SS.m
vd2 <- ss.se^2 + SS.se^2
pd2 <- 1/vd2
covd12 <- SS.se^2
d1low <- d1 - qnorm(.975)*sqrt(vd1)
d1upp <- d1 + qnorm(.975)*sqrt(vd1)
d2low <- d2 - qnorm(.975)*sqrt(vd2)
d2upp <- d2 + qnorm(.975)*sqrt(vd2)
sed1 <- sqrt(vd1)
sed2 <- sqrt(vd2)

# scale the phenotype differences
scd1 <- d1/0.05
scd2 <- d2/0.05
scvd1 <- vd1/0.05^2
scpd1 <- 1/scvd1
scvd2 <- vd2/0.05^2
scpd2 <- 1/scvd2
sccovd12 <- covd12/0.05^2
scd1low <- d1low/0.05
```

```
scd1upp <- d1upp/0.05
scd2low <- d2low/0.05
scd2upp <- d2upp/0.05
scsed1 <- sed1/0.05
scsed2 <- sed2/0.05

type <- c(rep(1,10), rep(2,3), rep(3,5))

data <- data.frame(data, logor1, vlogor1, plogor1,
        logor2, vlogor2, plogor2, covlogor12,
        logor1low, logor1upp, logor2low, logor2upp,
        selogor1, selogor2,
        d1, vd1, pd1,
        d2, vd2, pd2, covd12,
        d1low, d1upp, d2low, d2upp,
        sed1, sed2,
        scd1, scvd1, scpd1,
        scd2, scvd2, scpd2, sccovd12,
        scd1low, scd1upp, scd2low, scd2upp,
        scsed1, scsed2,
        type)
save(data, file="mann.Rdata")

rm(list=ls())

load("mann.Rdata")

cdata <- data[data$type==1,]

#
# common eta - complete data
#

reml.log.like <- function(beta, logor1, logor2, vlogor1, vlogor2,
    covlogor12, d1, d2, vd1, vd2, covd12){

eta <- beta[1]
mu1 <- beta[2]
mu2 <- beta[3]
logtau1sq <- beta[4]
logtau2sq <- beta[5]
transrho <- beta[6]

tau1 <- sqrt(exp(logtau1sq))
tau2 <- sqrt(exp(logtau2sq))
rho <- (exp(transrho) - 1)/(exp(transrho) + 1)

loglike <- NULL
BETA <- c(eta*mu1, mu1, eta*mu2, mu2)
W <- B <- SIGMA <- matrix(nrow=4,ncol=4)

for (i in 1:length(logor1)) {
    Y <- c(logor1[i], d1[i], logor2[i], d2[i])

    W[1,] <- c(vlogor1[i], 0, covlogor12[i], 0)
    W[2,] <- c(0, vd1[i], 0, covd12[i])
    W[3,] <- c(covlogor12[i], 0, vlogor2[i], 0)
    W[4,] <- c(0, covd12[i], 0, vd2[i])

    Bleft <- matrix(c(tau1^2, rho*tau1*tau2, rho*tau1*tau2, tau2^2), nrow=2)
    Bright <- matrix(c(eta^2, eta, eta, 1), nrow=2)
    B <- kronecker(Bleft, Bright)

    SIGMA <- W + B
    detSIGMA <- det(SIGMA)
    if(detSIGMA == 0){break} # check of invertibility
    invSIGMA <- solve(SIGMA)
```

```
    loglike[i] <- -0.5*(log(detSIGMA) + t(Y - BETA)%*%invSIGMA%*%(Y - BETA))
}
-1*sum(loglike)
}

inits <- rep(0,6)
remlopt <- optim(inits, reml.log.like, hessian=T, control=list(maxit=5000),
        logor1=cdata$logor1, logor2=cdata$logor2,
        vlogor1=cdata$vlogor1, vlogor2=cdata$vlogor2, covlogor12=cdata$covlogor12,
        d1=cdata$scd1, d2=cdata$scd2, vd1=cdata$scvd1, vd2=cdata$scvd2, covd12=cdata$sccovd12)

cat("mr model - complete outcome data \n")
remlopt$convergence
remlopt$hessian
det(remlopt$hessian)
INFO <- solve(remlopt$hessian)
SE <- diag(chol(INFO))
CI <- remlopt$par + cbind(rep(-1,6), rep(1,6))*qnorm(.975)*SE

cat("logORpd: ", remlopt$par[1], "\n")
cat("logORpd CI: ", CI[1,], "\n")
cat("ORpd: ", exp(remlopt$par[1]), "\n")
cat("ORpd CI: ", exp(CI[1,]), "\n")

cat("delta2: ", remlopt$par[2], "\n")
cat("delta2 CI: ", CI[2,], "\n")

cat("delta3: ", remlopt$par[3], "\n")
cat("delta3 CI: ", CI[3,], "\n")

t1sq <- exp(remlopt$par[4])
cat("tau1sq: ", t1sq, "\n")
cat("tau1sq CI:", exp(CI[4,]), "\n")

t2sq <- exp(remlopt$par[5])
cat("tau2sq: ", t2sq, "\n")
cat("tau2sq CI: ", exp(CI[5,]), "\n")

rho <- (exp(remlopt$par[6]) - 1)/(exp(remlopt$par[6]) + 1)
cat("rho: ", rho, "\n")
rhoCI <- c((exp(CI[6,1]) - 1)/(exp(CI[6,1]) + 1), (exp(CI[6,2]) - 1)/(exp(CI[6,2]) + 1))
cat("rho CI: ", rhoCI, "\n")

cat("lambda: ", remlopt$par[2]/remlopt$par[3], "\n")
lambdaCI <- c(CI[2,1]/CI[3,1], CI[2,2]/CI[3,2])
cat("lambda CI: ", lambdaCI, "\n")

B1left <- matrix(c(t1sq, rho*sqrt(t1sq)*sqrt(t2sq), rho*sqrt(t1sq)*sqrt(t2sq), t2sq), nrow=2)
B1right <- matrix(c(remlopt$par[1]^2, remlopt$par[1], remlopt$par[1], 1), nrow=2)
B1 <- kronecker(B1left, B1right)
B1left; B1right; B1
det(B1left); det(B1right); det(B1)

loglike1 <- -remlopt$value # log likelihood
cat("log-likelihood: ", loglike1, "\n")
cat("\n")
```

The following code is for studies with complete data or missing either gene-disease or gene-phenotype data.

```
#
# MVMR model - all data
#

reml.log.like.all <- function(beta, logor1, logor2, vlogor1, vlogor2,
    covlogor12, d1, d2, vd1, vd2, covd12, type){
```

```
eta <- beta[1]
mu1 <- beta[2]
mu2 <- beta[3]
logtau1sq <- beta[4]
logtau2sq <- beta[5]
transrho <- beta[6]

tau1 <- sqrt(exp(logtau1sq))
tau2 <- sqrt(exp(logtau2sq))
rho <- (exp(transrho) - 1)/(exp(transrho) + 1)

loglike <- NULL

n <- length(logor1)

for (i in 1:n) {
    if(type[i] == 1){ # complete outcomes
    BETA <- c(eta*mu1, mu1, eta*mu2, mu2)
    W <- B <- SIGMA <- matrix(nrow=4, ncol=4)
    Y <- c(logor1[i], d1[i], logor2[i], d2[i])

    W[1,] <- c(vlogor1[i], 0, covlogor12[i], 0)
    W[2,] <- c(0, vd1[i], 0, covd12[i])
    W[3,] <- c(covlogor12[i], 0, vlogor2[i], 0)
    W[4,] <- c(0, covd12[i], 0, vd2[i])

    Bleft <- matrix(c(tau1^2, rho*tau1*tau2, rho*tau1*tau2, tau2^2), nrow=2)
    Bright <- matrix(c(eta^2, eta, eta, 1), nrow=2)
    B <- kronecker(Bleft, Bright)
    }

    if(type[i] == 2){ # gene-disease outcomes only
    BETA <- c(eta*mu1, eta*mu2)
    W <- B <- SIGMA <- matrix(nrow=2, ncol=2)
    Y <- c(logor1[i], logor2[i])

    W[1,] <- c(vlogor1[i], covlogor12[i])
    W[2,] <- c(covlogor12[i], vlogor2[i])

    B[1,] <- c(eta^2*tau1^2, eta^2*rho*tau1*tau2)
    B[2,] <- c(eta^2*rho*tau1*tau2, eta^2*tau2^2)
    }

    if(type[i] == 3){ # gene-phenotype outcomes only
    BETA <- c(mu1, mu2)
    W <- B <- SIGMA <- matrix(nrow=2, ncol=2)
    Y <- c(d1[i], d2[i])

    W[1,] <- c(vd1[i], covd12[i])
    W[2,] <- c(covd12[i], vd2[i])

    B[1,] <- c(tau1^2, rho*tau1*tau2)
    B[2,] <- c(rho*tau1*tau2, tau2^2)
    }

    SIGMA <- W + B
    detSIGMA <- det(SIGMA)
    if(detSIGMA == 0){break} # check of invertibility
    invSIGMA <- solve(SIGMA)
    loglike[i] <- -0.5*(log(detSIGMA) + t(Y - BETA)%*%invSIGMA%*%(Y - BETA))
}
-1*sum(loglike)
}

inits <- rep(0.1, 6)
fit <- optim(inits, reml.log.like.all, hessian=T, control=list(maxit=5000),
```

```
        logor1=data$logor1, logor2=data$logor2,
        vlogor1=data$vlogor1, vlogor2=data$vlogor2, covlogor12=data$covlogor12,
        d1=data$scd1, d2=data$scd2,
        vd1=data$scvd1, vd2=data$scvd2,
        covd12=data$sccovd12, type=data$type)

cat("\n mr model - all data \n")
fit$convergence
fit$hessian
det(fit$hessian)
INFO3 <- solve(fit$hessian)
SE3 <- diag(chol(INFO3))
CI3 <- fit$par + cbind(rep(-1,6), rep(1,6))*qnorm(.975)*SE3

cat("logORpd: ", fit$par[1], "\n")
cat("logORpd CI: ", CI3[1,], "\n")
cat("ORpd: ", exp(fit$par[1]), "\n")
cat("ORpd CI: ", exp(CI3[1,]), "\n")

cat("delta2: ", fit$par[2], "\n")
cat("delta2 CI: ", CI3[2,], "\n")

cat("delta3: ", fit$par[3], "\n")
cat("delta3 CI: ", CI3[3,], "\n")

t1sq <- exp(fit$par[4])
cat("tau1sq: ", t1sq, "\n")
cat("tau1sq CI:", exp(CI3[4,]), "\n")

t2sq <- exp(fit$par[5])
cat("tau2sq: ", t2sq, "\n")
cat("tau2sq CI: ", exp(CI3[5,]), "\n")

rho <- (exp(fit$par[6]) - 1)/(exp(fit$par[6]) + 1)
cat("rho: ", rho, "\n")
rhoCI <- c((exp(CI3[6,1]) - 1)/(exp(CI3[6,1]) + 1), (exp(CI3[6,2]) - 1)/(exp(CI3[6,2]) + 1))
cat("rho CI: ", rhoCI, "\n")

cat("lambda: ", fit$par[2]/fit$par[3], "\n")
lambdaCI3 <- c(CI3[2,1]/CI3[3,1], CI3[2,2]/CI3[3,2])
cat("lambda CI: ", lambdaCI3, "\n")

B3left <- matrix(c(t1sq, rho*sqrt(t1sq)*sqrt(t2sq), rho*sqrt(t1sq)*sqrt(t2sq), t2sq), nrow=2)
B3right <- matrix(c(fit$par[1]^2, fit$par[1], fit$par[1], 1), nrow=2)
B3 <- kronecker(B3left, B3right)
B3left; B3right; B3
det(B3left); det(B3right); det(B3)

loglike3 <- -fit$value # log likelihood
cat("log-likelihood: ", loglike3, "\n")
cat("\n")
```

### D.3.2 R code for the maximum likelihood estimation of the MVMR-GMF model

The following code is for studies with complete data.

```
#
# model with gmf - complete data
#

gmf.reml.log.like <- function(beta, logor1, logor2, vlogor1, vlogor2,
    covlogor12, d1, d2, vd1, vd2, covd12){
```

```
eta <- beta[1]
trlambda <- beta[2]
delta <- beta[3]
logtausq <- beta[4]

lambda <- (exp(trlambda) - 1)/(exp(trlambda) + 1)
tausq <- exp(logtausq)

loglike <- NULL
BETA <- c(eta*lambda*delta, lambda*delta, eta*delta, delta)
W <- B <- SIGMA <- matrix(nrow=4,ncol=4)

for (i in 1:length(logor1)) {
    Y <- c(logor1[i], d1[i], logor2[i], d2[i])

    W[1,] <- c(vlogor1[i], 0, covlogor12[i], 0)
    W[2,] <- c(0, vd1[i], 0, covd12[i])
    W[3,] <- c(covlogor12[i], 0, vlogor2[i], 0)
    W[4,] <- c(0, covd12[i], 0, vd2[i])

    Bleft <- matrix(c(lambda^2*tausq, lambda*tausq, lambda*tausq, tausq), nrow=2)
    Bright <- matrix(c(eta^2, eta, eta, 1), nrow=2)
    B <- kronecker(Bleft, Bright)

    SIGMA <- W + B
    detSIGMA <- det(SIGMA)
    if(detSIGMA == 0){break} # check of invertibility
    invSIGMA <- solve(SIGMA)

    loglike[i] <- -0.5*(log(detSIGMA) + t(Y - BETA)%*%invSIGMA%*%(Y - BETA))
}
-1*sum(loglike)
}

inits <- rep(0,4)
gmfremlopt <- optim(inits, gmf.reml.log.like, hessian=T, control=list(maxit=5000),
        logor1=cdata$logor1, logor2=cdata$logor2, vlogor1=cdata$vlogor1,
        vlogor2=cdata$vlogor2, covlogor12=cdata$covlogor12,
        d1=cdata$scd1, d2=cdata$scd2, vd1=cdata$scvd1, vd2=cdata$scvd2, covd12=cdata$sccovd12)

cat("\n gmf model - complete outcome data \n")
gmfremlopt$convergence
gmfremlopt$hessian
det(gmfremlopt$hessian)

INFO2 <- solve(gmfremlopt$hessian)
SE2 <- diag(chol(INFO2))
CI2 <- gmfremlopt$par + cbind(rep(-1,4), rep(1,4))*qnorm(0.975)*SE2

cat("logORpd: ", gmfremlopt$par[1], "\n")
cat("logORpd: ", CI2[1,], "\n")
cat("ORpd: ", exp(gmfremlopt$par[1]), "\n")
cat("ORpd CI: ", exp(CI2[1,]), "\n")

lambda <- (exp(gmfremlopt$par[2]) - 1)/(exp(gmfremlopt$par[2]) + 1)
cat("lambda: ", lambda, "\n")
lambdaCI2 <- c((exp(CI2[2,1]) - 1)/(exp(CI2[2,1]) + 1), (exp(CI2[2,2]) - 1)/(exp(CI2[2,2]) + 1))
cat("lambda CI: ", lambdaCI2, "\n")

cat("delta: ", gmfremlopt$par[3], "\n")
cat("delta CI: ", CI2[3,], "\n")

tsq <- exp(gmfremlopt$par[4])
cat("tausq: ", tsq, "\n")
cat("tausq CI: ", exp(CI2[4,]), "\n")

B2left <- matrix(c(lambda^2*tsq, lambda*tsq, lambda*tsq, tsq), nrow=2)
```

```
B2right <- matrix(c(remlopt$par[1]^2, remlopt$par[1], remlopt$par[1], 1), nrow=2)
B2 <- kronecker(B2left, B2right)
B2left; B2right; B2
det(B2left); det(B2right); det(B2)

loglike2 <- -gmfremlopt$value # log likelihood
cat("log-likelihood: ", loglike2, "\n")

# REML likelihood ratio test
remllrtp <- pchisq(-2*(loglike2 - loglike1), 2, lower.tail=F)
cat("REML LRT: ", remllrtp, "\n")
```

The following code is for studies with complete data or missing either gene-disease or gene-phenotype outcomes.

```
#
# MVMR-GMF model - complete and incomplete data
#

gmf.reml.log.like.all <- function(beta, logor1, logor2, vlogor1, vlogor2,
    covlogor12, d1, d2, vd1, vd2, covd12, type){

eta <- beta[1]
trlambda <- beta[2]
delta <- beta[3]
logtausq <- beta[4]

lambda <- (exp(trlambda) - 1)/(exp(trlambda) + 1)
tausq <- exp(logtausq)
loglike <- NULL

for (i in 1:length(logor1)) {
    if(type[i] == 1){
    BETA <- c(eta*lambda*delta, lambda*delta, eta*delta, delta)
    W <- B <- SIGMA <- matrix(nrow=4, ncol=4)
    Y <- c(logor1[i], d1[i], logor2[i], d2[i])

    W[1,] <- c(vlogor1[i], 0, covlogor12[i], 0)
    W[2,] <- c(0, vd1[i], 0, covd12[i])
    W[3,] <- c(covlogor12[i], 0, vlogor2[i], 0)
    W[4,] <- c(0, covd12[i], 0, vd2[i])

    Bleft <- matrix(c(lambda^2*tausq, lambda*tausq, lambda*tausq, tausq), nrow=2)
    Bright <- matrix(c(eta^2, eta, eta, 1), nrow=2)
    B <- kronecker(Bleft, Bright)
    }
    if(type[i] == 2){ # g-d only
    BETA <- c(eta*lambda*delta, eta*delta)
    W <- B <- SIGMA <- matrix(nrow=2, ncol=2)
    Y <- c(logor1[i], logor2[i])

    W[1,] <- c(vlogor1[i], covlogor12[i])
    W[2,] <- c(covlogor12[i], vlogor2[i])

    B[1,] <- c(eta^2*lambda^2*tausq, eta^2*lambda*tausq)
    B[2,] <- c(eta^2*lambda*tausq, eta^2*tausq)
    }
    if(type[i] == 3){ # g-p only
    BETA <- c(lambda*delta, delta)
    W <- B <- SIGMA <- matrix(nrow=2, ncol=2)
    Y <- c(d1[i], d2[i])

    W[1,] <- c(vd1[i], covd12[i])
    W[2,] <- c(covd12[i], vd2[i])

    B[1,] <- c(lambda^2*tausq, lambda*tausq)
    B[2,] <- c(lambda*tausq, tausq)
```

```
    }

    SIGMA <- W + B
    detSIGMA <- det(SIGMA)
    if(detSIGMA == 0){break} # check of invertibility
    invSIGMA <- solve(SIGMA)
    loglike[i] <- -0.5*(log(detSIGMA) + t(Y - BETA)%*%invSIGMA%*%(Y - BETA))
}
-1*sum(loglike)
}

inits <- rep(0,4)
gmf <- optim(inits, gmf.reml.log.like.all, hessian=T, control=list(maxit=5000),
        logor1=data$logor1, logor2=data$logor2,
        vlogor1=data$vlogor1, vlogor2=data$vlogor2, covlogor12=data$covlogor12,
        d1=data$scd1, d2=data$scd2,
        vd1=data$scvd1, vd2=data$scvd2, covd12=data$sccovd12,
        type=data$type)

cat("\n gmf model - complete and incomplete outcome data \n")
gmf$convergence
gmf$hessian
det(gmf$hessian)

INFO4 <- solve(gmf$hessian)
SE4 <- diag(chol(INFO4))
CI4 <- gmf$par + cbind(rep(-1,4), rep(1,4))*qnorm(0.975)*SE4

cat("logORpd: ", gmf$par[1], "\n")
cat("logORpd: ", CI4[1,], "\n")
cat("ORpd: ", exp(gmf$par[1]), "\n")
cat("ORpd CI: ", exp(CI4[1,]), "\n")

lambda <- (exp(gmf$par[2]) - 1)/(exp(gmf$par[2]) + 1)
cat("lambda: ", lambda, "\n")
lambdaCI4 <- c((exp(CI4[2,1]) - 1)/(exp(CI4[2,1]) + 1), (exp(CI4[2,2]) - 1)/(exp(CI4[2,2]) + 1))
cat("lambda CI: ", lambdaCI4, "\n")

cat("delta: ", gmf$par[3], "\n")
cat("delta CI: ", CI4[3,], "\n")

tsq <- exp(gmf$par[4])
cat("tausq: ", tsq, "\n")
cat("tausq CI: ", exp(CI4[4,]), "\n")

B4left <- matrix(c(lambda^2*tsq, lambda*tsq, lambda*tsq, tsq), nrow=2)
B4right <- matrix(c(gmf$par[1]^2, gmf$par[1], gmf$par[1], 1), nrow=2)
B4 <- kronecker(B4left, B4right)
B4left; B4right; B4
det(B4left); det(B4right); det(B4)

loglike4 <- -gmf$value # log likelihood
cat("log-likelihood: ", loglike4, "\n")

# likelihood ratio test
lr.test.reml.p <- pchisq(-2*(loglike4 - loglike3), 2, lower.tail=F)
cat("LRT: ", lr.test.reml.p, "\n")
```

### D.3.3 WinBUGS model statement for the product normal formulation

```
model{
    for(i in 1:10){
        theta1[i] <- eta*lambda*delta[i]
        logor1[i] ~ dnorm(theta1[i], plogor1[i])
```

```
        delta1[i] <- lambda*delta[i]
        scd1[i] ~ dnorm(delta1[i], scpd1[i])

        theta2[i] <- eta*delta[i]
        logor2[i] ~ dnorm(theta2[i], plogor2[i])

        delta2[i] <- delta[i]
        scd2[i] ~ dnorm(delta2[i], scpd2[i])

        delta[i] ~ dnorm(delta3, invtausq)
    }
    for(i in 11:13){
        theta1[i] <- eta*lambda*delta[i]
        logor1[i] ~ dnorm(theta1[i], plogor1[i])

        theta2[i] <- eta*delta[i]
        logor2[i] ~ dnorm(theta2[i], plogor2[i])

        delta[i] ~ dnorm(delta3, invtausq)
    }
    for(i in 14:18){
        delta1[i] <- lambda*delta[i]
        scd1[i] ~ dnorm(delta1[i], scpd1[i])

        delta2[i] <- delta[i]
        scd2[i] ~ dnorm(delta2[i], scpd2[i])

        delta[i] ~ dnorm(delta3, invtausq)
    }

    # priors
    delta3 ~ dnorm(0, 1.0E-6)
    invtausq ~ dgamma(0.001, 0.001)
    tausq <- 1/invtausq
    eta ~ dnorm(0, 1.0E-6)
    lambda ~ dbeta(1, 1)
    or <- exp(eta)
}
```

### D.3.4 Stata code for a multi-column forest plot

To draw a four column forest plot in Stata the plot region needs to split into five columns, one for the study labels and the other four for each forest plot for The use of the `graph twoway` forced size option `fxsize(#)` (although the number in the brackets has to be guessed) is useful in setting the aspect ratio for each of the plots which are then combined into a single plot using `graph combine` at the end. Note the `fxsize()` option is documented at the bottom of the `graph combine` help-file.

```
* plot for the y-axis
twoway rcap or1low or1upp studynumericorder, horizontal lc(none) || ///
    , legend(off) ylab(1(1)19, noticks value ang(h) labsize(small)) ///
    ytitle("") ///
    xtitle(" ", justification(left)) ///
    xscale(off fill) xlab(.25, labsize(vsmall)) ///
```

```
    yscale(rev noline) ///
    fxsize(15) ///
    subtitle(" ") ///
    plotr(ls(none))
graph save ./Plots/yaxis.gph,replace

* plot for the gg vs Gg gene-disease outcome
twoway rspike or1low or1upp studynumericorder, horizontal lc(gs0) || ///
    scatter studynumericorder or1 if studynumericorder!=19 [aw=or1prec], msize(vsmall) mc(gs0) m(s) || ///
    scatter studynumericorder or1 if studynumericorder==19, xline(1) msize(small) mc(gs0) m(d) || ///
    , legend(off) ylab(1(1)19,noticks nolabels) ytitle("") ///
    yscale(rev noline) plotr(ls(none)) ///
    xscale(log range(.25 30)) xlab(.25 1 4 16, labsize(small)) ///
    xtitle(OR, size(small)) ///
    subtitle("Fracture risk: Gg vs gg") ///
    fxsize(30)
graph save ./Plots/gdma1.gph,replace

* plot for the gg vs GG gene-disease outcome
twoway rspike or2lowres or2uppres studynumericorder, horizontal lc(gs0) || ///
    scatter studynumericorder or2 if studynumericorder!=19 [aw=or2prec], msize(vsmall) mc(gs0) m(s) || ///
    scatter studynumericorder or2 if studynumericorder==19, xline(1) msize(small) mc(gs0) m(d) || ///
    , ylab(1(1)19, nolabels noticks) legend(off) ///
    fxsize(30) ytitle("") yscale(rev noline) ///
    xscale(log range(.25 30)) xlab(.25 1 4 16, labsize(small)) ///
    xtitle(OR,size(small)) ///
    subtitle("Fracture risk: GG vs gg") plotr(ls(none))
graph save ./Plots/gdma2.gph,replace

* plot for the gg vs Gg gene-phenotype outcome
twoway rspike d1low d1upp studynumericorder, horizontal lc(gs0) || ///
    scatter studynumericorder d1 if studynumericorder!=19 [aw=d1prec], msize(vsmall) mc(gs0) m(s) || ///
    scatter studynumericorder d1 if studynumericorder==19, xline(0) msize(small) mc(gs0) m(d) || ///
    , legend(off) ytitle("") ///
    yscale(rev noline) ylab(1(1)19, nolabels noticks) ///
    plotr(ls(none)) ///
    xtitle(BMD, size(small)) ///
    xscale(range(-.2 .2)) xlab(, labsize(small)) ///
    subtitle("BMD: Gg vs gg") ///
    fxsize(30)
graph save ./Plots/gpma1.gph,replace

* plot for the GG vs gg gene-phenotype outcome
twoway rspike d2lowres d2uppres studynumericorder, horizontal lc(gs0) || ///
    scatter studynumericorder d2 if studynumericorder!=19 [aw=d2prec], msize(vsmall) mc(gs0) m(s) || ///
    scatter studynumericorder d2 if studynumericorder==19, xline(0) msize(small) mc(gs0) m(d) || ///
    , legend(off) ytitle("") ///
    yscale(rev noline) ylab(1(1)19, noticks nolabels) ///
    plotr(ls(none)) ///
    xtitle(BMD,size(small)) ///
    xlab(,labsize(small)) ///
    xscale(range(-.2 .2)) ///
    subtitle("BMD: GG vs gg") ///
    fxsize(30)
graph save ./Plots/gpma2.gph,replace

* combine the 5 plots
graph combine "./Plots/yaxis.gph" "./Plots/gdma1.gph" ///
    "./Plots/gpma1.gph" "./Plots/gdma2.gph" "./Plots/gpma2.gph" ///
    , cols(5) imargin(vsmall)
```

### D.3.5 R code for a multi-column forest plot

In R the plotting region can be divided into the 5 required columns by setting `par(mfrow=c(1,5))` before the plots are drawn.

```
library(plotrix) # for plotCI() function

load("mann.Rdata")
id <- 1:dim(data)[1]
sdata <- data[c(15,18,17,14,16,12,13,11,7,10,1,8,2,4,5,3,6,9),] # order studies

# plot
par(mfrow=c(1,5), mar=c(5,0,3,0.25), lend="square", font.main=1)
plot(id , yaxt="n", xaxt="n", xlab="", ylab="", bty="n", col=NA)
axis(side=2, labels=sdata$study, at=id, tick=F, las=1, line=-10)

plotCI(x=sdata$logor1, y=id,
    li=sdata$logor1low, ui=sdata$logor1upp, err="x",
    xlim=c(-2,3.5), bty="n", yaxt="n",
    xlab="G-D log odds ratio",
    ylab="",
    main="Gg versus gg",
    cex=1.1, pch=15)
abline(v=0)

plotCI(x=sdata$scd1, y=id,
    li=sdata$scd1low, ui=sdata$scd1upp, err="x",
    xlim=c(-4,4), bty="n", yaxt="n",
    xlab="G-P mean difference",
    ylab="",
    main="Gg versus gg",
    cex=1.1, pch=15)
abline(v=0)

plotCI(x=sdata$logor2, y=id,
    li=sdata$logor2low, ui=sdata$logor2upp, err="x",
    xlim=c(-2,3.5), bty="n", yaxt="n",
    xlab="G-D log odds ratio",
    ylab="",
    main="GG versus gg",
    cex=1.1, pch=15)
abline(v=0)

plotCI(x=sdata$scd2, y=id,
    li=sdata$scd2low, ui=sdata$scd2upp, err="x",
    xlim=c(-4,4), bty="n", yaxt="n",
    xlab="G-P mean difference",
    ylab="",
    main="GG versus gg",
    cex=1.1, pch=15)
abline(v=0)
```

## D.4 R code for the simulations in Chapter 6

### D.4.1 Single cohort simulations

```
# function for taylor series and fieller's method  - ignoring correlation terms

tys <- function(my,vy,mx,vx){
    n <- max(length(my), length(vy), length(mx), length(vx))
    m <- (my/mx) + (vx*my)/(mx^3)
    v <- (vx*my^2)/(mx^4) + vy/(mx^2)
    ci <- m + cbind(rep(-1,n),rep(1,n))*qnorm(0.975)*sqrt(v)
    a <- 1 - qnorm(0.975)^2*(vx/mx^2)
    b <- (vy/my^2) + (vx/mx^2) - qnorm(0.975)*(vy/my^2)*(vx/mx^2)
    fci <- ((my/mx)/a)*(1 + cbind(rep(-1,n),rep(1,n))*qnorm(0.975)*sqrt(b))
    return(list(m=m,v=v,ci=ci,fci=fci))
}

# cohort study function

chstudy <- function(N,alpha0,alpha1,alpha2,psd,beta0,beta1,beta2,usd,q){
NSIM <- N
genotype1 <- rbinom(n=NSIM, size=1, prob=q)
genotype2 <- rbinom(n=NSIM, size=1, prob=q)
genotype <- genotype1 + genotype2
confounder <- runif(NSIM, 0, usd)
phenotype <- rnorm(NSIM, alpha0 + alpha1*genotype + alpha2*confounder, psd)
lp <- beta0 + beta1*phenotype + beta2*confounder
pd <- exp(lp)/(1 + exp(lp))
d <- as.numeric(runif(NSIM) < pd)
ch <- data.frame(genotype=genotype, phenotype=phenotype, lp=lp, pd=pd, d=d, confounder=confounder)

tab <- table(ch$d, ch$genotype)
a <- tab[1,1]
b <- tab[1,2]
c <- tab[1,3]
d <- tab[2,1]
e <- tab[2,2]
f <- tab[2,3]
OR2 <- (e/b)/(d/a)
logOR2 <- log(OR2)
se2 <- sqrt(sum(1/tab[1:2,1:2]))
OR3 <- (f/c)/(d/a)
logOR3 <- log(OR3)
se3 <- sqrt(sum(1/tab[,-2]))
m0 <- mean(ch$phenotype[ch$genotype == 0 & ch$d == 0])
m1 <- mean(ch$phenotype[ch$genotype == 1 & ch$d == 0])
m2 <- mean(ch$phenotype[ch$genotype == 2 & ch$d == 0])
sd0 <- sd(ch$phenotype[ch$genotype == 0 & ch$d == 0])
sd1 <- sd(ch$phenotype[ch$genotype == 1 & ch$d == 0])
sd2 <- sd(ch$phenotype[ch$genotype == 2 & ch$d == 0])
d2 <- m1 - m0
d3 <- m2 - m0
sed2 <- sqrt(sd0^2 + sd1^2)
sed3 <- sqrt(sd0^2 + sd2^2)
cov23 <- sum(1/tab[,1])
covd23 <- sd0^2

# calculations for difference between marginal and cond OR
pd0 <- ch$pd[ch$genotype == 0]
pd1 <- ch$pd[ch$genotype == 1]
pd2 <- ch$pd[ch$genotype == 2]
mpd0 <- mean(pd0)
mpd1 <- mean(pd1)
mpd2 <- mean(pd2)
ml0 <- mean(log(pd0/(1 - pd0)))
```

```
ml1 <- mean(log(pd1/(1 - pd1)))
ml2 <- mean(log(pd2/(1 - pd2)))
lm0 <- log(mpd0/(1 - mpd0))
lm1 <- log(mpd1/(1 - mpd1))
lm2 <- log(mpd2/(1 - mpd2))
lg2c <- ml1 - ml0
lg3c <- 0.5*(ml2 - ml0)
lg2m <- lm1 - lm0
lg3m <- 0.5*(lm2 - lm0)

lr <- coef(glm(ch$d ~ ch$phenotype, family="binomial"))
st <- lm(ch$phenotype ~ ch$genotype)
ad <- coef(glm(ch$d ~ ch$phenotype + st$residuals, family="binomial"))

eta2 <- logOR2/d2
eta3 <- logOR3/d3

err2 <- (eta2 - beta1)^2
err3 <- (eta3 - beta1)^2

out <- c(m0,m1,m2,d2,d3,
    sd0,sd1,sd2,sed2,sed3,
    logOR2,se2,
    logOR3,se3,
    eta2,eta3,
    lg2c,lg3c,lg2m,lg3m,
    a,b,c,d,e,f,
    lr,ad,cov23,covd23,err2,err3)
names(out) <- c("mu0","mu1","mu2","d2","d3",
        "sd0","sd1","sd2","sed2","sed3",
        "theta2","setheta2",
        "theta3","setheta3",
        "eta2","eta3",
        "lg2c","lg3c","lg2m","lg3m",
        "a","b","c","d","e","f",
        "b0","b1","ab0","ab1","ab2",
        "cov23","covd23","err2","err3")
return(out)
}



# 1 - cohort size 3000
set.seed(12345)
its <- 10000
sims <- replicate(its,
    chstudy(N=3000, alpha0=0, alpha1=1, alpha2=0, psd=1,
        beta0=log(0.05/0.95), beta1=log(1.25), beta2=0, usd=1, q=0.3),
    simplify=F)
res <- matrix(unlist(sims), nrow=length(sims), byrow=T)
colnames(res) <- names(sims[[1]])
mns <- colMeans(res, na.rm=TRUE)
print(mns, 4)
sds <- apply(res, 2, sd, na.rm=TRUE)
n <- length(mns)
cis <- mns + cbind(rep(-1,n),rep(1,n))*qnorm(.975)*sds/sqrt(its)
rownames(cis) <- names(mns)
print(cis, 4)
mns[c("eta2","eta3")] - log(1.25)
cis[c("eta2","eta3"),] - log(1.25)
save(res,mns,sds,cis,file="ch3000_rev.Rdata")

tys(mns["theta2"], sds["theta2"]^2/its, mns["d2"], sds["d2"]^2/its)
tys(mns["theta3"], sds["theta3"]^2/its, mns["d3"], sds["d3"]^2/its)

# 2 - cohort size 3e3 with confounder effect
set.seed(12345)
```

```
its <- 10000
sims <- replicate(its,
    chstudy(N=3000, alpha0=0, alpha1=1, alpha2=1, psd=1,
        beta0=log(0.05/0.95), beta1=log(1.25), beta2=1, usd=1, q=0.3),
    simplify=F)
res <- matrix(unlist(sims), nrow=length(sims), byrow=T)
colnames(res) <- names(sims[[1]])
mns <- colMeans(res)
print(mns,4)
sds <- apply(res, 2, sd)
n <- length(mns)
cis <- mns + cbind(rep(-1,n),rep(1,n))*qnorm(.975)*sds/sqrt(its)
rownames(cis) <- names(mns)
print(cis,4)
mns[c("eta2","eta3")] - log(1.25)
cis[c("eta2","eta3"),] - log(1.25)
save(res,mns,sds,cis,file="ch3000conf_rev.Rdata")

tys(mns["theta2"], sds["theta2"]^2/its, mns["d2"], sds["d2"]^2/its)
tys(mns["theta3"], sds["theta3"]^2/its, mns["d3"], sds["d3"]^2/its)

# 3 - cohort size 3000 - no confounder - smaller psd
set.seed(12345)
its <- 10000
sims <- replicate(its,
    chstudy(N=3000, alpha0=0, alpha1=1, alpha2=0, psd=0.1,
        beta0=log(0.05/0.95), beta1=log(1.25), beta2=0, usd=1, q=0.3),
    simplify=F)
res <- matrix(unlist(sims), nrow=length(sims), byrow=T)
colnames(res) <- names(sims[[1]])
mns <- colMeans(res)
print(mns,4)
sds <- apply(res, 2, sd)
n <- length(mns)
cis <- mns + cbind(rep(-1,n),rep(1,n))*qnorm(.975)*sds/sqrt(its)
rownames(cis) <- names(mns)
print(cis,4)
mns[c("eta2","eta3")] - log(1.25)
cis[c("eta2","eta3"),] - log(1.25)
save(res,mns,sds,cis,file="ch3000sm_rev.Rdata")

tys(mns["theta2"], sds["theta2"]^2/its, mns["d2"], sds["d2"]^2/its)
tys(mns["theta3"], sds["theta3"]^2/its, mns["d3"], sds["d3"]^2/its)


# 4 - cohort size 3e3 with confounder effect - smaller psd
set.seed(12345)
its <- 10000
sims <- replicate(its,
    chstudy(N=3000, alpha0=0, alpha1=1, alpha2=1, psd=0.1,
        beta0=log(0.05/0.95), beta1=log(1.25), beta2=1, usd=1, q=0.3),
    simplify=F)
res <- matrix(unlist(sims), nrow=length(sims), byrow=T)
colnames(res) <- names(sims[[1]])
mns <- colMeans(res)
print(mns,4)
sds <- apply(res, 2, sd)
n <- length(mns)
cis <- mns + cbind(rep(-1,n),rep(1,n))*qnorm(.975)*sds/sqrt(its)
rownames(cis) <- names(mns)
print(cis,4)
mns[c("eta2","eta3")] - log(1.25)
cis[c("eta2","eta3"),] - log(1.25)
save(res,mns,sds,cis,file="ch3000confsm_rev.Rdata")

tys(mns["theta2"], sds["theta2"]^2/its, mns["d2"], sds["d2"]^2/its)
tys(mns["theta3"], sds["theta3"]^2/its, mns["d3"], sds["d3"]^2/its)
```

### D.4.2 Meta-analysis simulations

```
library(Matrix)

# loglikelihood for mvmr model
mvmrLoglike <- function(beta,logor1,logor2,vlogor1,vlogor2,covlogor12,d1,d2,vd1,vd2,covd12){
eta <- beta[1]
mu1 <- beta[2]
mu2 <- beta[3]
logtau1sq <- beta[4]
logtau2sq <- beta[5]
transrho <- beta[6]
tau1 <- sqrt(exp(logtau1sq))
tau2 <- sqrt(exp(logtau2sq))
rho <- (exp(transrho) - 1)/(exp(transrho) + 1)
loglike <- NULL
BETA <- c(eta*mu1, mu1, eta*mu2, mu2)
W <- B <- SIGMA <- matrix(nrow=4,ncol=4)
for (i in 1:length(logor1)) {
    Y <- c(logor1[i], d1[i], logor2[i], d2[i])
    W[1,] <- c(vlogor1[i], 0, covlogor12[i], 0)
    W[2,] <- c(0, vd1[i], 0, covd12[i])
    W[3,] <- c(covlogor12[i], 0, vlogor2[i], 0)
    W[4,] <- c(0, covd12[i], 0, vd2[i])
    Bleft <- matrix(c(tau1^2, rho*tau1*tau2, rho*tau1*tau2, tau2^2), nrow=2)
    Bright <- matrix(c(eta^2, eta, eta, 1), nrow=2)
    B <- kronecker(Bleft, Bright)
    SIGMA <- W + B
    invSIGMA <- try(solve(SIGMA))
    x <- 1
    if(class(invSIGMA) == "try-error"){
        next
    }
    loglike[i] <- -0.5*(log(det(SIGMA)) + t(Y - BETA)%*%invSIGMA%*%(Y - BETA))
}
-1*sum(loglike)
}

tsrc <- function(my,vy,mx,vx,rhoxy,b1){
    m <- my/mx + (vx*my)/(mx^3) - (rhoxy*sqrt(vx)*sqrt(vy))/(mx^2)
    v <- (vx*my^2)/(mx^4) + vy/(mx^2) - (2*rhoxy*sqrt(vx)*sqrt(vy)*my)/(mx^3)
    ci <- m + c(-1,1)*qnorm(0.975)*sqrt(v)
    bias <- m - b1
    return(list(m=m,v=v,ci=ci,bias=bias))
}

# scenario 1

set.seed(12345)
its <- 100
sims <- matrix(nrow=its, ncol=13)
colnames(sims) <- c("eta","mu2","mu3","logtau2sq","logtau3sq","transrho",
            "conv","bias","mse","rc2","rc2bias","rc3","rc3bias")
for(i in 1:its){
ma <- replicate(10, chstudy(3e3,0,1,0,1,log(0.05/0.95),log(1.25),0,1,0.3), simplify=FALSE)
meta <- matrix(unlist(ma), nrow=length(ma), byrow=T)
colnames(meta) <- names(ma[[1]])
sum(is.infinite(meta[,"theta3"]))
meta <- meta[is.finite(meta[,"theta3"]),]
inits <- rep(0,6)
fit <- try(optim(inits, fn=mvmrLoglike, hessian=T, method="BFGS",
    logor1=meta[,"theta2"], logor2=meta[,"theta3"],
    vlogor1=meta[,"setheta2"]^2, vlogor2=meta[,"setheta3"]^2,
    covlogor12=meta[,"cov23"],
    d1=meta[,"d2"], d2=meta[,"d3"],
    vd1=meta[,"sed2"]^2, vd2=meta[,"sed3"]^2, covd12=meta[,"covd23"]))
bias <- fit$par[1] - log(1.25)
```

```
mse <- (fit$par[1] - log(1.25))^2
rc2 <- tsrc(mean(meta[,"theta2"]), var(meta[,"theta2"]),
        mean(meta[,"d2"]), var(meta[,"d2"]),
        cor(meta[,"theta2"],meta[,"d2"]), log(1.25))
rc3 <- tsrc(mean(meta[,"theta3"]), var(meta[,"theta3"]),
        mean(meta[,"d3"]), var(meta[,"d3"]),
        cor(meta[,"theta3"],meta[,"d3"]), log(1.25))
sims[i,] <- c(fit$par,fit$convergence,bias,mse,rc2$m,rc2$bias,rc3$m,rc3$bias)
print(i)
}
sims <- sims[sims[,"conv"] == 0,]
avs <- colMeans(sims, na.rm=TRUE)
avs
sds <- apply(sims, 2, sd)
n <- length(avs)
cis <- avs + cbind(rep(-1,n),rep(1,n))*qnorm(.975)*sds/sqrt(its)
cis
```

# Bibliography


Amemiya, T. 1974. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, **2**, 105–110.

Anderson, T. W. 1958. *An Introduction to Multivariate Statistical Analysis*. Chichester: Wiley.

Andrews, D. W. K., Moreira, M. J., & Stock, J. H. 2007. Performance of conditional Wald tests in IV regression with weak instruments. *Journal of Econometrics*, **139**(1), 116–132.

Ardlie, K. G., Lunetta, K. L., & Seielstad, M. 2002. Testing for Population Subdivision and Association in Four Case-Control Studies. *The American Journal of Human Genetics*, **71**(2), 304–311.

Balding, D. J., Bishop, M., & Cannings, C. (eds). 2007. *Handbook of Statistical Genetics*. Third edn. Vol. 1. Chichester, UK: Wiley.

Balke, A., & Pearl, J. 1994. Counterfactual probabilities: Computational methods, bounds, and applications. *Uncertainty in Artificial Intelligence*, **10**, 46–54.

Baltagi, B. H. 1998. *Econometrics*. New York: Springer.

Bammann, K., & Wawro, N. 2006. Die Einbeziehung genetischer Faktoren in Studien der Epidemiologie. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, **49**(10), 974–981.

Basmann, R. L. 1957. A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation. *Econometrica*, **25**(1), 77–83.

Baum, C. F. 2006. *An Introduction to Modern Econometrics Using Stata*. College Station, Texas: Stata Press.

Baum, C. F., Schaffer, M. E., & Stillman, S. 2003. Instrumental variables and GMM: Estimation and testing. *The Stata Journal*, **3**(1), 1–31.

Baum, C. F., Schaffer, M. E., & Stillman, S. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *The Stata Journal*, **7**(4), 465–506.

Bautista, L. E., Smeeth, L., Hingorani, A. D., & Casas, J. P. 2006. Estimation of bias in nongenetic observational studies using Mendelian Triangulation. *Annals of Epidemiology*, **16**(9), 675–680.

Binder, D. A. 1983. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**(3), 279–292.

Bochud, M. 2008. On the use of Mendelian randomization to infer causality in observational epidemiology. *European Heart Journal*, **29**, 2456–2457.

Bochud, M., Chiolero, A., Elston, R. C., & Paccaud, F. 2008. A cautionary note on the use of Mendelian randomization to infer causation in observational epidemiology. *International Journal of Epidemiology*, **37**, 414–417. in press, Letter to the Editor.

Bound, J., Jaeger, D. A., & Baker, R. M. 1995. Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak. *Journal of the American Statistical Association*, **90**(430).

Bowden, R. J., & Turkington, D. 1981. Comparative Study of Instrumental Variables Estimators for Nonlinear Simultaneous Models. *Journal of the American Statistical Association*, **76**(376), 988–995.

Bowden, R. J., & Turkington, D. A. 1984. *Instrumental variables*. Cambridge: Cambridge University Press.

Bradburn, M. J., Deeks, J. J., Berlin, J. A., & Localio, A. R. 2007. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, **26**(1), 53–77.

Bradford Hill, A. 1965. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, **58**, 295–300.

Brennan, P. 2004. Commentary: Mendelian randomization and gene-environment interaction. *International Journal of Epidemiology*, **33**(1), 17–21.

Brennan, P., Hsu, C. C., Moullan, N., Szeszenia-Dabrowska, N., Lissowska, J., Zaridze, D., Rudnai, Peter, Fabianova, E., Mates, D., Bencko, V., Foretova, L., Janout, V., Gemignani, F., Chabrier, A., Hall, J., Hung, R. J., Boffetta, P., & Canzian, F. 2005. Effect of cruciferous vegetables on lung cancer in patients stratified by genetic status: a mendelian randomisation approach. *Lancet*, **366**, 1558–1560.

Breslow, N. E., & Clayton, D. G. 1993. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **88**(421), 9–25.

Briggs, A, & Fenn, P. 1998. Confidence intervals of surfaces? Uncertainty on the cost-effectiveness plane. *Health Economics*, **7**, 723–740.

Brunner, E. J., Kivimäki, M., Witte, D. R., Lawlor, D. A., Davey Smith, G., Cooper, J. A., Miller, M., Lowe, G. D. O., Rumley, A., Casas, J. P., Shah, T., Humphries, S. E., Hingorani, A. D., Marmot, M. G., Timpson, N. J., & Kumari, M. 2008. Inflammation, Insulin Resistance, and Diabetes - Mendelian Randomization Using CRP Haplotypes Points Upstream. *Plos Medicine*, **5**(8), e155.

Bubela, T. 2006. Science communication in transition: genomics hype, public engagement, education and commercialization pressures. *Clinical Genetics*, **70**(5), 445–450.

Burnett, A. K., Goldstone, A. H., & Wheatley, K. 2005 (January). *MRC Working Parties on Leukaemia in Adults and Children Actute Myloid Leukaemia Trial 15: Protocol for patients under 60 (Trial reference ISCTN 17161961).* Tech. rept. University of Birmingham. Version 3.

Cameron, A. C., & Trivedi, P. K. 2005. *Microeconometrics: methods and applications.* New York: Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. 2009. *Microeconometrics Using Stata.* College Station, Texas: Stata Press.

Cardon, L. R., & Bell, J. I. 2001. Association study designs for complex diseases. *Nature Reviews Genetics*, **2**, 91–99.

Cardon, L. R., & Palmer, L. J. 2003. Population stratification and spurious allelic association. *The Lancet*, **361**, 598–604.

Carroll, R. J., & Stefanski, L. A. 1994. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine*, **13**, 1265–1282.

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T., & Abbott, R. D. 1984. On errors-in-variables for binary regression models. *Biometrika*, **71**(1), 19–25.

Carroll, R. J., Ruppert, D., & Stefanski, L. A. 1995. *Measurement Error in Nonlinear Models.* London: Chapman & Hall.

Carroll, R. J., Rupert, D., Stefansky, L. A., & Crainiceanu, C. M. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective.* Second edn. Boca Raton, FL, USA: CRC Press.

Casas, J. P., Bautista, L. E., S., L., Sharma, P., & Hingorani, A. D. 2005. Homocysteine and stroke: evidence on a causal link from mendelian randomisation. *The Lancet*, **365**, 224–232.

Casas, J. P., Shah, T., Cooper, J., Hawe, E., McMahon, A. D., Gaffney, D., Packard, C. J., O'Reilly, D. S., Juhan-Vague, I., Yudkin, J. S., Tremoli, E., Margaglione, M., Di Minno, G., Hamsten, A., Kooistra, T., Stephens, J. W., Hurel, S. J., Livingstone, S., Colhoun, H. M., Miller, G. J., Bautista, L. E., Meade, T., Sattar, N., Humphries, S. E., & Hingorani, A. D. 2006. Insight into the nature of the CRP-coronary event association using Mendelian randomization. *International Journal of Epidemiology*, **35**(4), 922–931.

Chao, W. H., Palta, M., & Young, T. 1997. Effect of Omitted Confounders on the Analysis of Correlated Binary Data. *Biometrics*, **53**(2), 678–689.

Chaudhary, M. A., & Stearns, S. C. 1996. Estimating Confidence Intervals For Cost-Effectiveness Ratios: An Example from a Randomized Trial. *Statistics in Medicine*, **15**, 1447–1458.

Chen, L., Davey Smith, G., Harbord, R. M., & Lewis, S. J. 2008. Alcohol Intake and Blood Pressure: A Systematic Review Implementing a Mendelian Randomization Approach. *Plos Medicine*, **5**(3), 461.

Chesher, A. 2007 (July). *Endogeneity and discrete outcomes*. CeMMAP working papers CWP05/07. Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Clayton, D. 2007. *Handbook of Statistical Genetics*. Third edn. Vol. 2. Chichester, UK: Wiley. Chap. Population Association, pages 1216–1237.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D., & Todd, J. A. 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, **37**, 1243–1246.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. 1999. *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag.

Cox, D. R., & Wermuth, N. 2001. Some statistical aspects of causality. *European Socio-*

*logical Review*, **17**(1), 65–74.

CRP CHD Genetics Collaboration, . 2008. Collaborative pooled analysis of data on C-reactive protein gene variants and coronary disease: judging causality by Mendelian randomisation. *European Journal of Epidemiology*, **23**, 531–540.

Curnow, R. N. 2005. Genetic biases in using 'Mendelian randomization' to compare transplantation with chemotherapy. *International Journal of Epidemiology*, **34**(5), 1167–1168.

Czeizel, A. E., & Dudàs, I. 1992. Prevention of the first occurrence of neural tube defects by periconceptual vitamin supplementation. *New England Journal of Medicine*, **327**, 1832–35.

Davey Smith, G. 2006. Randomised by (your) god: robust inference from an observational study design. *Journal of Epidemiology and Community Health*, **60**(5), 382–388.

Davey Smith, G. 2007. Capitalizing on Mendelian randomization to assess the effects of treatments. *Journal of the Royal Society of Medicine*, **100**, 432–435.

Davey Smith, G., & Ebrahim, S. 2002. Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers. *BMJ*, **325**(7378), 1437–1438.

Davey Smith, G., & Ebrahim, S. 2003. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease. *International Journal of Epidemiology*, **32**, 1–22.

Davey Smith, G., & Ebrahim, S. 2004. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, **33**(1), 30–42.

Davey Smith, G., & Ebrahim, S. 2005. What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures? *British Medical Journal*, **330**(7499), 1076–1079.

Davey Smith, G., & Ebrahim, S. 2008. Reply: A cautionary note on the use of Mendelian

randomization to infer causation in observational epidemiology by Bochud et al. *International Journal of Epidemiology*, **37**, 416–417.

Davey Smith, G., Harbord, R., & Ebrahim, S. 2004. Fibrinogen, C-reactive protein and coronary heart disease: does Mendelian randomization suggest the associations are non-causal? *QJM*, **97**, 163–166.

Davey Smith, G., Lawlor, D. A., Harbord, R., Timpson, N., Rumley, A., Lowe, G. D. O., Day, I. N. M., & Ebrahim, S. 2005a. Association of C-Reactive Protein with Blood Pressure and Hypertension: Life Course Confounding and Mendelian Randomization Tests of Causality. *Arteriosclerosis, Thrombosis and Vascular Biology*, **25**, 1051–1056.

Davey Smith, G., Harbord, R., Milton, J., Ebrahim, S., & Sterne, J. A. C. 2005b. Does Elevated Plasma Fibrinogen Increase the Risk of Coronary Heart Disease? Evidence from a Meta-Analysis of Genetic Association Studies. *Arteriosclerosis, Thrombosis and Vascular Biology*, **25**, 2228–2233.

Davidson, R., & MacKinnon, J. G. 2004. *Econometric Theory and Methods*. New York: Oxford University Press.

Dawid, A. P. 2002. Influence diagrams for causal modelling and inference. *International statistical review*, **70**(2), 161–189.

Day, I. N. M., Gu, D., Ganderton, R. H., Spanakis, E., & Ye, S. 2001. Epidemiology and the genetic basis of disease. *International Journal of Epidemiology*, **30**(4), 661–667.

Dehghan, Abbas, Kardys, Isabella, de Maat, Moniek P. M., Uitterlinden, Andre G., Sijbrands, Eric J. G., Bootsma, Aart H., Stijnen, Theo, Hofman, Albert, Schram, Miranda T., & Witteman, Jacqueline C. M. 2007. Genetic Variation, C-Reactive Protein Levels, and Incidence of Diabetes. *Diabetes*, **56**(3), 872–878.

DerSimonian, R., & Laird, N. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**(3), 177–188.

Didelez, V., & Sheehan, N. 2007a. *Causality and Probability in the Sciences*. London: College Publications. Chap. Mendelian randomisation: why epidemiology needs a formal language for causality, pages 263–292.

Didelez, V., & Sheehan, N. 2007b. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, **16**, 309–330.

Diggle, P. J., Heagerty, P., Liang, K-Y., & Zeger, S. L. 2002. *Analysis of Longitudinal Data*. Second edn. Oxford University Press.

Doll, R., & Hill, A. B. 1952. A study of the aetiology of carcinoma of the lung. *British Medical Journal*, **2**(4797), 1271–1286.

Ducimetière, P., & Cambien, F. 2007. Coronary heart disease aetiology: associations and causality. *Comptes Rendus Biologies*, **330**(4), 299–305.

Dunn, G., & Bentall, R. 2007. Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine*, **26**, 4719–4745.

Dunn, G., Maracy, M., & Tomenson, B. 2005. Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. *Statistical Methods in Medical Research*, **14**(4), 369.

Duval, S., & Tweedie, R. L. 2000a. A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Society*, **95**, 89–98.

Duval, S., & Tweedie, R. L. 2000b. Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**, 455–463.

Ebrahim, S., & Clarke, M. 2007. STROBE: new standards for reporting obervational epidemiology, a chance to improve. *International Journal of Epidemiology*, **36**(5), 946–948.

Ebrahim, S., & Davey Smith, G. 2007. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics*. in press.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. 1997. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, **315**(7109), 629–634.

Egger, M., Ebrahim, S., & Davey Smith, G. 2002. Where now for meta-analysis. *International Journal of Epdiemiology*, **31**, 1–5.

Egger, M., Juni, P., Bartlett, C., Holenstein, F., & Sterne, J. 2003. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment*, **7**(1), 1–76.

Elston, R., Olson, J., & Palmer, L. 2002. *Biostatistical Genetics and Genetic Epidemiology*. New York: Wiley.

Elwood, M. 2007. *Critical appraisal of epidemiological studies and clinical trials*. third edn. Oxford: Oxford University Press.

Farewell, V. T. 1979. Some results on the estimation of logistic models based on retrospective data. *Biometrika*, **66**(1), 27–32.

Fieller, E. C. 1954. Some Problems in Interval Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, **16**(2), 175–185.

Fleiss, J. L. 1993. The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, **2**, 121–145.

Foster, E. M. 1997. Instrumental variables for logistic regression: an illustration. *Social Science Research*, **26**, 487–504.

Fox, J. 1979. Simultaneous Equation Models and Two-Stage Least Squares. *Sociological Methodology*, **10**, 130–150.

Fox, J. 2008. *sem: Structural Equation Models*. R package version 0.9-12.

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A-M., Ness, A. R., Ebrahim, S., Lawlor, D. A., Ring, S. M., Ben-Shlomo, Y., Jarvelin, M-R., Sovio, U., Bennett, A. J., Melzer, D., Ferrucci, L., Loos, R. J. F., Barroso, I., Wareham, N. J., Owen, F. Karpeand K. R., Cardon, L. R., Walker, M., Hitman, G. A., Palmer, C. N. A., Doney, A. S. F., Morris, A. D., Davey Smith, G., Consortium, The Wellcome Trust Case Control, Hattersley, A. T., & McCarthy, M. I. 2007a. A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science*, **316**(5826), 889–894.

Frayling, T. M., Rafiq, S., Murray, A., Hurst, A. J., Weedon, M. N., Henley, W., Bandinelli, S., Corsi, A. M., Ferrucci, L., Guralnik, J. M., Wallace, R. B., & Melzer, D. 2007b. An Interleukin-18 Polymorphism Is Associated With Reduced Serum Concentrations and Better Physical Functioning in Older People. *Journals of Gerontology Series A: Biological and Medical Sciences*, **62**(1), 73–78.

Freathy, R. M., Timpson, N. J., Lawlor, D. A., Pouta, A., Ben-Shlomo, Y., Ruokonen, A., Ebrahim, S., Shields, B., Zeggini, E., Weedon, M. N., Lindgren, C. M., Lango, H., Melzer, D., Ferrucci, L., Paolisso, G., Neville, M. J., Karpe, F., Palmer, C. N. A., Morris, A. D., Elliott, P., Jarvelin, M-R., Davey Smith, G., McCarthy, M. I., Hattersley, A. T., & Frayling, T. M. 2008. Common Variation in the FTO Gene Alters Diabetes-Related Metabolic Traits to the Extent Expected Given Its Effect on BMI. *Diabetes*, **57**(5), 1419–1426.

Gail, M. H. 1988. The Effect of Pooling Across Strata in Perfectly Balanced Studies. *Biometrics*, **44**(1), 151–162.

Geman, S., & Geman, D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Gilmour, A. R., Anderson, R. D., & Rae, A. L. 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika*, **72**, 593–599.

Grant, S. F. A., Reid, D. M., Blake, G., Herd, R., Fogelman, I., & Ralston, S. H. 1996. Reduced bone density and osteoporosis associated with a polymorphic Sp 1 binding site in the collagen type I $\alpha$ 1 gene. *Nature Genetics*, **14**(2), 203–205.

Grassi, M., Assanelli, D., & Pezzini, A. 2007. Direct, reverse or reciprocal causation in the relation between homocysteine and ischemic heart disease. *Thrombosis Research*, **120**(1), 61–69.

Gray, R., & Wheatley, K. 1991. How to avoid bias when comparing cone marrow transplantation with chemotherapy. *Bone Marrow Transplantation*, **7**, 9–12. Suppl. 3.

Greene, W. H. 1999. *Econometric Analysis*. Fourth edn. New York: Prentice Hall.

Greene, W. H. 2008. *Econometric Analysis*. Sixth edn. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Greenland, S. 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, **29**, 722–729.

Greenland, S., & Brumback, B. 2002. An overview of relations among causal modelling methods. *International Journal of Epidemiology*, **31**(5), 1030–1037.

Greenland, S., & Morgenstern, H. 1989. Ecological Bias, Confounding, and Effect Modification. *International Journal of Epidemiology*, **18**(1), 269–274.

Greenland, S., Morgenstern, H., Poole, C., & Robins, J. M. 1989. RE: Confounding Confounding. *American Journal of Epidemiology*, **129**(5), 1086–1091.

Greenland, S., Pearl, J., & Robins, J. M. 1999a. Causal Diagrams for Epidemiologic Research. *Epidemiology*, **10**(1), 37–48.

Greenland, S., Robins, J. M., & Pearl, J. 1999b. Confounding and Collapsibility in Causal

Inference. *Statistical Science*, **14**, 29–46.

Grootendorst, P. 2007. A review of instrumental variables estimation of treatment effects in the applied health sciences. *Health Services and Outcomes Research Methodology*, **7**, 159–179.

Hammer Bech, B., Autrup, H., Nohr, E. A., Henriksen, T. B., & Olsen, J. 2006. Stillbirth and slow metabolizers of caffeine: comparison by genotypes. *International Journal of Epidemiology*, **35**(4), 948–953.

Hansen, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica*, **50**(4), 1029–1054.

Harbord, R. M., Egger, M., & Sterne, J. A. C. 2006. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine*, **25**(20), 3443–3457.

Hardin, J. W. 2002. The robust variance estimator for two-stage models. *The Stata Journal*, **2**(3), 253–265.

Hardin, J. W., & Carroll, R. J. 2003a. Measurement error, GLMs, and notational conventions. *The Stata Journal*, **3**(4), 329–341.

Hardin, J. W., & Carroll, R. J. 2003b. Variance estimation for the instrumental variables approach to measurement error in generalized linear models. *The Stata Journal*, **3**(4), 342–350.

Hardin, J. W., & Hilbe, J. M. 2003. *Generalized Estimating Equations*. Chapman and Hall/CRC.

Hardy, G. H. 1908. Mendelian proportions in a mixed population. *Science*, **28**, 49–50.

Hauck Jr, W. W., & Donner, A. 1977. Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*, **72**(360), 851–853.

Hausman, J. A. 1978. Specification Tests in Econometrics. *Econometrica*, **46**(6), 1251–1271.

Hayya, J., Armstrong, D., & Gressis, N. 1975. A Note on the Ratio of Two Normally Distributed Variables. *Management Science*, **21**(11), 1338–1341.

Heckman, J. J., & Hotz, V. J. 1989. Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association*, **84**(408), 862–874.

Henningsen, A., & Hamann, J. D. 2007. systemfit: A Package for Estimating Systems of Simultaneous Equations in R. *Journal of Statistical Software*, **23**(4), 1–40.

Herder, C., Klopp, N., Baumert, J., Müller, M., Khuseyinova, N., Meisinger, C., Martin, S., Illig, T., Koenig, W., & Thorand, B. 2008. Effect of macrophage migration inhibitory factor (MIF) gene variants and MIF serum concentrations on the risk of type 2 diabetes: results from the MONICA/KORA Augsburg Case–Cohort Study, 1984–2002. *Diabetologia*, **51**(2), 276–284.

Hernán, M. A., & Robins, J. M. 2006. Instruments for Causal Inference. An Epidemiologist's Dream? *Epidemiology*, **17**, 360–372.

Hernán, M.A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. A. E., & Robins, J. M. 2008. Observational Studies Analyzed Like Randomized Experiments: An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease. *Epidemiology*, **19**(6), 766–779.

Higgins, J. P. T., & Thompson, S. G. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, **21**, 1539–1558.

Hingorani, A., & Humphries, S. 2005. Nature's randomised trials. *The Lancet*, **366**(9501), 1906–1908.

Hinkley, D. V. 1969. On the Ratio of Two Correlated Normal Random Variables.

*Biometrika*, **56**(3), 635–639.

Hinkley, D. V. 1970. Correction: On the Ratio of Two Correlated Normal Random Variables. *Biometrika*, **57**(3), 683.

Hirschfield, G. M., & Pepys, M. B. 2003. C-reactive protein and cardiovascular disease: new insights from an old molecule. *QJM*, **96**, 793–807.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. 2002. A comprehensive review of genetic association studies. *Genetics in Medicine*, **4**(2), 45–61.

Hole, A. R. 2006. Calculating Murphy-Topel variance estimates in Stata: a simplified procedure. *The Stata Journal*, **6**(4), 521–529.

Holland, P. W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**(396), 945–960.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. 1998. Comparison of Population-Averaged and Subject-Specific Approaches for Analyzing Repeated Binary Outcomes. *American Journal of Epidemiology*, **147**(7), 694–703.

Huber, P. J. 1967. The behaviour of maximum likelihood estimators under nonstandard conditions. *In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley: University of California Press.

Ioannidis, J. P. A., Gwinn, M., Little, J., Higgins, J. P. T., Bernstein, J. L., Boffetta, P., Bondy, M., Bray, M. S., Brenchley, P. E., Buffler, P. A., Casas, J. P., Chokkalingam, A., Danesh, J., Davey Smith, G., Dolan, S., Duncan, R., Gruis, N. A., Hartge, P., Hashibe, M., Hunter, D. J., Jarvelin, M-R., Malmer, B., Maraganore, D. M., Newton-Bishop, J. A., O'Brien, T. R., Petersen, G., Riboli, E., Salanti, G., Seminara, D., Smeeth, L., Taioli, E., Timpson, N., Uitterlinden, A. G., Vineis, P., Wareham, N., Winn, D. M., & Zimmern, R. 2006. A road map for efficient and reliable human genome epidemiology. *Nature Genetics*, **38**, 3–5.

Ioannidis, J. P. A., Patsopoulos, N. A., & Evangelou, E. 2007. Uncertainty in heterogeneity estimates in meta-analyses. *British Medical Journal*, **335**(7626), 914–916.

Ishak, K. J., Platt, R. W., Joseph, L., & Hanley, J. A. 2008. Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine*, **27**(5), 670–686.

Jewell, N. P. 2003. *Statistics for Epidemiology.* Boca Raton, US: Chapman & Hall/CRC.

Jewell, N. P., & Shiboski, S. C. 1990. Statistical Analysis of HIV Infectivity Based on Partner Studies. *Biometrics*, **46**(4), 1133–1150.

Johnson, N. L., & Kotz, S. 1970. *Distributions in Statistics, Continuous Univariate Distributions.* Vol. 2. Boston: Houghton-Mifflin.

Johnson, N. L., & Kotz, S. 1994. *Distributions in Statistics, Continuous Univariate Distributions.* Second edn. Vol. 1. Boston: Houghton-Mifflin.

Johnston, K. M., Gustafson, P., Levy, A. R., & 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**, 1539–1556.

Jones, D. R. 1995. Meta-analysis: Weighing the evidence. *Statistics in Medicine*, **14**(2), 137–149.

Jousilahti, P., & Salomaa, V. 2004. Fibrinogen, social position, and "Mendelian randomisation". *Journal of Epidemiology and Community Health*, **58**(10), 883.

Kamath, S., & Lip, G. Y. H. 2003. Fibrinogen: biochemistry, epidemiology and determinants. *QJM*, **96**, 711–729.

Karaca-Mandic, P., & Train, K. 2003 (July). *Standard Error Correction in Two-Stage Estimation with Nested Samples.* Tech. rept. University of California, Berkeley.

Katan, M. B. 1986. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*, **327**, 507–508.

Katan, M. B. 2004. Apolipoprotein E isoforms, serum cholesterol, and cancer. *International Journal of Epidemiology*, **33**, 9.

Kavvoura, F. K., & Ioannidis, J. P. A. 2008. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Human Genetics*, **123**(1), 1–14.

Keavney, B., Palmer, A., Parish, S., Clark, S., Youngman, L., Danesh, J., Mckenzie, C., Delepine, M., Lathrop, M., Peto, R., & Collins, R. 2004. Lipid-related genes and myocardial infarction in 4685 cases and 3460 controls: discrepancies between genotype, blood lipid concentrations, and coronary disease risk. *International Journal of Epidemiology*, **33**(5), 1002–1013.

Keavney, B., Danesh, J., Parish, S., Palmer, A., Clark, S., Youngman, L., Delepine, M., Lathrop, M., Peto, R., & Collins, R. 2006. Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'. *International Journal of Epidemiology*, **35**, 935–943.

Kendall, M., & Stuart, A. 1977. *The advanced theory of statistics. Vol. 1: Distribution theory.* New York: Macmillan.

Keogh-Brown, M. R., Bachmann, M. O., Shepstone, L., Hewitt, C., Howe, A., Ramsay, C. R., Song, F., Miles, J. N. V., Torgerson, D. J., Miles, S., Elbourne, D., Harvey, I., & Campbell, M. J. 2007. Contamination in trials of educational interventions. *Health Technology Assessment*, **11**(43).

Keshk, O. M. G. 2003. CDSIMEQ: A program to implement two-stage probit least squares. *The Stata Journal*, **3**(2), 157–167.

Keys, A., Aravanis, C., Blackburn, H., Buzina, R., Dontas, A. S., Fidanza, F., Karvonen, M. J., Menotti, A., Nedeljkovic, S., Punsar, S., & Toshima, H. 1985. Serum cholesterol

and cancer mortality in the Seven Countries Study. *American Journal of Epidemiology*, **121**(6), 870–883.

Khoury, M. J., Davies, R., Gwinn, M., Lindegren, M. L., & Yoon, P. 2005. Do We Need Genomic Research for the Prevention of Common Diseases with Environmental Causes? *American Journal of Epidemiology*, **161**, 799–805.

Kivimäki, M., Lawlor, D. A., Eklund, C., Smith, G. Davey, Hurme, M., Lehtimäki, T., Viikari, J. S. A., & Raitakari, O. T. 2007. Mendelian Randomization Suggests No Causal Association Between C-reactive Protein and Carotid Intima-media Thickness in the Young Finns Study. *Arteriosclerosis, Thrombosis and Vascular Biology*, **27**, 978–979.

Kivimäki, M., Davey Smith, G., Timpson, N. J., Lawlor, D. A., Batty, G. D., Kahonen, M., Juonala, M., Ronnemaa, T., Viikari, J. S. A., Lehtimaki, T., & Raitakari, O. T. 2008. Lifetime body mass index and later atherosclerosis risk in young adults: examining causal links using Mendelian randomization in the Cardiovascular Risk in Young Finns study. *European Heart Journal*, ehn252.

Knowler, W. C., Williams, R. C., Pettitt, D. J., & Steinberg, A. G. 1988. Gm3; 5, 13, 14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American Journal of Human Genetics*, **43**(4), 520.

Kolz, M., Koenig, W., Muller, M., Andreani, M., Greven, S., Illig, T., Khuseyinova, N., Panagiotakos, D., Pershagen, G., Salomaa, V., Sunyer, J., Peters, A., & for the AIRGENE Study Group. 2008. DNA variants, plasma levels and variability of C-reactive protein in myocardial infarction survivors: results from the AIRGENE study. *Eur Heart J*, **29**(10), 1250–1258.

Lauritzen, S. L., & Sheehan, N. A. 2007. *Handbook of Statistical Genetics*. Third edn. Vol. 2. Chichester, UK: Wiley. Chap. Graphical Models in Genetics, pages 808–842.

Lawlor, D. A., Davey Smith, G., & Ebrahim, S. 2004. Commentary: The hormone

replacement-coronary heart disease conundrum: is this the death of observational epidemiology? *International Journal of Epidemiology*, **33**(3), 464–467.

Lawlor, D. A., Timpson, N., Davey Smith, G., Harbord, R., & Sterne, J. A. C. 2008a. Comments on 'Mendelian randomization: Using genes as instruments for making causal inference in epidemiology': Author's response. *Statistics in Medicine*, **27**, 2976–2978.

Lawlor, D. A., Timpson, N. J., Harbord, R. M., Leary, S., Ness, A., McCarthy, M. I., Frayling, T. M., Hattersley, A. T., & Davey Smith, G. 2008b. Exploring the Developmental Overnutrition Hypothesis Using ParentalOffspring Associations and FTO as an Instrumental Variable. *Plos Medicine*, **5**(3), e33.

Lawlor, D. A., Windmeijer, F., & Davey Smith, G. 2008c. Is Mendelian randomization 'lost in translation?': Comments on 'Mendelian randomization equals instrumental variable analysis with genetic instruments' by Wehby et al. *Statistics in Medicine*, **27**, 2750–2755.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Davey Smith, G. 2008d. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, **27**(8), 1133–1163.

Lewis, S. J., & Davey Smith, G. 2005. Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiology, Biomarkers & Prevention*, **14**(8), 1967–1971.

Lewis, S. J., Ebrahim, S., & Davey Smith, G. 2005. Meta-analysis of MTHFR 677CT polymorphism and coronary heart disease: does totality of evidence support causal role for homocysteine and preventive potential of folate? *British Medical Journal*, **331**(7524), 1053–1056.

Lewis, S. J., Harbord, R. M., Harris, R., & Davey Smith, G. 2006. Meta-analyses of Observational and Genetic Association Studies of Folate Intakes or Levels and Breast Cancer Risk. *Journal of the National Cancer Institute*, **98**(22), 1607–1622.

Liang, K-Y., & Liu, X-H. 1991. *Estimating Functions*. Clarendon Press. Chap. Estimating

equations in generalized linear models with measurement error, pages 47–64.

Little, J., & Higgins, J. P. T. (eds). 2006. *The HuGENet HuGE Review Handbook, version 1.0*. The Human Genome Epidemiology Network. accessed 28 February 2006.

Little, J., & Khoury, M. J. 2003. Mendelian randomisation: a new spin or real progress? *The Lancet*, **362**, 930–931.

Little, J., Khoury, M. J., Bradley, L., Clyne, M., Gwinn, M., Lin, B., Lindegren, M-L., & Yoon, P. 2003. The Human Genome Project is complete. How do we de develop a handle for the pump? *American Journal of Epdiemiology*, **157**, 667–673.

Little, S., Konig, I., Franjkovic, I., Stricker, J., Colaris, T., Martens, F., Langefeld, T., Weismuller, K., Focke, J., Hackstein, H., Bohnert, A., Hempelmann, G., Menges, T., Chakraborty, T., & Bein, G. 2006. Genetic variation of TNF is associated with sepsis syndrome and death in severely injured patients. *Critical Care*, **10**(Suppl 1), P145.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Mann, V., Hobson, E. E., Li, B., Stewart, T. L., Grant, S. F. A., Robins, S. P., Aspden, R. M., & Ralston, S. H. 2001. A *COL1A1* Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality. *The Journal of Clinical Investigation*, **107**(7), 899–907.

Manski, C. F. 1990. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, **80**, 319–323.

Mantel, N., & Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719–748.

Marsaglia, G. 1965. Ratios of Normal Variables and Ratios of Sums of Uniform Variables. *Journal of the American Statistical Association*, **60**(309), 193–204.

Marsaglia, G. 2006. Ratios of Normal Variables. *Journal of Statistical Software*, **16**(4), 1–10.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V., & Klungel, O. H. 2006. Instrumental Variables: Application and Limitations. *Epidemiology*, **17**, 260–267.

McMichael, A. J., Jensen, O. M., Parkin, D. M., & Zaridze, D. G. 1984. Dietary and endogenous cholesterol and human cancer. *Epidemiologic Reviews*, **6**(1), 192–216.

Mendel, G. 1865. *Experiments in Plant Hybridization.*

Meng, S. 2008 (July). Causal Inference from Observational Data using Mendelian Randomisation. *In: XXIV$^{th}$ International Biometric Conference.*

Menges, T., König, I. R., Hossain, H., Little, S., Tchatalbachev, S., Hackstein, H., Franjkovic, I., Colaris, T., Martens, F., Weismüller, K., Langefeld, T., Stricker, J., Hemplemann, G., Vos, P. E., Zeigler, A., Jacobs, R., Chakraborty, T., & Bein, G. 2008. Sepsis syndrome and death in trauma patients is associated with variations in the gene encoding TNF. *Critical Care Medicine.* under review.

Miettinen, O. S., & Cook, E. F. 1981. Confounding: essence and detection. *American Journal of Epidemiology*, **114**(4), 593–603.

Mikusheva, A., & Poi, B. 2006. Tests and confidence sets with correct size when instruments are potentially weak. *The Stata Journal*, **6**(3), 335–347.

Minelli, C., Thompson, J. R., Tobin, M. D., & Abrams, K. R. 2003. Meta-analytical methods for the synthesis of genetic studies using Mendelian randomisation. *Technical Report 2003/GE2, Univeristy of Leicester.*

Minelli, C., Thompson, J. R., Tobin, M. D., & Abrams, K. R. 2004. An integrated approach

to the meta-analysis of Genetic Association studies using Mendelian randomization. *American Journal of Epidemiology*, **160**(5), 445–452.

Minelli, C., Thompson, J. R., Abrams, K. R., & Lambert, P. C. 2005a. Bayesian implementation of a genetic model-free approach to the meta-analysis of genetic association studies. *Statistics in Medicine*, **24**, 3845–3861.

Minelli, C., Thompson, J. R., Abrams, K. R., Thakkinstian, A., & Attia, J. 2005b. The choice of a genetic model in the meta-analysis of molecular association studies. *International Journal of Epidemiology*, **34**, 1319–1328.

Mullahy, J. 1997. Instrumental–variable estimation of count data models: Applications to models of cigarette smoking behaviour. *The Review of Economics and Statistics*, **79**(4), 568–593.

Murphy, K. M., & Topel, R. H. 1985. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, **3**(4), 370–379.

Nagelkerke, N., Fidler, V., Bernsen, R., & Borgdorff, M. 2000. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, **19**(14), 1849–64. Erratum Stat Med 2001; 20: 982.

Naylor, C. D. 1997. Meta-analysis and the meta-epidemiology of clinical research. *British Medical Journal*, **315**(7109), 617.

Neuhaus, J. M. 1993. Estimation Efficiency and Tests of Covariate Effects with Clustered Binary Data. *Biometrics*, **49**(4), 989–996.

Neuhaus, J. M. 1998. Estimation Efficiency with Omitted Covariates in Generalized Linear Models. *Journal of the American Statistical Association*, **93**, 1124–1129.

Neuhaus, J. M., & Jewell, N. P. 1993. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, **80**(4), 807–815.

Neuhaus, J. M., Kalbfleisch, J. D., & Hauck, W. W. 1991. A Comparison of Cluster-

Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. *International Statistical Review*, **59**(1), 25–35.

Nichols, A. 2006. Weak instruments: An overview and new techniques. *In: North American Stata Users' Group Meeting.*

Nichols, A. 2007a. Causal inference with observational data. *The Stata Journal*, **7**(4), 507–541.

Nichols, A. 2007b. *IVPOIS: Stata module to estimate an instrumental variables Poisson regression via GMM.* Statistical Software Components, Boston College Department of Economics. available at http://ideas.repec.org/c/boc/bocode/s456890.html.

Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B. L., Whittaker, J. C., & Leon, D. A. 2006. Limits to Causal Inference based on Mendelian Randomization: A Comparison with Randomized Controlled Trials. *American Journal of Epidemiology*, **163**, 397–403.

Norby, S. 2005. Mendelian randomization. *Ugeskr Laeger*, **167**(19), 2074.

Novotny, L., & Bencko, V. 2007. Genotype-disease association and possibility to reveal environmentally modifiable disease causes: the use of mendelian randomization principle. *Cas Lek Cesk*, **146**(4), 343–50.

O'Brien, B. J., Drummond, M. F., Labelle, R. J., & Willan, A. 1994. In Search of Power and Significance: Issues in the Design and Analysis of Stochastic Cost-Effectiveness Studies in Health Care. *Medical Care*, **32**(2), 150–163.

Olsen, J., & Thulstrup, A. M. 2005. Mendelsk randomisering. Dansk Epidemiologisk Selskab. *Ugeskr Laeger*, **167**, 1383.

Owen, D. B., Craswell, K. J., & Hanson, D. L. 1964. Nonparametric Upper Confidence Bounds for $P(Y < X)$ and Confidence Limits for $P(Y < X)$ When $X$ and $Y$ are Normal. *Journal of the American Statistical Association*, **59**(307), 906–924.

Pagan, A. 1984. Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review*, **25**(1), 221–247.

Palmer, L. J. 2007. UK Biobank: bank on it. *The Lancet*, **369**(9578), 1980–1982.

Palmer, T. M., Thompson, J. R., Tobin, M. D., Sheehan, N. A., & Burton, P. R. 2008a. Adjusting for bias and unmeasured confounding in the analysis of Mendelian randomization studies with binary responses. *International Journal of Epidemiology*, **37**(5), 1161–1168.

Palmer, T. M., Peters, J. L., Sutton, A. J., & Moreno, S. G. 2008b. Contour enhanced funnel plots for meta-analysis. *The Stata Journal*, **8**(2), 242–254.

Palmer, T. M., Thompson, J. R., & Tobin, M. D. 2008c. Meta-analysis of Mendelian randomization studies incorporating all three genotypes. *Statistics in Medicine*, **27**(30), 6570–6582.

Pawitan, Y. 2001. *In All Likelihood*. Oxford: Oxford University Press.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, **82**(4), 669–688.

Pearl, J. 1998 (July). *Why There Is No Statistical Test For Confounding, Why Many Think There Is, And Why They Are Almost Right*. Tech. rept. R-256. University of California, Los Angeles.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearl, J. 2001. Causal Inference in Statistics: A Gentle Introduction. *In:* Wegman, Edward J., Braverman, Amy, Goodman, Arnold, & Smyth, Padhraic (eds), *Proceedings of the 33$^{rd}$ Symposium on the Interface*. Computing Science and Statistics, vol. 33. Interface Foundation of North America.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. 2006. Comparison

of Two Methods to Detect Publication Bias in Meta-analysis. *Journal of the American Medical Association*, **295**(6), 676–680.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. 2008. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, **61**(10), 991–996.

Petrin, A., & Train, K. 2003. *Omitted product attributes in discrete choice models.* Tech. rept. National Bureau of Economic Research.

PHOEBE. 2007. *Mendelian randomisation: inferring causality in observational epidemiology.* Tech. rept. Promoting Harmonization of Epidemiological Biobanks in Europe.

Qi, L., Rifai, N., & Hu, F. B. 2007. Interleukin-6 Receptor Gene Variations, Plasma Interleukin-6 Levels, and Type 2 Diabetes in U.S. Women. *Diabetes*, **56**(12), 3075–3081.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ralston, S. H., Uitterlinden, A. G., Brandi, M. L., Balcells, S., Langdahl, B. L., Lips, P., Lorenc, R., Obermayer-Pietsch, B., Scollen, S., Bustamante, M., Husted, L. B., Carey, A. H., Diez-Perez, A., Dunning, A. M., Falchetti, A., Karczmarewicz, E., Kruk, M., van Leeuwen, J. P. T. M., van Meurs, J. B. J., Mangion, J., McGuigan, F. E. A., Mellibovsky, L., del Monte, F., Pols, H. A. P., Reeve, J., Reid, D. M., Renner, W., Rivadeneira, F., van Schoor, N. M., Sherlock, R. E., & Ioannidis, J. P. A. 2006. Large-Scale Evidence for the Effect of the COLIA1 Sp1 Polymorphism on Osteoporosis Outcomes: The GENOMOS Study. *PLoS Medicine*, **3**(4), e90, 0515–0523.

Reiersol, O. 1941. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, **9**(1), 1–24.

Reiersol, O. 1945. Confluence analysis by means instrumental sets of variables. *Arkiv for Mathematik Astronomi och Fysik*, **32**, 1–119.

Richardson, S., Stücker, I., & Hémon, D. 1987. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology*, **16**(1), 111–120.

Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C., & Thompson, J. R. 2007a. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, **7**, 3.

Riley, R. D., Abrams, K. R., Lambert, P. C., Sutton, A. J., & Thompson, J. R. 2007b. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*, **26**(1), 78–97.

Riley, R. D., Thompson, J. R., & Abrams, K. R. 2008. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, **9**(1), 172–186.

Ritz, J., & Spiegelman, D. 2004. A note about the equivalence of conditional and marginal regression models. *Statistical Methods in Medical Research*, **13**, 309–23.

Rivers, D., & Vuong, Q. H. 1988. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, **39**, 347–366.

Robins, J. M., Hernan, M. A., & Brumback, B. 2000. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, **11**(5), 550–560.

Rodriguez, S., Gaunt, T. R., & Day, I. N. M. 2009. Hardy-Weinberg Equilibrium Testing of Biological Ascertainment for Mendelian Randomization Studies. *American Journal of Epidemiology*. in press.

Roodman, D. M. 2008. *CMP: Stata module to implement conditional (recursive) mixed process estimator*. Tech. rept. Centre for Global Development, Washington DC.

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., Jackson, R. D., Beresford, S. A., Howard, B. V., Johnson, K. C., Kotchen,

J. M., & Ockene, J. 2002. Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial. *Journal of the American Medical Association*, **288**(3), 321–33.

Rothman, K. J., Greenland, S., & Lash, T. L. 2008. *Modern Epidemiology.* Philadelphia, US: Lippincott, Williams and Wilkins. Chap. Validity in Epidemiologic Studies, pages 128–147.

Salanti, G., & Higgins, J. P. T. 2008. Meta-analysis of genetic association studies under different inheritance models using data reported as merged genotypes. *Statistics in Medicine*, **27**(5), 764–777.

Salanti, G., Amountza, G., Ntzani, E. E., & Ioannidis, J. P. A. 2005a. Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. *European Journal of Human Genetics*, **13**, 840–848.

Salanti, G., Sanderson, S., & Higgins, J. P. T. 2005b. Obstacles and opportunities in meta-analysis of genetic association studies. *Genetics in Medicine*, **7**, 13–20.

Salanti, G., Higgins, J. P. T., Trikalinos, T. A., & Ioannidis, J. P. A. 2007. Bayesian meta-analysis and meta-regression for gene-disease associations and deviations from Hardy–Weinberg Equilibrium. *Statistics in Medicine*, **26**, 553–567.

Sargan, J. D. 1958. The Estimation of Economic Relationships Using Instrumental Variables. *Econometrica*, **26**(3), 393–415.

Schaffer, M. E. 2005 (November). *XTIVREG2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models.* Statistical Software Components, Boston College Department of Economics.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association*, **273**(5), 408–412.

Sheehan, N. A., Didelez, V., Burton, P. R., & Tobin, M. D. 2008. Mendelian Randomisation and Causal Inference in Observational Epidemiology. *PLoS Medicine*, **5**(8), e177. in press.

Smith, D. M., & Diggle, P. J. 1998. Compliance in an anti-hypertension trial: a latent process model for binary longitudinal data. *Statistics in Medicine*, **17**(3), 357–370.

Speed, T. P., & Zhao, H. 2007. *Handbook of Statistical Genetics*. Third edn. Vol. 1. Chichester, UK: Wiley. Chap. Chromosome Maps, pages 3–39.

Spiegel, M. R. 1971. *Schaum's Outline of Theory and Problems of Advanced Mathematics for Engineers and Scientists*. McGraw-Hill Book Company.

Spiegelhalter, D. J. 1998. Bayesian graphical modelling: a case-study in monitoring health outcomes. *Applied Statistics*, **47**(1), 115–133.

Staiger, D., & Stock, J. H. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica*, **65**(3), 557–586.

Stata Corp. 2007. *Stata Reference Manual, Release 10, I–P*. Stata Press. Chap. ivprobit - Probit model with endogenous regressors, pages 15–25.

Stefanski, L. A. 1985. The effects of measurement error on parameter estimation. *Biometrika*, **72**(3), 583–592.

Stefanski, L. A., & Boos, D. D. 2002. The Calculus of M-Estimation. *The American Statistician*, **56**(1), 29–38.

Sterne, J. A. C., Jueni, P., Schulz, K. F., Altman, D. G., Bartlett, C., & Egger, M. 2002. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Statistics in Medicine*, **21**(11), 1513–1524.

Stewart, J., & Gill, L. 1998. *Econometrics*. second edn. Englewood Cliffs, New Jersey: Prentice and Hall.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T. A., & Thacker, S. B. 2000. Meta-analysis of Observational Studies in Epidemiology. *Journal of the American Medical Association*, **283**(15), 2008–2012.

Sunyer, J., Pistelli, R., Plana, E., Andreani, M., Baldari, F., Kolz, M., Koenig, W., Pekkanen, J., Peters, A., & Forastiere, F. 2008. Systemic inflammation, genetic susceptibility and lung function. *European Respiratory Journal*, **32**, 92–97.

Sutton, A. J., & Higgins, J. P. T. 2008. Recent developments in meta-analysis. *Statistics in Medicine*, **27**(5), 625–650.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. 2000. *Methods for Meta-Analysis in Medical Research*. Chichester: Wiley.

Sweeting, M.J., Sutton, A.J., & Lambert, P.C. 2004. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, **23**, 1351–1375.

Ten Have, T. R., Richard Landis, J., & Hartzel, J. 1996. Population-averaged and cluster-specific models for clustered ordinal response data. *Statistics in Medicine*, **15**(23), 2573–2588.

Ten Have, T. R., Joffe, M., & Cary, M. 2003. Causal logistic models for non-compliance under randomized treatment with univariate binary response. *Statistics in Medicine*, **22**, 1255–1283.

Thakkinstian, A., McElduff, P., D'Este, C., Duffy, D., & Attia, J. 2005. A method for meta-analysis of molecular association studies. *Statistics in Medicine*, **24**, 1291–1306.

The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Theil, H. 1953. Repeated Least Squares Applied to Complete Equation Systems. *The*

*Hague: Central Planning Bureau.*

Thomas, D. C., & Conti, D. V. 2004. Commentary: The concept of 'Mendelian randomization'. *International Journal of Epidemiology*, **33**, 21–25.

Thomas, D. C., & Witte, J. S. 2002. Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations? *Cancer Epidemiology Biomarkers & Prevention*, **11**(6), 505–512.

Thomas, D. C., Lawlor, D. A., & Thompson, J. R. 2007. Re: Estimation of Bias in Nongenetic Observational Studies Using "Mendelian Triangulation" by Bautista et al. *Annals of Epidemiology*, **17**(7), 511–513.

Thompson, J. R., Tobin, M. D., & Minelli, C. 2003. *GE1: On the accuracy of estimates of the effect of phenotype on disease derived from Mendelian randomisation studies.* Tech. rept. University of Leicester, Leicester.

Thompson, J. R., Minelli, C., Abrams, K. R., Tobin, M. D., & Riley, R. D. 2005. Meta-analysis of genetic studies using Mendelian randomization - a multivariate approach. *Statistics in Medicine*, **24**, 2241–2254.

Thompson, J. R., Minelli, C., Abrams, K. R., Thakkinstian, A., & Attia, J. 2008. Combining information from related meta-analyses of genetic association studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**(1), 103–115.

Thompson, S. G., Smith, T. C., & Sharp, S. J. 1997. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine*, **16**(23), 2741–2758.

Timpson, N. J., Lawlor, D. A., Harbord, R. M., Gaunt, T. R., Day, I. N. M., Palmer, L. J., Hattersley, A. T., Ebrahim, S., Lowe, G. D. O., Rumley, A., & Davey Smith, G. 2005. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. *The Lancet*, **366**, 1954–1959.

Tobin, M. D., Minelli, C., Burton, P. R., & Thompson, J. R. 2004. Commentary: Develop-

ment of Mendelian randomization: from hypothesis test to 'Mendelian deconfounding'. *International Journal of Epidemiology*, **33**(1), 26–29.

Uitterlinden, A. G., Burger, H., Huang, Q., Yue, F., McGuigan, F. E., Grant, S. F., Hofman, A., van Leeuwen, J. P., Pols, H. A., & Ralston, S. H. 1998. Relation of alleles of the collagen type I $\alpha$ 1 gene to bone density and the risk of osteoporotic fractures in postmenopausal women. *N Engl J Med*, **338**(15), 1016–21.

van Houwelingen, H. C., & Senn, S. 1999. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine*, **18**(1), 110–5.

van Houwelingen, H. C., Zwindermann, K. H., & Stijnen, T. 1993. A bivariate approach to meta-analysis. *Statistics in Medicine*, **12**, 2273–2284.

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. 2002. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, **21**, 589–624.

Vandenbroucke, J. P. 2004. When are observational studies as credible as randomised trials? *The Lancet*, **363**(9422), 1728–1731.

Vansteelandt, S., & Goetghebeur, E. 2003. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B*, **65**(4), 817–835.

Venables, W. N., & Ripley, B. D. 2002. *Modern Applied Statistics with S*. Fourth edn. New York: Springer.

Verzilli, C., Shah, T., Casas, J. P., Chapman, J., Sandhu, M., Debenham, S. L., Boekholdt, M. S., Khaw, K. T., Wareham, N. J., Judson, R., Benjamin, Emelia J., Kathiresan, Sekar, Larson, Martin G., Rong, Jian, Sofat, Reecha, Humphries, Steve E., Smeeth, Liam, Cavalleri, Gianpiero, Whittaker, John C., & Hingorani, Aroon D. 2008. Bayesian meta-analysis of genetic association studies with different sets of markers. *American Journal of Human Genetics*, **82**(4), 859–72.

Vineis, P. 2004. A self-fulfilling prophecy: are we underestimating the role of the environment in gene-environment interaction research? *International Journal of Epidemiology*, **33**, 945–946.

Wacholder, S., Rothman, N., & Caporaso, N. 2002. Counterpoint: Bias from Population Stratification Is Not a Major Threat to the Validity of Conclusions from Epidemiological Studies of Common Polymorphisms and Cancer. *Cancer Epidemiology Biomarkers & Prevention*, **11**(6), 513–520.

Wald, A. 1940. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, **11**(3), 284–300.

Wald, N., & Sneddon, J. 1991. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. *Lancet*, **338**(8760), 131–137.

Walter, S. D., Gafni, A., & Birch, S. 2008. A geometric confidence ellipse approach to the estimation of the ratio of two variables. *Statistics in Medicine*. in press.

Wang, Z., & Louis, T. A. 2003. Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika*, **90**(4), 765–775.

Wehby, G. L., Ohsfeldt, R. L., & Murray, J. C. 2008. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Statistics in Medicine*, **27**, 2745–2749.

Weinberg, C. R. 1993. Toward a clearer definition of confounding. *American Journal of Epidemiology*, **137**(1), 1–8.

Wermuth, N. 1987. Parametric Collapsibility and the Lack of Moderating Effects in Contingency Tables with a Dichotomous Response Variable. *Journal of the Royal Statistical Society. Series B (Methodological)*, **49**(3), 353–364.

Wheatley, K. 2002. Current controversies: which patients with acute myeloid leukaemia

should receive a bone marrow transplantation?- A statistician's view. *British Journal of Haematology*, **118**(2), 351–356.

Wheatley, K., & Gray, R. 2004. Commentary: Mendelian randomization - an update on its use to evaluate allogenic stem cell transplantation in leukaemia. *International Journal of Epidemiology*, **33**(1), 15–17.

Whitehead, A., & Whitehead, J. 1991. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, **10**(11), 1665–1677.

Wiggins, V. 2000. *Must I use all of my exogenous variables as instruments when estimating instrumental variables regression?* Tech. rept. Stata Corp., College Station, Texas, US.

Wijsman, E. M. 2002. *Biostatistical Genetics and Genetic Epidemiology.* Chichester: Wiley. Chap. Mendel's Laws, pages 527–529.

Windmeijer, F. A. G., & Santos Silva, J. M. C. 1997. Endogeneity in Count Data Models: An Application to Demand for Health Care. *Journal of Applied Econometrics*, **12**(3), 281–294.

Wolter, K. M. 2003. *Introduction to Variance Estimation.* Springer.

Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Juni, P., Altman, D. G., Gluud, C., Martin, R. M., Wood, A. J. G., & Sterne, J. A. C. 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *British Medical Journal*, **336**(7644), 601–605.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data.* MIT Press.

Youngman, L. D., Keavney, B. D., Palmer, A., Parish, S., Clark, S., Danesh, J., Delepine, M., Lathrop, M., Peto, R., & Collins, R. 2000. Plasma fibrinogen and fibrinogen genotypes in 4685 cases of myocardial infarction and in 6002 controls: test of causality by Mendelian randomization. *Circulation*, **102**(Supplement II), 31–2.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. 1985. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases*, **27**(5), 335–371.

Zeger, S. L., Liang, K-Y., & Albert, P. S. 1988. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, **44**, 1049–1060.

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R. B., Rayner, N. W., Freathy, R. M., Barrett, J. C., Shields, B., Morris, A. P., Ellard, S., Groves, C. J., Harries, L. W., Marchini, J. L., Owen, K. R., Knight, B., Cardon, L. R., Walker, M., Hitman, G. A., Morris, A. D., Doney, A. S. F., (WTCCC), The Wellcome Trust Case Control Consortium, McCarthy, M. I., & Hattersley, A. T. 2007. Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. *Science*, **316**(5829), 1336–1341.

Ziegler, A., & König, I. R. 2006. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Weinheim, Germany: Wiley-VCH.

Ziegler, A., König, I. R., & Thompson, J. R. 2008a. Biostatistical Aspects of Genome-Wide Association Studies. *The Biometrical Journal*, **1**, 1–21.

Ziegler, A., Pahlke, F., & König, I. 2008b. Comments on 'Mendelian randomization: using genes as instruments for making causal inferences in epidemiology'. *Statistics in Medicine*, **27**, 2974–2976.

Zohoori, N., & Savitz, D. A. 1997. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. *Annals of Epidemiology*, **7**(4), 251–257.

# Addenda

From the list of publications, on page vii, papers 4 and 6 are included on the following pages. Paper 4 relates to the work about the adjusted instrumental variable estimator in Chapters 3 and 4 and Appendices B and C. Paper 6 relates to the work on the meta-analysis of Mendelian randomization studies in Chapter 5.

Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2008;37:1161–1168
doi:10.1093/ije/dyn080


# Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses

**Tom M Palmer,[1]* John R Thompson,[1] Martin D Tobin,[2] Nuala A Sheehan[2] and Paul R Burton[2]**

**Accepted** 3 April 2008

**Background** Mendelian randomization uses a carefully selected gene as an instrumental-variable (IV) to test or estimate an association between a phenotype and a disease. Classical IV analysis assumes linear relationships between the variables, but disease status is often binary and modelled by a logistic regression. When the linearity assumption between the variables does not hold the IV estimates will be biased. The extent of this bias in the phenotype-disease log odds ratio of a Mendelian randomization study is investigated.

**Methods** Three estimators termed direct, standard IV and adjusted IV, of the phenotype-disease log odds ratio are compared through a simulation study which incorporates unmeasured confounding. The simulations are verified using formulae relating marginal and conditional estimates given in the Appendix.

**Results** The simulations show that the direct estimator is biased by unmeasured confounding factors and the standard IV estimator is attenuated towards the null. Under most circumstances the adjusted IV estimator has the smallest bias, although it has inflated type I error when the unmeasured confounders have a large effect.

**Conclusions** In a Mendelian randomization study with a binary disease outcome the bias associated with estimating the phenotype-disease log odds ratio may be of practical importance and so estimates should be subject to a sensitivity analysis against different amounts of hypothesized confounding.

**Keywords** Instrumental-variable analysis, Mendelian randomization, bias, unobserved confounding

## Introduction

In traditional epidemiological studies the associations between biological phenotypes and diseases can be distorted by confounding or reverse causation. The aim of Mendelian randomization analysis is to test or estimate the association between a biological phenotype and a disease in the presence of unmeasured confounding.[1–3] This is achieved using a carefully selected gene as an instrumental-variable (IV).[4–7] When certain assumptions hold Mendelian randomization will remove the distorting effects and produce unconfounded estimates of the association between a phenotype and a disease.[3,8] Genes that influence the disease through their effect on the biological phenotype of interest can be used as instrumental-variables in the analysis because a subject's genotype is essentially

[1] Department of Health Sciences, University of Leicester, UK.
[2] Departments of Health Sciences and Genetics, University of Leicester, UK.
* Corresponding author. University of Leicester, Department of Health Sciences, 2nd Floor, Adrian Building, University Road, Leicester LE1 7RH, UK. E-mail: tmp8@le.ac.uk

randomly assigned before birth and thus should not be influenced by the many environmental and lifestyle factors that typically act as confounders in epidemiology.[9]

In this article, we show that, for binary outcomes, the observed bias towards the null in Mendelian randomization estimates is due to the impact of random effects that are not explicitly included in the linear predictor. This is analogous to the discrepancy between marginal and conditional parameter estimates in generalized linear mixed models with a logistic link.[10,11] Theoretical formulae for approximating this difference are provided for each of three different estimators and their accuracy is verified by simulation. In theory, knowledge of the difference between marginal and conditional estimates could provide a correction for the bias that pertains in Mendelian randomization analyses. However, the extent of this bias depends on the properties of the unmeasured confounders, which are always unknown. An adjusted instrumental-variable estimator is applied to Mendelian randomization analyses to produce an improved estimate of the phenotype-disease association. The adjusted IV estimator partially compensates for the unknown confounders by exploiting information from the residuals of the regression of the intermediate phenotype on the genotype.

# Methods

## Estimators for Mendelian randomization studies with binary responses

The key variables in describing the Mendelian randomization model are; the disease status ($Y$), intermediate phenotype ($X$), genotype ($G$) and confounder ($U$). The assumed relationship between these variables is shown in Figure 1. For the $i$th subject in a cohort, let $y_i$ represent their binary disease status, $p_i$ represent their probability of having the disease, $x_i$ represent the level of the biological phenotype and $g_i$ represent their genotype, which is coded 0, 1 and 2 to indicate the number of copies of the relevant risk allele. Typically there will be many unmeasured confounders, so it is assumed that they can be represented by a single variable, $u_i$, that captures their combined effect. This confounding variable is arbitrarily assumed to be standardized to have a mean of zero and a standard deviation of one. For simplicity, we assume an additive effect of genotype on the intermediate phenotype, although the argument would apply equally to any known mode of inheritance. It is also assumed that the confounder acts additively in the linear predictors of the associations between the genotype and phenotype and between the phenotype and the disease.

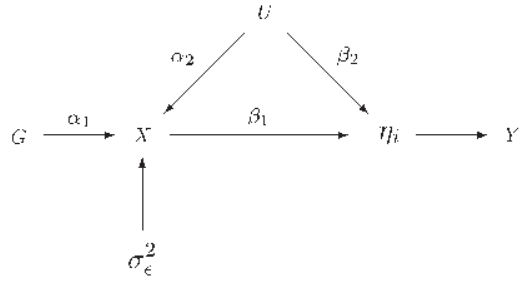



**Figure 1** The relationship between the variables ($\eta_i$ is the linear predictor of the logistic regression)

The coefficients in the regression of phenotype on genotype are denoted by $\alpha$'s so that,

$$x_i = \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i, \text{ with } \epsilon_i \sim N(0, \sigma^2_\epsilon), \quad (1)$$

and $\epsilon$ represents the effects of measurement error and unmeasured factors that are not confounders because they do not influence disease. The coefficients in the linear predictor between phenotype and disease are denoted by $\beta$'s, so that the disease status follows a Bernoulli distribution,

$$y_i \sim \text{Bern}(p_i), \text{ with } \log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i + \beta_2 u_i. \quad (2)$$

Implicit in the notation is the idea that $\epsilon_i$ and $u_i$ are independent of one another. The primary interest in this paper is to recover $\beta_1$.

If both regressions were linear, ignoring the confounder in the instrumental-variable analysis would not bias the estimate of $\beta_1$, but this is not the case for a non-linear relationship between phenotype and disease.[12] Substituting the formula for $x_i$ in Equation (1) into the logistic regression in Equation (2) gives,

$$\log\frac{p_i}{1-p_i} = \beta_0 + \beta_1(\alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i) + \beta_2 u_i. \quad (3)$$

The coefficient of $g_i$ in this relationship is $\beta_1\alpha_1$ while the coefficient of $g_i$ in the linear regression in Equation (1) is $\alpha_1$. In principle the ratio of the estimates of these coefficients should give an estimate of $\beta_1$,[4] which is the effect of the phenotype on disease risk after adjusting for confounding. Unfortunately $u_i$ and $\epsilon_i$ are unknown, so the estimate of $\beta_1\alpha_1$ is taken from the logistic regression without those terms, thus in effect replacing the true conditional model with a marginal model which averages over the unknown terms, $u_i$ and $\epsilon_i$.

An alternative to the ratio estimate of $\beta_1$ is obtained by taking the predicted values of the intermediate phenotype from the first regression ignoring the confounding,

$$\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 g_i \approx \alpha_0 + \alpha_1 g_i \quad (4)$$

and substituting those into the logistic regression in Equation (2), in which case,

$$\log\frac{p_i}{1-p_i} \approx \beta_0 + \beta_1(\hat{x}_i + \alpha_2 u_i + \epsilon_i) + \beta_2 u_i. \quad (5)$$

In this two-stage approach, the estimate of interest is just the coefficient of the predicted phenotype $\hat{x}_i$,

but the biases will be similar to those that occur for the ratio estimator.

In an attempt to correct for this difference between marginal and conditional parameter estimates, and thus improve upon the standard instrumental-variable estimator an adjusted IV estimator is applied. The estimated residuals from the first stage linear regression in Equation (1) are,

$$r_i = x_i - \hat{x}_i. \tag{6}$$

These estimated residuals capture some of the variability contained in the unknown confounders and the phenotype error term, $\epsilon$. This information can be used in the second regression by fitting,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \hat{x}_i + \beta_r r_i. \tag{7}$$

The information about the confounding contained in the residuals should, in part, compensate for the missing terms in the marginal form of the logistic regression model and therefore reduce the difference between the conditional and marginal estimates of $\beta_1$.

This article considers three estimators of $\beta_1$. First, the direct estimator, that does not use Mendelian randomization but performs a logistic regression of disease status on the intermediate as in a traditional epidemiological study. The direct estimator of $\beta_1$ is derived from the linear predictor,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i. \tag{8}$$

The standard IV estimator uses Mendelian randomization so that the linear predictor is,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \hat{x}_i. \tag{9}$$

The third estimator is the adjusted IV estimator obtained from Equation (7). In the Appendix, formulae are given for calculating the size of the bias in $\beta_1$ under the three estimators.

### Data simulation

A simulation study was performed to validate the formulae for the three estimators. In a cohort of size 10 000, subjects were each randomly assigned two alleles in Hardy-Weinberg equilibrium with the allele frequency of the risk allele set to 30%. The confounding variable was simulated to be normally distributed with mean zero and variance equal to one, $u_i \sim N(0,1)$. The phenotype, $x_i$, was generated as a Normal random variable with mean equal to, $\alpha_0 + \alpha_1 g_i + \alpha_2 u_i$ following Equation (1), and the standard deviation of the phenotype error term, $\sigma_\epsilon$, was set to one. Each subject's probability of disease was simulated, following Equation (2) such that $\log p_i/(1-p_i) = \beta_0 + \beta_1 y_i + \beta_2 u_i$.

The baseline prevalence of disease was set to 5% by fixing $\beta_0$. Different amounts of confounding were considered by changing the values of $\alpha_2$ and $\beta_2$. In particular, four confounding scenarios were considered by setting the confounding effect on the phenotype, $\alpha_2$, to 0, 1, 2 and 3 whilst the confounding effect on the disease, $\beta_2$, was varied between zero and three for each scenario. The other parameters were fixed as follows; $\alpha_0 = 0$, $\alpha_1 = 1$ and $\beta_1 = 1$. For each set of parameter values 10 000 simulations were performed. Statistical analysis was performed using R (version 2.6.1).[13]

## Results

The three estimators are assessed using the median parameter estimates, coverage probabilities and type I errors of the phenotype-disease log odds ratio, $\beta_1$. The coverage probability of $\beta_1$ was calculated as the proportion of simulations whose confidence interval included the true value of $\beta_1$. A set of simulations was performed with $\beta_1$ equal to 0 to represent the situation in which there is no association between phenotype and disease. For those simulations, the proportion of statistically significant estimates of $\beta_1$ is an estimate of the type I error of the Wald test of $\beta_1$.

### Assessment of the bias of the estimators

Figure 2 shows the median of $\beta_1$ for the three estimators from the simulations, represented by the symbols, and the values of the estimators calculated from the formulae given in the Appendix represented by the lines.

Figure 2 shows that the median values from the simulations are in close agreement with the theoretical predictions, there is the same pattern to the estimates of $\beta_1$ for the different values of $\alpha_2$ except when $\alpha_2$ is equal to zero. When $\alpha_2$ is equal to zero the direct and adjusted estimators are equivalent due to the assumptions underlying the relationship between the confounder and the phenotype. When $\alpha_2$ is non-zero, allowing the confounder to take effect, the direct estimate of $\beta_1$ is greater than the set value of one. However, the effect the unmeasured confounding has on the standard IV estimates is to bias them towards zero, producing estimates that are always below the true value of one. The values of the adjusted IV estimator are between the other two sets of estimates and have the smallest bias of the three estimators. For the adjusted IV estimates the bias in $\beta_1$ reduces with largest values of $\alpha_2$ because the estimated residuals are more informative.

### Assessment of the coverage probabilities of the estimators

Figure 3 shows the coverage probabilities of the three estimators, when the nominal level was 95%. The direct estimator and the standard IV estimator demonstrate very low coverage for all four scenarios due to the bias in $\beta_1$. The adjusted IV estimator

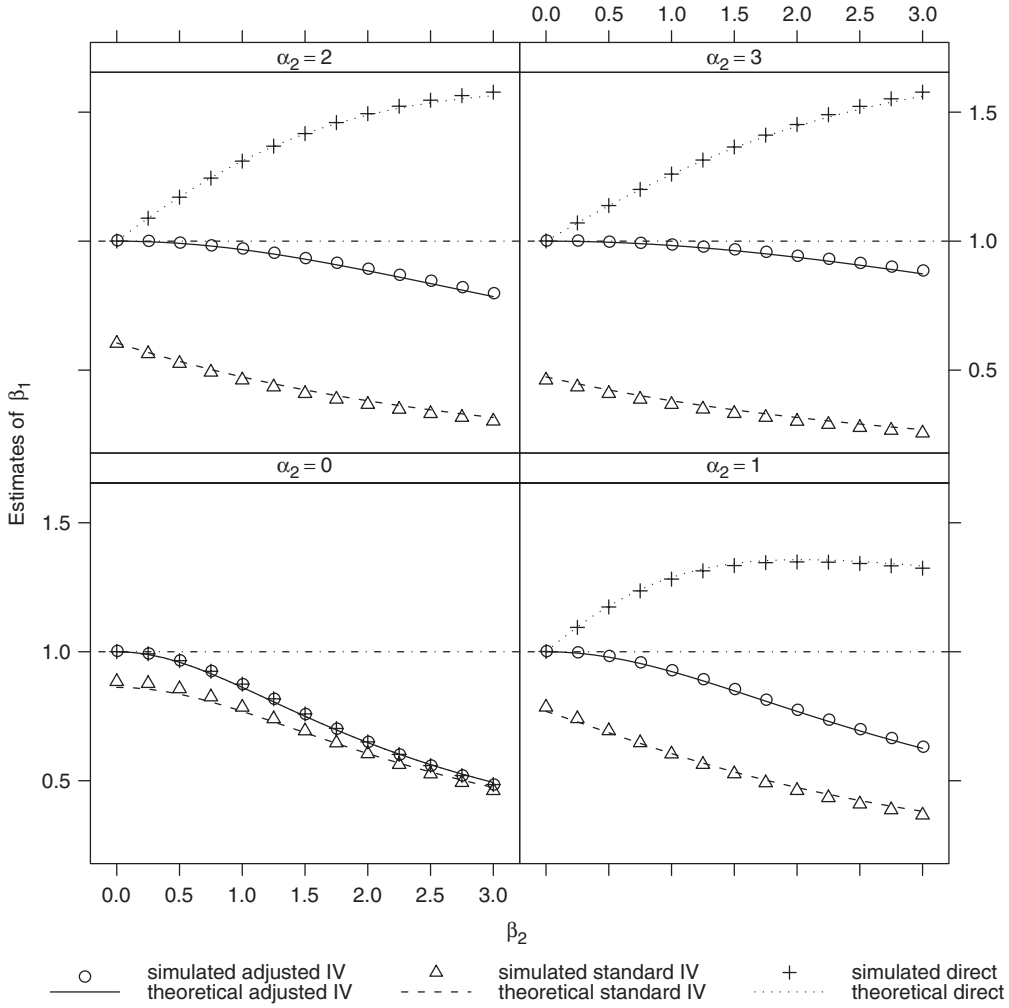



**Figure 2** Simulated and theoretical values of $\beta_1$

demonstrates the best coverage properties with levels around 95% over the range of values of $\beta_2$ for which its estimate of $\beta_1$ was approximately equal to the set value of one in Figure 2.

### Assessment of type I error

Figure 4 shows the type I error of the standard IV and adjusted IV estimators when the nominal rate is 5%. The type I error of the direct estimator is not shown on Figure 4 because the values were very large. Under the three scenarios with non-zero values of $\alpha_2$ the adjusted IV estimator has a substantially higher type I error rate than the standard IV estimator because the inclusion of the estimated residuals in the adjusted IV estimator reduced its estimated standard error.

## Discussion

This article considers the bias in the estimates from Mendelian randomization studies with binary outcomes. Three estimators of the phenotype-disease log odds ratio, termed; direct, standard IV and adjusted IV, have been evaluated through a simulation study. The simulations are in agreement with formulae relating conditional and marginal parameter estimates from logistic regression given in the Appendix. The adjusted IV estimator was the least biased, but it had high type I error when the effect of the unmeasured confounder was large. Further, unreported simulations show that the difference between marginal and conditional parameter estimates would also exist with probit regression and hence a similar but not identical adjustment between the conditional and marginal estimates of $\beta_1$ would be required if probit regressions were used in place of logistic regressions for the three estimators.[10]

The simulations investigated the performance of the estimators over a range of values of the confounder. Over the four panels in Figure 2, when $\alpha_2 = 0$, 1, 2 and 3, the confounder accounted for approximately 0%, 45%, 80% and 90% of the phenotype variance. For the log odds of disease the confounder accounted for between 0% and 90% of the variance in the linear

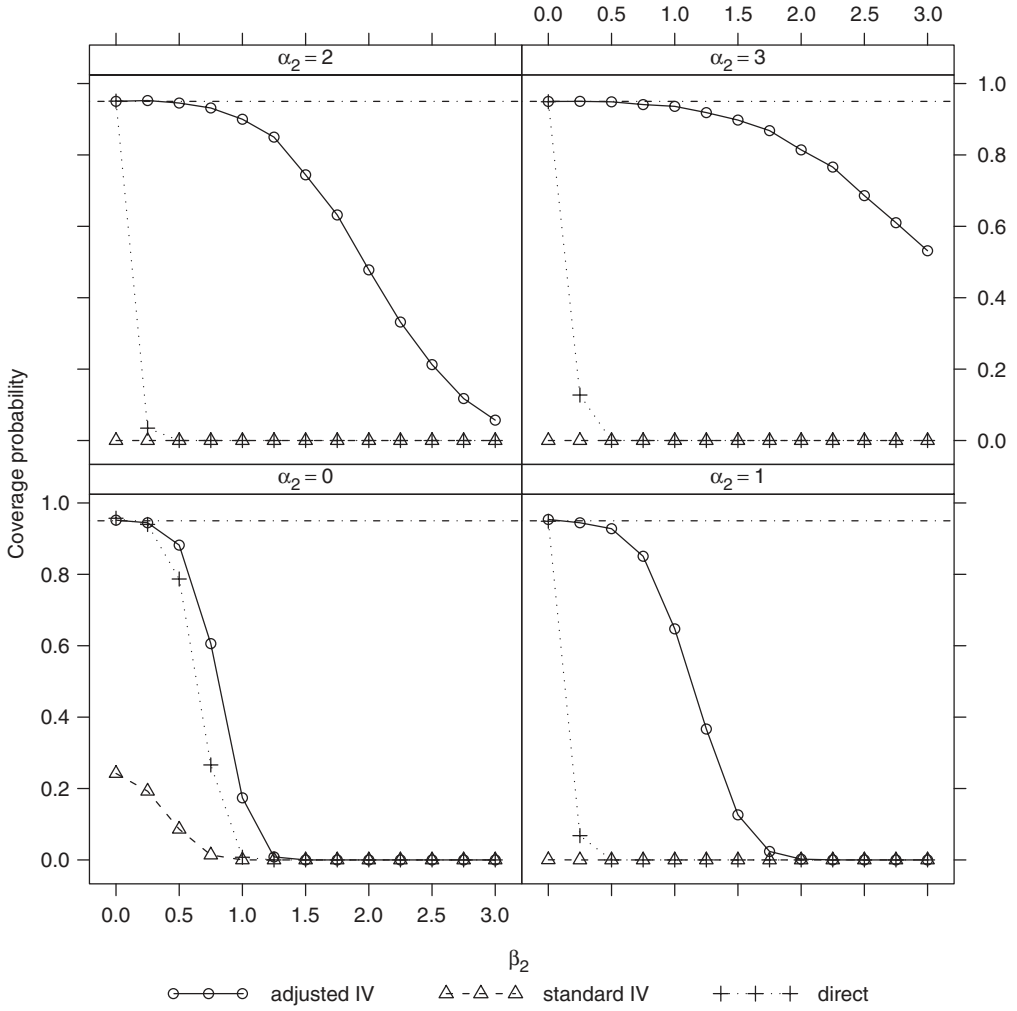



**Figure 3** Coverage probabilities of the three estimators

predictor when $\alpha_2 = 0$ and $\beta_2$ varied from 0 to 3, between 45% and 90% when $\alpha_2 = 1$, between 80% and 90% when $\alpha_2 = 2$ and between 85% and 95% when $\alpha_2 = 3$. Typically the gene used in a Mendelian randomization study will only explain a small percentage of the variance in the phenotype, perhaps $<10\%$. The impact of the confounders can therefore be large causing large bias. If it is possible to include measured confounders in the analysis this will reduce the importance of the unmeasured confounders and so reduce the bias in all of the estimators.

The adjusted IV estimator uses the estimated residuals as well as the predicted values from the first stage regression of the genotype on the phenotype as covariates in the second stage logistic regression between the phenotype and the disease outcome. A similar adjusted IV estimator was introduced in the context of clinical trials subject to non-compliance.[14] The first stage residuals contain some information about the unmeasured confounder since they capture the variance in the phenotype that is not explained by the genotype. The argument used in the clinical trials context was that these first stage residuals meet Pearl's back-door criterion and their inclusion in the model results in the adjusted IV estimate having a causal interpretation.[14]

Point estimates of causal effects from instrumental variable analyses require strong parametric and distributional assumptions, e.g. all relationships are linear without interactions.[6,15] Although the relationship between a gene and an intermediate phenotype might well be approximated by a linear regression, the final response variable in epidemiological studies is often a binary indicator of disease status and so the phenotype-disease relationship is typically non-linear. Instrumental variable theory has not been fully generalized to non-linear situations[6] so the practical implications of such a violation of the core assumptions have not yet been clearly defined. Most crucially, both the specification of the relevant causal parameter and identification of how it relates to what can be estimated in the observational regime are not

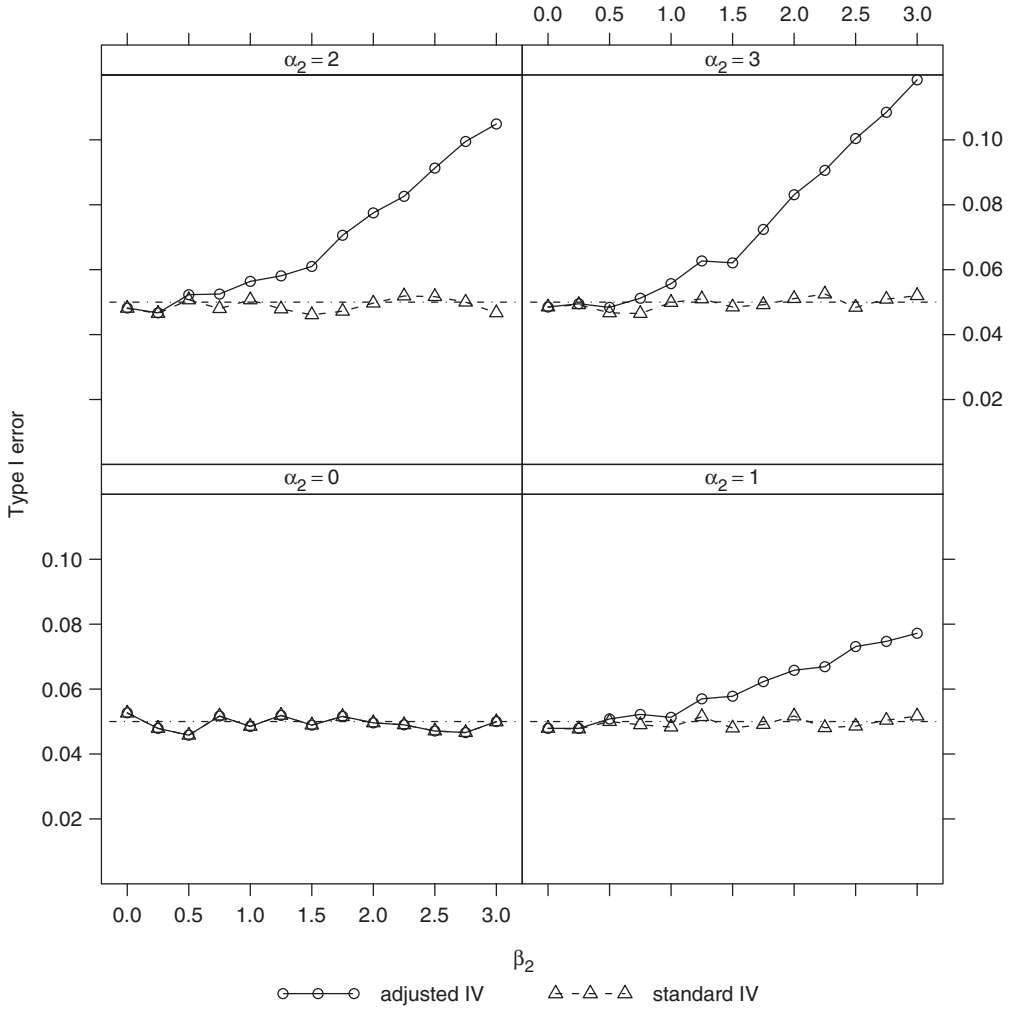



**Figure 4** Type I error rate of the Wald test for the three estimators of $\beta_1$

generally straightforward.[12] There are many examples where causal estimates have been obtained for binary outcomes but the particular parameter that can be estimated depends on the situation being considered and the assumptions that can be made.[16–22] Whilst, this is an important issue, our focus here is simply on improving the estimates of the parameter for the effect of phenotype on disease in the relevant logistic regression equation when contemporary Mendelian randomization methods are applied to binary outcome data. For now, we ignore the issue of whether, and under what conditions, this parameter has a strictly causal interpretation.

The bias associated with binary outcomes in a Mendelian randomization study may be of practical importance, so more detailed sensitivity analyses should be performed in which the biasing effects of hypothesized amounts of confounding are investigated using the formulae given in the Appendix. The three estimators considered here give different values of the phenotype-disease log odds ratio under different scenarios of confounding. The differences between the estimates are greater when the effects of the unmeasured confounders are larger. There are now several published examples of Mendelian randomization analyses, and the collection of genotype, phenotype and disease status information is becoming increasingly common, especially with the creation of large-scale Biobanks such as the UK Biobank. Large-scale collaborative genetic epidemiological studies[23,24] will ensure that there will be many genes available for use as instrumental variables in future Mendelian randomization analyses.

## Acknowledgements

TMP is funded by a Medical Research Council Capacity Building studentship in Genetic Epidemiology (G0501386). MDT is funded by a Medical Research

Council Clinician Scientist Fellowship (G0501942). The methodological research programme in Genetic Epidemiology at the University of Leicester forms one part of broader research programmes supported by: an MRC Program Grant (G0601625) addressing causal inference in Mendelian randomization; PHOEBE (Promoting Harmonization Of Epidemiological Biobanks in Europe) funded by the European Commission under Framework 6 (LSHG-CT-2006-518418); $P^3G$ (Public Population Project in Genomics) funded under an International Consortium Initiative from Genome Canada and Genome Quebec; and an MRC Cooperative Grant (G9806740). The simulation study was performed using the University of Leicester Mathematical Modelling Centre's supercomputer which was purchased through the HEFCE Science Research Investment Fund. The authors would like to thank three anonymous referees whose comments helped improve the article.

# References

[1] Katan MB. Apolipoprotein e isoforms, serum cholesterol, and cancer. *Lancet* 1986;**327:**507–8.

[2] Davey Smith G, Ebrahim S. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease. *Int J Epidemiol* 2003;**32:**1–22.

[3] Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* 2008;**27:**1133–63.

[4] Thomas DC, Conti DV. Commentary: The concept of 'mendelian randomization'. *Int J Epidemiol* 2004;**33:**21–25.

[5] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996;**91:**444–55.

[6] Pearl J. *Causality*. Cambridge: Cambridge University Press, 2000.

[7] Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;**29:**722–29.

[8] Tobin MD, Minelli C, Burton PR, Thompson JR. Commentary: Development of mendelian randomization: from hypothesis test to 'mendelian deconfounding'. *Int J Epidemiol* 2004;**33:**26–29.

[9] Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 2005;**366:**1484–98.

[10] Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988;**44:**1049–60.

[11] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993;**88:**9–25.

[12] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res* 2007;**16:**309–330.

[13] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. R Foundation for Statistical Computing, 2007. ISBN 3-900051-07-0.

[14] Nagelkerke N, Fidler V, Bernsen R, Borgdorff M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Stat Med* 2000;**19:**1849–64, [Erratum, Stat Med 2001;**20**:982].

[15] Bowden RJ, Turkington DA. *Instrumental Variables*. Cambridge: Cambridge University Press, 1984.

[16] Amemiya T. The nonlinear two-stage least-squares estimator. *J Econom* 1974;**2:**105–10.

[17] Hansen LP, Singleton RJ. Generalized instrumental variable estimation of non-linear rational expectation models. *Econometrica* 1982;**50:**1269–86.

[18] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;**14:**29–46.

[19] Robins JM, Rotnitzky A. Estimation of treatment effects in randomised trials with non-compliance and dichotomous outcomes using structural mean models. *Biometrika* 2004;**91:**763–83.

[20] Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. Limits to causal inference based on mendelian randomization: a comparison with randomized controlled trials. *Am J Epidemiol* 2006;**163:**397–403.

[21] Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006;**17:**260–67.

[22] Hernán MA, Robins JM. Instruments for causal inference. An epidemiologist's dream? *Epidemiology* 2006;**17:**360–72.

[23] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared controls. *Nature* 2007;**447**:661–78.

[24] The GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the genetic association information network. *Nat Genet* 2007; **39:**1045–51.

[25] Hardin JW, Hilbe JM. *Generalized Estimating Equations*. Boca Raton, US: Chapman and Hall/CRC, 2003.

[26] Thomas DC, Lawlor DA, Thompson JR. Re: Estimation of Bias in Nongenetic Observational Studies Using Mendelian Triangulation by Bautista *et al*. *Ann Epidemiol* 2007;**17:** 511–13.

[27] Anderson TW. *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.

# Appendix

## Formulae for the difference between the marginal and conditional parameter estimates of the three estimators

The difference between marginal and conditional parameter estimates has been investigated for the case of linear, logistic, probit and Poisson regression models.[10,25] In the case of logistic regression this difference can be expressed by a multiplicative factor,

$$\beta_{\text{marg}} \approx \beta_{\text{cond}} \cdot \frac{1}{\sqrt{1 + c^2 V}}, \text{ where } c = \frac{16\sqrt{3}}{15\pi}. \qquad (10)$$

where $\beta_{\text{marg}}$ and $\beta_{\text{cond}}$ are the marginal and conditional parameter estimates and $V$ is the variance of the

covariates over which the marginal estimates are averaged. The formulae for the three estimators are derived by approximating the logistic regression as a simple regression of the log odds ratio, $\theta = \log(p/(1-p))$ on the covariates and confounders.[26] If the terms included in the linear predictor of the logistic regression are denoted by $Z$ then the remaining variance after allowing for these terms will be given by,

$$V = \mathrm{var}(\theta|Z) = \mathrm{var}(\theta) - \frac{\mathrm{cov}(\theta, Z)^2}{\mathrm{var}(Z)} \tag{11}$$

since $\theta$ and $Z$ can both be assumed to be normally distributed.[27] From Equation (3),

$$\theta_i = \beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 g_i + (\beta_1\alpha_2 + \beta_2)u_i + \beta_1\epsilon_i \tag{12}$$

and because $u$ is standardized, it follows that

$$\mathrm{var}(\theta) = (\beta_1\alpha_1)^2\mathrm{var}(g) + (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2 \tag{13}$$

and we can approximate $\mathrm{var}(g)$ by $2q(1-q)$ where $q$ is the minor allele frequency. Hence to apply Equation (10) it is necessary to derive $V$ for each of the three estimators.

## The direct estimator

The direct estimator performs a logistic regression of disease on the intermediate phenotype. In this case $Z = x_i$ where,

$$x_i = \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i \tag{14}$$

so,

$$\mathrm{var}(Z) = \alpha_1^2\mathrm{var}(g) + \alpha_2^2 + \sigma_\epsilon^2. \tag{15}$$

The covariance between the log odds and the terms in the linear predictor is given by

$$\begin{aligned}\mathrm{cov}(\theta, Z) &= \begin{bmatrix}\alpha_1 & \alpha_2 & 1\end{bmatrix} \cdot \begin{bmatrix}\mathrm{var}(g) & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & \sigma_\epsilon^2\end{bmatrix} \cdot \begin{bmatrix}\beta_1\alpha_1\\ \beta_1\alpha_2 + \beta_2\\ \beta_1\end{bmatrix}\\ &= \alpha_1^2\beta_1\mathrm{var}(g) + \alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_\epsilon^2.\end{aligned} \tag{16}$$

Hence $V_{\mathrm{direct}}$ can be formed using Equations (13), (16) and (15).

## The standard IV estimator

For the standard IV estimator the log odds are regressed on the fitted values from the linear regression of the phenotype on the genotype. Thus $Z \approx \alpha_0 + \alpha_1 g$ and,

$$\mathrm{var}(Z) = \alpha_1^2\mathrm{var}(g), \tag{17}$$

$$\mathrm{cov}(\theta, Z) = \alpha_1^2\beta_1\mathrm{var}(g). \tag{18}$$

Hence for the standard IV estimator $V$ is given by,

$$V_{\mathrm{standard}} = (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2. \tag{19}$$

## The adjusted IV estimator

The adjusted IV estimator makes use of the estimated residuals, $r$, from the regression of the phenotype on genotype to capture some of the variance explained by confounding variables not included in the standard IV estimator. Therefore the value of $V$ is reduced compared with the standard IV estimator. For the adjusted IV estimator $V$ is given by,

$$V = \mathrm{var}(\theta|Z) - \frac{\mathrm{cov}(\theta|Z, r)^2}{\mathrm{var}(r)}. \tag{20}$$

If the confounder $u$ is standardized the estimated residuals and their variance are given by,

$$r_i = \alpha_2 u_i + \epsilon_i \tag{21}$$

$$\mathrm{var}(r_i) = \alpha_2^2 + \sigma_\epsilon^2 \tag{22}$$

The covariance between the log odds given the phenotype information and the estimated residuals is given by,

$$\mathrm{cov}(\theta|Z, r) = \begin{bmatrix}\beta_1\alpha_2 + \beta_2 & \beta_1\end{bmatrix} \cdot \begin{bmatrix}1 & 0\\ 0 & \sigma_\epsilon^2\end{bmatrix} \cdot \begin{bmatrix}\alpha_2\\ 1\end{bmatrix} \tag{23}$$

$$= \alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_\epsilon^2. \tag{24}$$

Since $\mathrm{var}(\theta|Z) = V_{\mathrm{standard}}$ from the standard IV estimator above, for the adjusted IV estimator we have,

$$V_{\mathrm{adjusted}} = (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2 - \frac{(\alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_\epsilon^2)^2}{\alpha_2^2 + \sigma_\epsilon^2}. \tag{25}$$

STATISTICS IN MEDICINE
*Statist. Med.* 2008; **27**:6570–6582
Published online 3 September 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3423

# Meta-analysis of Mendelian randomization studies incorporating all three genotypes

Tom M. Palmer[1,*,†], John R. Thompson[1] and Martin D. Tobin[1,2]

[1]*Department of Health Sciences, University of Leicester, Leicester, U.K.*
[2]*Department of Genetics, University of Leicester, Leicester, U.K.*

SUMMARY

In Mendelian randomization a carefully selected gene is used as an instrumental variable in the estimation of the association between a biological phenotype and a disease. A study using Mendelian randomization will have information on an individual's disease status, the genotype and the phenotype. The phenotype must be on the causal pathway between gene and disease for the instrumental-variable analysis to be valid. For a biallelic polymorphism there are three possible genotypes with which to compare disease risk. Existing methods select two of the three possible genotypes for use in a Mendelian randomization analysis. Multivariate meta-analysis models for Mendelian randomization case–control studies are proposed, which extend previous methods by estimating the pooled phenotype–disease association across both genotype comparisons by using the gene–disease log odds ratios and differences in mean phenotypes. The methods are illustrated using a meta-analysis of the effect of a gene related to collagen production on bone mineral density and osteoporotic fracture. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: Mendelian randomization; meta-analysis; instrumental-variable analysis

## 1. INTRODUCTION

Epidemiological studies investigating the relationship between biological risk factors and disease can be affected by confounding or reverse causation. The method known as Mendelian randomization has been proposed as a way of overcoming these difficulties [1, 2]. There has been a growing

*Correspondence to: Tom M. Palmer, Department of Health Sciences, University of Leicester, Leicester, U.K.
†E-mail: tmp8@le.ac.uk

Contract/grant sponsor: Medical Research Council capacity building Ph.D. Studentship in Genetic Epidemiology; contract/grant number: G0501386
Contract/grant sponsor: Medical Research Council Clinician Scientist Fellowship; contract/grant number: G0501942
Contract/grant sponsor: MRC Program Grant; contract/grant number: G0601625

*Received 17 July 2008*
*Accepted 22 July 2008*

interest in the application of Mendelian randomization because of the increased availability of genetic data.

Mendelian randomization analyses use an individual's genotype as an instrumental variable in order to estimate the association between a phenotype and the risk of disease. To fulfill the conditions for an instrumental variable the selected gene must be associated with the disease through the intermediate phenotype [3, 4]. The associations between the genotype and the phenotype and between the genotype and the disease should not be confounded by lifestyle or environmental factors because the genotype is assigned at conception before these exposures. As such an instrumental-variable estimate of the association between the phenotype and the disease derived from the gene–disease and gene–phenotype associations should also be free from confounding.

Statistical power can be low in individual Mendelian randomization studies, and large sample sizes are required to produce precise estimates of the phenotype–disease association [5, 6]. Therefore, it is an advantage if the genotype–disease and genotype–phenotype estimates are derived from meta-analyses.

## 2. METHODS

This section describes the information available from a case–control study and the estimation of the phenotype–disease association using Mendelian randomization. Methods are proposed for the meta-analysis of Mendelian randomization studies incorporating all three genotypes by using two genotype comparisons and an extension is given incorporating the genetic model-free approach [7, 8].

### *2.1. The ratio of coefficients approach for case–control studies*

Suppose that the genotype and phenotype information are collected in the same study. For a genetic polymorphism with two alleles, the common and risk alleles denoted by $g$ and $G$, there are three possible genotypes; the common or wild-type homozygote ($gg$), the heterozygote ($Gg$), and the mutant or uncommon homozygote ($GG$). Table I summarizes the genotype–disease and genotype–phenotype associations in a case–control study. In the table the counts of cases and controls are denoted by $n_{dj}$, subscript $d$ indicates case or control status (1 or 0) and subscript $j$ denotes the genotype (1, 2, or 3 corresponding to $gg$, $Gg$, and $GG$). The phenotype should be measured in the controls since reverse causation might affect the level of the phenotype in the cases. The observed mean phenotype levels in the controls are denoted by $\overline{x}_j$, which are estimates of the true mean phenotype levels denoted by $\mu_j$. The observed standard deviations of the phenotype levels are denoted by $\mathrm{sd}_j$. The observed mean phenotype differences between either the heterozygotes or the rare homozygotes versus the common homozygotes are given by $\widehat{\delta}_j = \overline{x}_j - \overline{x}_1$, the subscript indicates the genotype with which the common homozygotes are compared. The true genotype–phenotype mean differences are given by $\delta_j = \mu_j - \mu_1$ and the genotype–disease log odds ratios are denoted by $\theta_j$.

In an individual study if the disease status variable were a continuous outcome measure, then the application of instrumental-variable methods would produce an unbiased estimate of the phenotype–disease association, assuming that the genotype met the core conditions to qualify as an instrumental variable [9, 10]. However, case–control studies typically rely on binary disease status variables



DOI: 10.1002/sim

Table I. Data available from a Mendelian randomization case–control study.

| | Genotypes | | |
|---|---|---|---|
| | *gg* | *Gg* | *GG* |
| Number of controls | $n_{01}$ | $n_{02}$ | $n_{03}$ |
| Number of cases | $n_{11}$ | $n_{12}$ | $n_{13}$ |
| Mean phenotypes in controls (s.d.) | $\overline{x}_1$ ($\mathrm{sd}_1$) | $\overline{x}_2$ ($\mathrm{sd}_2$) | $\overline{x}_3$ ($\mathrm{sd}_3$) |

that cause the instrumental-variable methods to produce biased estimates. The proposed approach uses gene–disease and gene–phenotype log odds ratios as continuous outcome measures in order to maintain linearity between studies [11]. The instrumental-variable method known as the ratio of coefficients approach is used to estimate the phenotype–disease log odds ratio, denoted by $\eta$, using equation (1) [12, 13]. Sometimes a unit increase in the phenotype will be biologically implausible and so an arbitrary constant $k$ can be included in the ratio so that $\eta$ represents the log odds ratio associated with a $k$-unit change in the phenotype [14]:

$$\eta_{[k]} \approx \frac{k\theta}{\delta} \tag{1}$$

From the data available from a Mendelian randomization case–control study reporting all three genotypes, two non-redundant estimates of the phenotype–disease log odds ratio are possible. One estimate of $\eta$ is based on the comparison of the common homozygotes with the heterozygotes, using $\theta_2$ and $\delta_2$. The other is based on the rare homozygotes compared with the common homozygotes, using $\theta_3$ and $\delta_3$. In many situations it will be sensible to assume that the two estimates relate to a common underlying log odds ratio. In the meta-analysis model these two estimates of $\eta$ can be combined into a single, more efficient, estimate.

### 2.2. *Meta-analysis incorporating two genotype comparisons*

The meta-analysis model incorporating two genotype comparisons builds on previous meta-analysis models for Mendelian randomization studies for a single genotype comparison [13, 15]. The model relates the pooled gene–disease log odds ratios and pooled gene–phenotype mean differences using the ratio of coefficients approach from equation (1) through the mean vector of a multivariate normal distribution. The model follows multivariate meta-analysis methodology, such as [16], through the specification of the marginal distribution of the study outcome measures by combining within- and between-study variance components. The approach is the multivariate analogue of the univariate random-effects meta-analysis model of DerSimonian and Laird [17].

In the following notation subscript $i$ denotes a study. It is assumed that the observed mean phenotype differences are normally distributed such that $\widehat{\delta}_{ji} \sim \mathrm{N}(\delta_{ji},\ \mathrm{var}(\widehat{\delta}_{ji}))$ and that the true study-specific mean differences are normally distributed such that $\delta_{ji} \sim \mathrm{N}(\delta_j,\ \tau_j^2)$, where $\tau_j^2$ is the between-study variance of the true study mean differences. Then the marginal distribution of the observed mean differences is given by $\widehat{\delta}_{ji} \sim \mathrm{N}(\delta_j,\ \mathrm{var}(\widehat{\delta}_{ji}) + \tau_j^2)$. Denoting the correlation between the pooled mean phenotype differences by $\rho$, the multivariate Mendelian randomization



DOI: 10.1002/sim

(MVMR) meta-analysis model then takes the following form:

$$\begin{bmatrix} \widehat{\theta}_{2i} \\ \widehat{\delta}_{2i} \\ \widehat{\theta}_{3i} \\ \widehat{\delta}_{3i} \end{bmatrix} \sim \text{MVN}\left( \begin{bmatrix} \eta\delta_2 \\ \delta_2 \\ \eta\delta_3 \\ \delta_3 \end{bmatrix}, \mathbf{V}_i + \mathbf{B}_1 \right) \tag{2}$$

$$\mathbf{V}_i = \begin{bmatrix} \text{var}(\widehat{\theta}_{2i}) & 0 & \text{cov}(\widehat{\theta}_{2i}, \widehat{\theta}_{3i}) & 0 \\ 0 & \text{var}(\widehat{\delta}_{2i}) & 0 & \text{cov}(\widehat{\delta}_{2i}, \widehat{\delta}_{3i}) \\ \text{cov}(\widehat{\theta}_{3i}, \widehat{\theta}_{2i}) & 0 & \text{var}(\widehat{\theta}_{3i}) & 0 \\ 0 & \text{cov}(\widehat{\delta}_{3i}, \widehat{\delta}_{2i}) & 0 & \text{var}(\widehat{\delta}_{3i}) \end{bmatrix} \tag{3}$$

$$\mathbf{B}_1 = \begin{bmatrix} \tau_2^2 & \tau_2\tau_3\rho \\ \tau_2\tau_3\rho & \tau_3^2 \end{bmatrix} \otimes \begin{bmatrix} \eta^2 & \eta \\ \eta & 1 \end{bmatrix} = \begin{bmatrix} \eta^2\tau_2^2 & \eta\tau_2^2 & \eta^2\tau_2\tau_3\rho & \eta\tau_2\tau_3\rho \\ \eta\tau_2^2 & \tau_2^2 & \eta\tau_2\tau_3\rho & \tau_2\tau_3\rho \\ \eta^2\tau_2\tau_3\rho & \eta\tau_2\tau_3\rho & \eta^2\tau_3^2 & \eta\tau_3^2 \\ \eta\tau_2\tau_3\rho & \tau_2\tau_3\rho & \eta\tau_3^2 & \tau_3^2 \end{bmatrix} \tag{4}$$

The terms in the within-study covariance matrix, $\mathbf{V}_i$, are assumed to be known from the data reported by the studies and it is also assumed that there is no correlation between the gene–phenotype and gene–disease outcome measures as in [15]. From the use of the Kronecker product, it is apparent that $\mathbf{B}_1$ is singular; however, $\mathbf{V}_i + \mathbf{B}_1$ is not, which allows the calculation of the likelihood.

The parameters of this model can be estimated by maximizing the log-likelihood. For $i = 1 \ldots n$ studies $Y_i$ represents the $(4 \times 1)$ vector of outcome measures, $\beta$ represents the $(4 \times 1)$ mean vector of the multivariate normal distribution, and $\mathbf{\Sigma}_i = \mathbf{V}_i + \mathbf{B}_1$. The log-likelihood of the multivariate normal distribution up to a constant is given by

$$\sum_{i=1}^{n} -1/2\{\log(|\mathbf{\Sigma}_i|) + (Y_i - \beta)'\mathbf{\Sigma}_i^{-1}(Y_i - \beta)\} \tag{5}$$

To improve the quadratic properties of the log-likelihood the log of $\tau_2^2$ and $\tau_3^2$ and the Fisher's $z$-transform of $\rho$ were used in the maximization that was performed using the optim function in R (version 2.7.0) [18].

### *2.3. Meta-analysis incorporating the genetic model-free approach*

In the analysis of genetic association studies the mode of inheritance is usually unknown and so an assumption is made about the underlying genetic model. In contrast, the genetic model-free approach estimates this underlying genetic model from the available data through a parameter $\lambda$ [7, 8]. When $\lambda$ is equal to 0, 0.5, and 1, this represents recessive, additive, and dominant models for the risk allele, respectively.

The genetic model-free approach was devised in the context of a meta-analysis of two genotype comparisons for gene–disease outcome measures [7, 8]. A consequence of assuming that the phenotype–disease association is constant across the comparison of the heterozygotes with the



DOI: 10.1002/sim

common homozygotes and the comparison of the rare homozygotes with the common homozygotes in equation (2) is that the genetic model is assumed to be equal using either gene–disease or gene–phenotype outcomes such that

$$\lambda=\frac{\theta_2}{\theta_3}=\frac{\delta_2}{\delta_3} \tag{6}$$

The multivariate Mendelian randomization meta-analysis model incorporating the genetic model-free approach (MVMR-GMF) is given by

$$\begin{bmatrix} \widehat{\theta}_{2i} \\ \widehat{\delta}_{2i} \\ \widehat{\theta}_{3i} \\ \widehat{\delta}_{3i} \end{bmatrix} \sim \text{MVN}\left( \begin{bmatrix} \eta\lambda\delta_3 \\ \lambda\delta_3 \\ \eta\delta_3 \\ \delta_3 \end{bmatrix},\ \mathbf{V}_i+\mathbf{B}_2 \right) \tag{7}$$

$$\mathbf{B}_2=\begin{bmatrix} \lambda^2\tau_3^2 & \lambda\tau_3^2 \\ \lambda\tau_3^2 & \tau_3^2 \end{bmatrix}\otimes\begin{bmatrix} \eta^2 & \eta \\ \eta & 1 \end{bmatrix}=\begin{bmatrix} \eta^2\lambda^2\tau_3^2 & \eta\lambda^2\tau_3^2 & \eta^2\lambda\tau_3^2 & \eta\lambda\tau_3^2 \\ \eta\lambda^2\tau_3^2 & \lambda^2\tau_3^2 & \lambda\eta\tau_3^2 & \lambda\tau_3^2 \\ \eta^2\lambda\tau_3^2 & \lambda\eta\tau_3^2 & \eta^2\tau_3^2 & \eta\tau_3^2 \\ \eta\lambda\tau_3^2 & \lambda\tau_3^2 & \eta\tau_3^2 & \tau_3^2 \end{bmatrix} \tag{8}$$

Similar to the previous model $\mathbf{B}_2$ is singular but again $\mathbf{V}_i+\mathbf{B}_2$ is not. When prior knowledge about the gene suggests that $0<\lambda<1$, then the $z$-transform of $\lambda$ can be used in the maximization along with the other transformations previously described to help improve the quadratic properties of the log-likelihood. This model was also fitted by maximizing the log-likelihood in equation (5).

It is also possible to estimate the parameters of this model using Bayesian methods. One Bayesian approach known as the product normal formulation (PNF) expresses the multivariate normal distribution for each study's outcome measures as a series of univariate normal distributions linked by the relationships between the means [19] such that

$$\begin{aligned} \widehat{\theta}_{2i} &\sim \text{N}(\eta\lambda\delta_{3i}, \text{var}(\widehat{\theta}_{2i})) \\ \widehat{\delta}_{2i} &\sim \text{N}(\lambda\delta_{3i}, \text{var}(\widehat{\delta}_{2i})) \\ \widehat{\theta}_{3i} &\sim \text{N}(\eta\delta_{3i}, \text{var}(\widehat{\theta}_{3i})) \\ \widehat{\delta}_{3i} &\sim \text{N}(\delta_{3i}, \text{var}(\widehat{\delta}_{3i})) \\ \delta_{3i} &\sim \text{N}(\delta_3, \tau_3^2) \end{aligned} \tag{9}$$

The following prior distributions were assumed for the parameters to be estimated:

$$\delta_3\sim\text{N}(0, 1\times10^6),\quad \tau_3^{-2}\sim\text{Gamma}(0.1, 0.1),\quad \eta\sim\text{N}(0, 1\times10^6),\quad \lambda\sim\text{Beta}(1, 1) \tag{10}$$

The prior distributions on $\delta_3$, $\tau_3^{-2}$, and $\eta$ were chosen to be non-informative; for example, the normal prior distribution is approximately uniform over a broad range. The Beta prior distribution restricts $\lambda$ to lie between 0 and 1.



DOI: 10.1002/sim

### 2.4. Missing outcomes

In a meta-analysis it is possible that some studies may not report all four outcomes. If studies are missing either gene–disease or gene–phenotype outcome measures these studies can be included in the model fitting using the appropriate bivariate log-likelihood derived by taking the appropriate rows and columns from equations (2)–(4) or equations (7), (3), and (8). This requires the assumption that the missing outcomes are missing at random and not missing for a systematic reason.

### 2.5. Diagnostic plots

The results of a bivariate Mendelian randomization meta-analysis have been presented using a two-column forest plot instead of two separate forest plots [13, 15]. For the models presented here

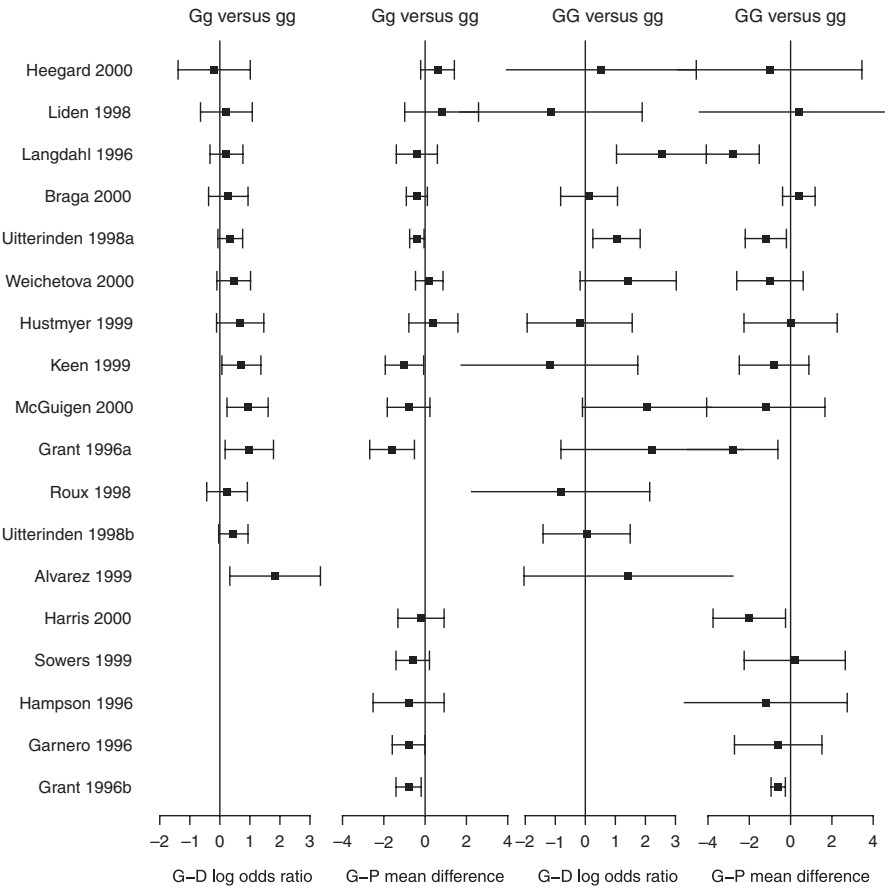



Figure 1. Four-column forest plot of the *COL1A1* multivariate meta-analysis. The genotype–phenotype (G-P) columns are on a per $0.05\,g/cm^2$ scale.



DOI: 10.1002/sim

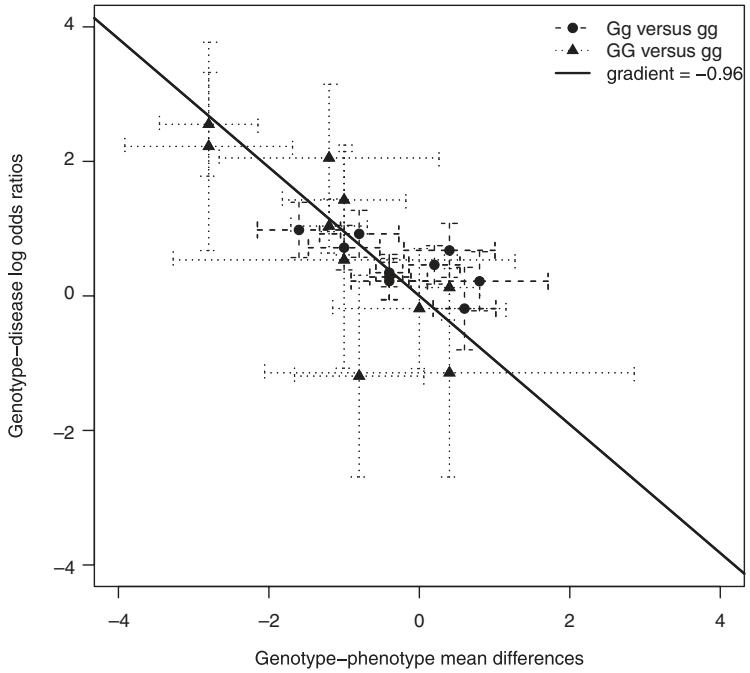



Figure 2. Gene–disease log odds ratios versus gene–phenotype mean differences (per $0.05\,g/cm^2$) plotted with one standard deviation error bars. The gradient of the line is given by $\widehat{\eta}$ from the MVMR meta-analysis model.

using four outcomes this can be extended to a four-column forest plot. To help compare the precision of the estimates, the two columns of gene–disease log odds ratios should use the same scale as should the two columns of gene–phenotype mean differences. This plot is shown in Figure 1.

In the meta-analysis models the assumption of the common phenotype–disease association in both genotype comparisons can be assessed by plotting the gene–disease outcome measures against the gene–phenotype measures [13]. From the ratio of coefficients approach, the phenotype–disease association can be expressed as the gradient of the line of best fit through the origin on this plot that is shown in Figure 2.

In the MVMR-GMF meta-analysis model the assumption that the genetic model is the same in the gene–disease and gene–phenotype outcomes can be assessed by plotting the Gg versus gg comparison against the GG versus gg comparison for each set of outcomes, respectively [7]. From the genetic model-free approach, $\lambda$ is given by the gradient of the line of best fit through the origin on these plots that are shown in Figure 3.

## 3. APPLICATION TO BONE MINERAL DENSITY AND OSTEOPOROTIC FRACTURE

A meta-analysis that investigated the relationship between a polymorphism in the COL1A1 gene and bone mineral density (BMD) and the risk of osteoporotic fracture is used to illustrate the methodology [20].



DOI: 10.1002/sim

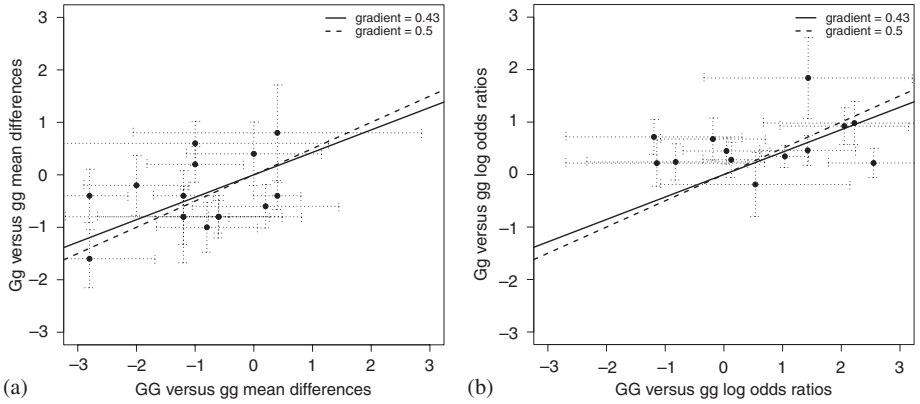



Figure 3. Graphical assessment of the estimated genetic model. The gradient of the bold lines is $\widehat{\lambda}$ from the MVMR-GMF model. A dashed line with gradient 0.5 representing the additive genetic model is also shown; lines with gradients 0 and 1 would represent the recessive and dominant genetic models, respectively: (a) genotype–phenotype information per $0.05\,g/cm^2$ and (b) genotype–disease information.

### *3.1. Description of the meta-analysis*

The *COL1A1* gene codes for one of the main forms of collagen and the Sp1 polymorphism has been shown in epidemiological studies to be associated with both BMD and the risk of fracture [21, 22]. This polymorphism is therefore a candidate for use as an instrumental variable in the estimation of the association between BMD and fracture risk. The *COL1A1* study presented two meta-analyses based on a single nucleotide, *G* to *T*, polymorphism affecting a binding site for the transcription factor Sp1 in the *COL1A1* gene. One meta-analysis investigated studies into *COL1A1* and BMD and the other meta-analysis investigated studies of *COL1A1* and osteoporotic fracture risk. It is therefore possible to apply Mendelian randomization meta-analysis to this example. The studies of the gene–phenotype and gene–disease associations should be free from confounding, whereas studies of the association of BMD with fracture may be confounded by factors such as the subject's age or the amount of exercise they take, and there may also be unknown confounders that cannot be controlled for in the analysis.

The *G* and *T* alleles of the polymorphism in the *COL1A1* gene are sometimes labelled as *S* and *s* for the common and risk alleles, respectively, but for consistency with the Methods section they are labelled as *g* and *G*. In estimating the phenotype–disease association using Mendelian randomization, a one-unit change in the phenotype can have a large impact on disease risk. In the example the standard deviation of the mean difference in BMD was $0.05\,g/cm^2$ between the homozygote genotypes and $0.03\,g/cm^2$ for comparison of the heterozygotes versus the common homozygotes. Therefore, the scaling constant, *k*, was set to 0.05 in the analysis to ensure that the pooled phenotype–disease odds ratio was estimated on an appropriate scale.

### *3.2. Results of the meta-analysis*

Figure 1 shows a four-column forest plot of the *COL1A1* meta-analyses. The first and second columns of the forest plot present the genotype–disease (G-D) and genotype–phenotype (G-P) outcomes for the *Gg* versus *gg* genotypes, while the third and fourth columns show the outcomes for the *GG* versus *gg* genotypes. The forest plot shows that there is an increased risk of fracture in



DOI: 10.1002/sim

the *Gg* over the *gg* genotype and an increased risk again in the *GG* genotype. The heterozygotes and the rare homozygotes had lower BMD than the common homozygotes. The forest plot shows that the comparison of the heterozygotes with the common homozygotes has more precise estimates because the confidence intervals around the point estimates are narrower and shows less between-study heterogeneity because the point estimates are more similar to one another.

The parameter estimates from the meta-analysis models incorporating all three genotypes are given in Table II. In the tables of parameter estimates, NA indicates a parameter that was not estimated in that particular model. The estimation of the PNF model was performed with a burn-in of 10 000 iterations followed by a chain of 50 000 iterations and MCMC convergence was assessed graphically. The estimates of $\eta$ were similar across the three models with odds ratios of osteoporotic fracture of 0.38 and 0.39 per $0.05\,g/cm^2$ increase in BMD. All three pooled odds ratios were statistically significant at the 5 per cent level. The parameters in the PNF model had wider 95 per cent credible intervals than the 95 per cent confidence intervals in the MVMR-GMF model. The estimates of $\lambda$ in the MVMR-GMF and PNF models were close to 0.5 with both 95 per cent intervals including 0.5 suggesting an additive model.

As a comparison parameter estimates from bivariate meta-analysis models similar to those considered by Thompson *et al*. [15] for the two genotype comparisons separately are given in Table III. The pooled odds ratio of fracture was 0.34 (95 per cent CI: 0.17, 0.68) per $0.05\,g/cm^2$ for the *Gg* versus *gg* comparison and 0.42 (95 per cent CI: 0.25, 0.72) for the *GG* versus *gg* comparison and the three estimates from the models in Table II are between the two values. The estimates in Table II are also more precise, as shown by the narrower confidence intervals, because of the inclusion of data for both genotype comparisons.

Parameter estimates from the bivariate meta-analysis models incorporating the genetic model-free approach using the gene–disease and gene–phenotype associations separately as in [7] are given in Table IV. The maximization of the gene–disease model failed to converge and so the between-study variance, $\tau^2_{\theta_3}$, was held constant. The fixed value of $\tau^2_{\theta_3}$ of 0.31 was taken from the univariate random-effects meta-analysis of the *GG* versus *gg* gene–disease log odds ratios. The estimate of $\lambda$ was 0.44 (95 per cent CI: 0.19, 0.64) from the gene–disease log odds ratios and 0.42 (95 per cent CI: 0.08, 0.67) from the gene–phenotype mean differences and the estimate of $\lambda$ from the MVMR-GMF model is between these two values with increased precision.

Table II. Parameter estimates for meta-analysis models using studies with complete and incomplete outcomes.

| Parameter | MVMR Estimate (95 per cent CI) ($n=18$) | MVMR-GMF Estimate (95 per cent CI) ($n=18$) | PNF Estimate (95 per cent CrI) ($n=18$) |
|---|---|---|---|
| $\eta$ | $-0.96$ $(-1.39, -0.53)$ | $-0.94$ $(-1.41, -0.47)$ | $-0.97$ $(-1.53, -0.58)$ |
| $\exp(\eta)$ | 0.38 (0.25, 0.59) | 0.39 (0.24, 0.63) | 0.38 (0.22, 0.56) |
| $\lambda$ | NA | 0.43 (0.20, 0.61) | 0.47 (0.28, 0.74) |
| $\delta_2$ | $-0.47$ $(-0.63, -0.30)$ | NA | NA |
| $\delta_3$ | $-0.85$ $(-1.35, -0.35)$ | $-0.94$ $(-1.34, -0.55)$ | $-0.92$ $(-1.44, -0.49)$ |
| $\tau^2_2$ | 0.03 (0.001, 1.17) | NA | NA |
| $\tau^2_3$ | 0.53 (0.15, 1.91) | 0.35 (0.10, 1.28) | 0.43 (0.07, 1.35) |
| $\rho$ | 0.05 $(-0.89, 0.91)$ | NA | NA |
| Log-likelihood | $-6.42$ | $-11.24$ | NA |



DOI: 10.1002/sim

Table III. Parameter estimates from bivariate Mendelian randomization meta-analysis models using studies with complete and incomplete outcomes.

| Parameter | *Gg* versus *gg* Estimate (95 per cent CI) ($n=18$) | *GG* versus *gg* Estimate (95 per cent CI) ($n=18$) |
|---|---|---|
| $\eta$ | $-1.08$ $(-1.76, -0.39)$ | $-0.86$ $(-1.39, -0.33)$ |
| $\exp(\eta)$ | 0.34 (0.17, 0.68) | 0.42 (0.25, 0.72) |
| $\delta_2$ | $-0.44$ $(-0.59, -0.28)$ | NA |
| $\delta_3$ | NA | $-0.90$ $(-1.42, -0.38)$ |
| $\tau_2^2$ | 0.02 (0.001, 2.27) | NA |
| $\tau_3^2$ | NA | 0.56 (0.16, 1.96) |

Table IV. Parameter estimates from bivariate genetic model-free meta-analysis models.

| Parameter | Gene–disease Estimate (95 per cent CI) ($n=13$) | Gene–phenotype Estimate (95 per cent CI) ($n=15$) |
|---|---|---|
| $\lambda$ | 0.44 (0.19, 0.64) | 0.42 (0.08, 0.67) |
| $\theta_3$ | 0.96 (0.50, 1.43) | NA |
| $\exp(\theta_3)$ | 2.62 (1.65, 4.16) | NA |
| $\tau_{\theta_3}^2$ | Fixed at 0.31 | NA |
| $\delta_3$ | NA | $-0.88$ $(-1.40, -0.37)$ |
| $\tau_3^2$ | NA | 0.48 (0.10, 2.31) |

Figure 2 shows the diagnostic plot to assess the pooled estimate of $\eta$ with the gene–phenotype outcome measures on the $x$-axis and the gene–disease outcome measures on the $y$-axis. Given that two genotype comparisons are assessed, each study can contribute two points to the plot. A line with gradient equal to the pooled estimate of $\eta$ is drawn on the plot to help assess the fit of the model. Only one point did not lie within one standard deviation of the fitted line. Figure 2 also shows that the point estimates from the *GG* versus *gg* comparison have greater between-study heterogeneity because the point estimates are spread over a wider range, and they are less precise than the point estimates from the *Gg* versus *gg* comparison.

Figure 3(a) and (b) assesses the estimated genetic model from the MVMR-GMF meta-analysis model. On both figures lines have been plotted with gradients equal to $\widehat{\lambda}$ from the MVMR-GMF model and 0.5 to represent the additive genetic model. For this meta-analysis these figures are sensitive to the fact that not all studies reported both sets of outcome measures and so not all studies are shown on each plot.

## 4. DISCUSSION AND CONCLUSIONS

In observational epidemiology estimates from a Mendelian randomization analysis can provide improved estimates of the association between a biological phenotype and a disease compared with direct estimates of this association. The proposed meta-analysis models extend previous literature



DOI: 10.1002/sim

by incorporating both genotype comparisons for a given genetic polymorphism into the same model. The MVMR-GMF and PNF meta-analysis models also incorporate the estimation of the underlying genetic model for the risk allele in a Mendelian randomization analysis.

The proposed meta-analysis models rely on two important assumptions, namely that the phenotype–disease association is the same in the *Gg* versus *gg* and the *GG* versus *gg* genotype comparisons and that the underlying genetic model is the same in the gene–phenotype and gene–disease associations. These assumptions are assessed in Figures 2 and 3. The modelling approach could be extended to allow the phenotype–disease log odds ratio, $\eta$, to vary across studies; this would most easily be implemented using Bayesian methodology. Figure 1 shows a four-column forest plot for a Mendelian randomization meta-analysis across two genotype comparisons. From the plot the relative precision of the estimates from the two genotype comparisons and the patterns in the estimates of individual studies can be assessed.

Incorporating multiple genotype comparisons into a Mendelian randomization analysis is advantageous because the comparison of the heterozygotes with the common homozygotes has the larger sample size, while the comparison of the rare homozygotes with the common homozygotes has the larger difference in disease risk. Therefore, the pooled estimates of the phenotype–disease association from the MVMR, MVMR-GMF, and PNF models in Table II were between the estimates for the two separate bivariate meta-analysis models using single genotype comparisons in Table III. The pooled estimate of the phenotype–disease association in the MVMR and MVMR-GMF models also showed increased precision over the single genotype comparison models because they included more information. Another advantage of incorporating all three genotypes is that if some of the studies omit to report either genotype–phenotype or genotype–disease outcome measures, then they can be accommodated in the meta-analysis model using the appropriate bivariate normal likelihood. This requires the additional assumption that the missing outcomes were missing at random and not missing for a systematic reason such as reporting bias.

The estimation of the underlying genetic model for the risk allele, known as the genetic model-free approach, can also be incorporated within this meta-analysis framework. The proposed approach extends previous literature through the joint synthesis of the genotype–disease and genotype–phenotype information to estimate the genetic model. This means that no strong assumptions about the genetic model are required prior to the analysis. In the example meta-analysis the genetic model was estimated close to the additive genetic model. The interpretation of estimates of $\lambda$ not at one of the standard genetic models has been discussed elsewhere [7].

The estimation of bivariate meta-analysis models has been shown to be problematic when correlation parameters are near $\pm 1$ [16, 23–25]. To overcome this problem an alternative form of the marginal distribution for a multivariate meta-analysis model has been proposed, which assumes a common correlation term both within and between studies; see model A in [15] or [25]. The advantage of this alternative covariance structure is that only study outcome measures and their respective variances are required to fit the multivariate meta-analysis model; the same information is required to perform the univariate meta-analyses for each outcome measure separately. A further discussion of how the relative magnitudes of the within- and between-study covariance matrices can affect parameter estimates in multivariate meta-analysis models is provided by Ishak *et al.* [26]. To fit multivariate meta-analysis models, the restricted log-likelihood could be used in the maximization as an alternative to the log-likelihood [25].

It would be possible to use these and the previously proposed bivariate meta-analysis models for Mendelian randomization studies reporting continuous disease outcome measures since the models assume that the log odds ratios are continuous and normally distributed. For case–control



DOI: 10.1002/sim

studies it would be possible to achieve similar pooled estimates of the phenotype–disease log odds ratio across two genotype comparisons using either a retrospective or a prospective likelihood for the genotype–disease outcome measures, which has previously been demonstrated for the genetic model-free approach [8]. Meta-analysis models have been used to estimate other parameters of interest from genetic data. For example, meta-regression has been used to investigate deviations from Hardy–Weinberg equilibrium [27] and merged genotype comparisons have been used to assess Hardy–Weinberg equilibrium and estimate the genetic model-free approach [28]. The work presented here also has parallels with modelling baseline risk in meta-analyses [29, 30].

The limitations that apply to the analysis of a single study using Mendelian randomization also apply to each of the studies in the meta-analysis. Therefore, it is important to assess that the selected genotype fulfills the conditions of an instrumental variable [10] and whether any of the factors that could potentially affect Mendelian randomization analyses such as pleiotropy or canalization are present [31]. Some further issues relating to the causal interpretation of meta-analyses of Mendelian randomization studies have been discussed by Nitsch *et al*. [32].

In conclusion, estimating the phenotype–disease association using separate genotype comparisons is often limited in that the comparison of the homozygote genotypes has a smaller sample size, whereas the comparison of the heterozygotes with the common homozygotes involves a smaller difference in disease risk. Pooling the phenotype–disease association across these comparisons produces an estimate that is a weighted average of the two but with increased precision. This meta-analysis framework can incorporate the estimation of the genetic model-free approach so that no strong prior assumptions about the underlying genetic model are required.

## ACKNOWLEDGEMENTS

Tom Palmer is funded by a Medical Research Council capacity building Ph.D. Studentship in Genetic Epidemiology (G0501386). Martin Tobin is funded by a Medical Research Council Clinician Scientist Fellowship (G0501942). John Thompson receives support from an MRC Program Grant (G0601625) addressing causal inference in Mendelian randomization. Tom Palmer would like to thank the ISCB subcommittee on Student Conference Awards for the receipt of a Student Conference Award to attend ISCB28. The authors would like to thank two anonymous referees whose comments greatly improved the paper.

## REFERENCES

1. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* 1986; **327**:507–508.
2. Davey Smith G, Ebrahim S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease. *International Journal of Epidemiology* 2003; **32**:1–22. DOI: 10.1093/ije/dyg070.
3. Bowden RJ, Turkington DA. *Instrumental Variables*. Cambridge University Press: Cambridge, 1984.
4. Greenland S. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* 2000; **29**:722–729.
5. Davey Smith G, Harbord R, Fibrinogen ES. C-reactive protein and coronary heart disease: does Mendelian randomization suggest the associations are non-causal? *The Quarterly Journal of Medicine* 2004; **97**:163–166.
6. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* 2008; **27**(8):1133–1163. DOI: 10.1002/sim.3034.
7. Minelli C, Thompson JR, Abrams KR, Thakkinstian A, Attia J. The choice of a genetic model in the meta-analysis of molecular association studies. *International Journal of Epidemiology* 2005; **34**:1319–1328.
8. Minelli C, Thompson JR, Abrams KR, Lambert PC. Bayesian implementation of a genetic model-free approach to the meta-analysis of genetic association studies. *Statistics in Medicine* 2005; **24**:3845–3861.

9. Greene WH. *Econometric Analysis* (4th edn). Prentice-Hall: New York, 1999.
10. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 2007; **16**:309–330.
11. Thomas DC, Lawlor DA, Thompson JR. Re: estimation of bias in nongenetic observational studies using 'Mendelian triangulation' by Bautista *et al*. *Annals of Epidemiology* 2007; **17**(7):511–513.
12. Thomas DC, Conti DV. Commentary: the concept of 'Mendelian randomization'. *International Journal of Epidemiology* 2004; **33**:21–25.
13. Minelli C, Thompson JR, Tobin MD, Abrams KR. An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. *American Journal of Epidemiology* 2004; **160**(5):445–452.
14. Thompson JR, Tobin MD, Minelli C. GE1: on the accuracy of estimates of the effect of phenotype on disease derived from Mendelian randomisation studies. *Technical Report*, University of Leicester, Leicester, 2003. Available from: http://www2.le.ac.uk/departments/health-sciences/extranet/BGE/genetic-epidemiology/genepi_tech_reports.
15. Thompson JR, Minelli C, Abrams KR, Tobin MD, Riley RD. Meta-analysis of genetic studies using Mendelian randomization—a multivariate approach. *Statistics in Medicine* 2005; **24**:2241–2254.
16. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**:589–624.
17. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.
18. R Development Core Team. *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. Available from: http://www.R-project.org. ISBN 3-900051-07-0.
19. Spiegelhalter DJ. Bayesian graphical modelling: a case-study in monitoring health outcomes. *Applied Statistics* 1998; **47**(1):115–133.
20. Mann V, Hobson EE, Li B, Stewart TL, Grant SFA, Robins SP, Aspden RM, Ralston SH. A *COL1A1* Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality. *The Journal of Clinical Investigation* 2001; **107**(7):899–907.
21. Grant SFA, Reid DM, Blake G, Herd R, Fogelman I, Ralston SH. Reduced bone density and osteoporosis associated with a polymorphic Sp 1 binding site in the collagen type I $\alpha$ 1 gene. *Nature Genetics* 1996; **14**(2):203–205.
22. Uitterlinden A, Burger H, Huang Q, Yue F, McGuigan F, Grant S, Hofman A, van Leeuwen J, Pols H, Ralston S. Relation of alleles of the collagen type I alpha 1 gene to bone density and the risk of osteoporotic fractures in postmenopausal women. *New England Journal of Medicine* 1998; **338**(15):1016–1021.
23. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**(1):78–97.
24. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**:3.
25. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* 2008; **9**(1):172–186.
26. Ishak KJ, Platt RW, Joseph L, Hanley JA. Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine* 2008; **27**:670–686.
27. Salanti G, Higgins JPT, Trikalinos TA, Ioannidis JPA. Bayesian meta-analysis and meta-regression for gene–disease associations and deviations from Hardy–Weinberg equilibrium. *Statistics in Medicine* 2007; **26**:553–567.
28. Salanti G, Higgins JPT. Meta-analysis of genetic association studies under different inheritance models using data reported as merged genotypes. *Statistics in Medicine* 2008; **27**(5):764–777. DOI: 10.1002/sim.2919.
29. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741–2758.
30. van Houwelingen HC, Senn S. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1999; **18**(1):110–115.
31. Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* 2004; **33**(1):30–42.
32. Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. Limits to causal inference based on mendelian randomization: a comparison with randomized controlled trials. *American Journal of Epidemiology* 2006; **163**:397–403.



DOI: 10.1002/sim