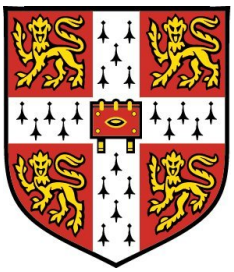
Statistical issues in Mendelian randomization: use of genetic instrumental variables for assessing causal associations



Stephen Burgess

MRC Biostatistics Unit

Emmanuel College, University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

22nd August 2011

"And furthermore, my son, be admonished: of making many books there is no end; and much study is a weariness of the flesh." Ecclesiastes 12:12.

Statement of Collaboration and Acknowledgements

I hereby declare that this thesis is the result of my own work, includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text and bibliography, is not substantially the same as any other work that I have submitted or will be submitting for a degree or diploma or other qualification at this or any other university, and does not exceed the prescribed word limit.

I would like to thank my supervisor, Simon G. Thompson, for guidance and support throughout this period of my life, and my advisors, Julian Higgins, Jack Bowden and Shaun Seaman, for helpful discussions. In particular, I acknowledge contributions from Shaun Seaman, Debbie Lawlor and Juan Pablo Casas, who are co-authors of the paper included as Appendix E, which forms the basis for Chapter 7. This paper was conceived, undertaken and written by me under the supervision of Simon Thompson, with editorial input from Shaun Seaman. Debbie Lawlor and Juan Pablo Casas provided data and comments on the manuscript.

I also acknowledge the contribution of the CRP CHD Genetics Collaboration (CCGC), specifically of Frances Wensley, who coordinated the collaboration, Mat Walker and Sarah Watson, who managed the data, and John Danesh, who oversaw the project. The papers included as Appendices B and C were conceived, undertaken and written by me under the supervision of Simon Thompson; several comments on penultimate versions of these manuscripts were provided by members of the collaboration. These papers form the basis of work in Chapters 3 and 5. The preliminary analyses undertaken in Chapter 8 were conceived, undertaken and written by me independently of the parallel analyses in the paper included as Appendix F, for which I contributed the instrumental variable analysis.

I would also like to thank my wife, Nina, all those with whom I have shared an office (Aidan, Alex, Ben, Dennis, Emma, Graham, Verena) and those who have brightened up the journey thus far, my family and friends.

Stephen Burgess: Statistical issues in Mendelian randomization: use of genetic instrumental variables for assessing causal associations

Mendelian randomization is an epidemiological method for using genetic variation to estimate the causal effect of the change in a modifiable phenotype on an outcome from observational data. A genetic variant satisfying the assumptions of an instrumental variable for the phenotype of interest can be used to divide a population into subgroups which differ systematically only in the phenotype. This gives a causal estimate which is asymptotically free of bias from confounding and reverse causation. However, the variance of the causal estimate is large compared to traditional regression methods, requiring large amounts of data and necessitating methods for efficient data synthesis. Additionally, if the association between the genetic variant and the phenotype is not strong, then the causal estimates will be biased due to the "weak instrument" in finite samples in the direction of the observational association. This bias may convince a researcher that an observed association is causal. If the causal parameter estimated is an odds ratio, then the parameter of association will differ depending on whether viewed as a population-averaged causal effect or a personal causal effect conditional on covariates.

We introduce a Bayesian framework for instrumental variable analysis, which is less susceptible to weak instrument bias than traditional two-stage methods, has correct coverage with weak instruments, and is able to efficiently combine gene–phenotype–outcome data from multiple heterogeneous sources. Methods for imputing missing genetic data are developed, allowing multiple genetic variants to be used without reduction in sample size. We focus on the question of a binary outcome, illustrating how the collapsing of the odds ratio over heterogeneous strata in the population means that the two-stage and the Bayesian methods estimate a population-averaged marginal causal effect similar to that estimated by a randomized trial, but which typically differs from the conditional effect estimated by standard regression methods. We show how these methods can be adjusted to give an estimate closer to the conditional effect.

We apply the methods and techniques discussed to data on the causal effect of C-reactive protein on fibrinogen and coronary heart disease, concluding with an overall estimate of causal association based on the totality of available data from 42 studies.

Abbreviations

2SLS	two-stage least squares
2SPS	two-stage predictor substitution
2SRI	two-stage residual inclusion
ACE	average causal effect
BMI	body mass index
CCGC	CRP CHD Genetics Collaboration
CRP	C-reactive protein
CHD	coronary heart disease
CI /CrI	confidence / credible interval
COR	causal odds ratio (Chapter 2)
C(L)OR	conditional (log) odds ratio (Chapter 4)
CRR	causal risk ratio
DIC	deviance information criterion
FE / RE	fixed-effects / random-effects
GMM	generalized method of moments
GWAS	genome-wide association study (or studies)
HDL-C	high-density lipoprotein cholesterol
HR	hazard ratio
HWE	Hardy–Weinberg equilibrium
I(L)OR	individual (log) odds ratio
IL6	interleukin-6
IPD	individual participant data
IV	instrumental variable
LIML	limited information maximum likelihood
LD	linkage disequilibrium
LDL-C	low-density lipoprotein cholesterol
lp(a)	lipoprotein(a)
MAB	median absolute bias
MAF	minor allele frequency
MAR	missing at random
MCAR	missing completely at random
MCMC	Monte Carlo Markov chain
MCSE	Monte Carlo standard error
MI	myocardial infarction
MNAR	missing not at random
M(L)OR	marginal (log) odds ratio
P(L)OR	population (log) odds ratio
RCT	randomized controlled trial
SE	standard error
(G)SMM	(generalized) structural mean model
SNP	single nucleotide polymorphism

Abbreviations for the various studies in the CCGC are given in Appendix H.

Notation

Throughout this dissertation, we use the notation:

X	phenotype: the risk factor, or protective factor, or intermediate phenotype of interest
Y	outcome
U	confounder in the X-Y association
V	unmeasured confounder (Chapter 3); covariate for Y (Chapters 4 and 6)
G	instrumental variable
α	parameter of genetic association: regression parameter in the G-X regression
β	regression parameter in the X-Y regression
β_1	causal effect of X on Y: the main parameter of interest
γ	parameter of genetic association for haplotypes: regression parameter in the G-X regression where G represents a haplotype or diplotype
ρ	correlation parameter
σ^2	variance parameter
τ^2	between-study heterogeneity variance parameter
ψ^2	genetic between-study heterogeneity variance parameter
F	F statistic from regression of X on G
i	subscript indexing individuals
j	subscript indexing genotypic subgroups
J	total number of genotypic subgroups
k	subscript indexing genetic variants (SNPs)
K	total number of genetic variants
m	subscript indexing studies in a meta-analysis, or imputed datasets (Chapter 7)
M	total number of studies, or imputed datasets (Chapter 7)
N	total number of individuals
n	total number of cases (individuals with a disease event)
t	time-to-event
$\mathcal{N}$	normal distribution
$\mathcal{U}$	uniform distribution

We follow the usual convention of using upper-case letters for random variables and lower-case letters for data values (except for N and n).

Contents

List of Tables

List of Figures

Chapter 1

Introduction to Mendelian randomization

The subject of this dissertation is statistical issues in the estimation of causal effects with genetic instrumental variables using observational data. The concept of assessing causal relations using genetic data is known as Mendelian randomization. In this introduction, we shall explore the epidemiological background of Mendelian randomization. We aim to illustrate the conceptual framework and motivation of Mendelian randomization, giving context to explain the relevance and timeliness of this dissertation. A genetic approach to epidemiology offers opportunities to deal with some of the difficulties of conventional epidemiology. We describe the specific characteristics of genetic data which give rise to this branch of epidemiology and in particular the Mendelian randomization approach, but which also lead to difficulties in statistical modelling of the resulting data from the approach. Finally, we introduce the dataset which gave rise to this PhD project and which forms the backbone of this dissertation, both illustrating and giving motivation to the findings.

1.1 The rise of genetic epidemiology

Genetic epidemiology is the study of the role of genetic factors in health and disease (1). We sketch the history and development of genetic epidemiology, giving a background and motivation as to why it is an important area of epidemiological and scientific research. A brief glossary of genetic terminology, reproduced from Lawlor et al. (2) is provided as Table 1.1. Similar glossaries can be found in (3) and (4).

- *Alleles* are the variant forms of a single-nucleotide polymorphism (SNP), a specific polymorphic site or a whole gene detectable at a locus.
- *Canalization* [also known as *developmental compensation*] is the process by which potentially disruptive influences on normal development from genetic (and environmental) variations are damped or buffered by compensatory developmental processes.
- A *chromosome* carries a collection of genes located on a long string of DNA. A non-homologous chromosome carries a unique collection of genes on a long string of DNA that is different from the gene collection of another non-homologue. Normal non-homologous chromosomes are not attached to each other during meiosis, and move independently of one another, each carrying its own gene collection. Two homologous chromosomes carry the same collection of genes, but each gene can be represented by a different allele on the two homologues (a heterozygous individual). A gamete will receive one of those homologues, but not both. Humans have 22 pairs of autosomal homologous chromosomes and 1 pair of sex chromosomes.
- *DNA* – deoxyribonucleic acid is a molecule that contains the genetic instructions used in the development and functioning of all living organisms. The main role of DNA is the long-term storage of information. It contains the instructions needed to construct other components of cells, including proteins and ribonucleic acid (RNA) molecules. DNA has four nucleotide bases A, T, G and C. The two strands of DNA in the double-helix structure are complementary (sense and anti-sense strands) such that A binds with T and G binds with C.
- A *gene* comprises a DNA sequence, including introns, exons and regulatory regions, related to transcription of a given RNA.
- [The] *genotype* of an individual refers to the two alleles inherited at a specific locus - if the alleles are the same, the genotype is homozygous, if different, heterozygous.
- [A] *haplotype* describes the particular combination of alleles from linked loci found on a single chromosome.
- *Linkage disequilibrium* (LD) is the correlation between allelic states at different loci within the population. The term LD describes a state that represents a departure from the hypothetical situation in which all loci exhibit complete independence (linkage equilibrium).
- A *locus* is the position in a DNA sequence and can be a SNP, a large region of DNA sequence, or a whole gene.
- *Pleiotropy* is the potential for polymorphisms to have more than one specific phenotypic effect
- *Polymorphism* is the existence of two or more variants (i.e. SNPs, specific polymorphic sites or whole genes) at a locus. Polymorphism is usually restricted to moderately common genetic variants, at least two alleles with frequencies of greater than 1 per cent in the population.
- *Recombination* is any process that generates new gene or chromosomal combinations not found previously in that cell or its progenitors. During meiosis, recombination is the process that generates haploid cells that have non-parental combinations of genes.
- *Single-nucleotide polymorphism[s]* (SNPs) are genetic variations in which one base in the DNA is altered, e.g. a T instead of an A.

Table 1.1: A glossary of genetic terminology, reproduced with permission from Lawlor et al. (2) with minor edits marked in square brackets

1.1.1 Historical background

The concept of inherited characteristics goes back to the dawn of time, although the mechanism for inheritance was long unknown[1]. When Charles Darwin proposed his theory of evolution in 1859 (6), one of its major problems was the lack of an underlying mechanism for heredity (7). Grigor Mendel in 1866 proposed two laws of inheritance: the law of segregation, that when any individual produces gametes (sex cells), the copies of a gene separate so that each gamete receives only one copy; and the law of independent assortment, that "unlinked or distantly linked segregating gene pairs assort independently at meiosis" (8). These laws are summarized by the term "Mendelian inheritance", and it is this which gives Mendelian randomization its name, specifically due to the second law, the law of 'independent assortment' (3). The two areas of evolution and Mendelian inheritance were brought together through the 1910s-30s in the "modern evolutionary synthesis", by amongst others Ronald Fisher, who helped to develop population genetics (9). The link between genetics and disease was established by Linus Pauling in 1949, who linked a specific genetic mutation in patients with sickle-cell anaemia to a demonstrated change in the haemoglobin of the red-blood cells of affected individuals (10). The discovery of the structure of deoxyribonucleic acid (DNA) in 1953 gave rise to the birth of molecular biology, which led to greater understanding of the genetic code (11). The Human Genome Project was established in 1990, leading to the publication of the entirety of the human genetic code by 2003 (12; 13). Recently, technological advances have reduced the cost of DNA sequencing to the level where it is now economically viable to measure genetic information for a large number of individuals (14).

1.1.2 Shortcomings of classical epidemiology

Epidemiology is the study of patterns of health and disease at the population level. We use the term 'classical epidemiology' meaning epidemiology without the use of genetic factors, to contrast with genetic epidemiology. A fundamental problem in epidemiological research, in common with other areas of social science, is the distinction between correlation and causation. If we want to address basic medical questions, such as to determine disease aetiology (that is, what is the cause of a disease?), to assess the impact of a public health intervention (that is, what would be the result of a change in treatment?), to inform public policy, to prioritize healthcare resources, to advise treatment practice, or to counsel on the impact of lifestyle choices, then we have to answer questions of cause and effect. The

[1]For example, Genesis 5:3 reads "When Adam had lived 130 years, he had a son in his own likeness, in his own image" (5).

optimal way to address these questions is by appropriate study design, such as the use of randomized trials and prospective data (15). However, such designs are not always possible, and often causal questions must be answered using only observational data.

Unfortunately, interpreting the association between a risk factor and a disease outcome in observational data as a causal association relies on untestable and often implausible assumptions. This has led to several high-profile cases where a risk factor has been widely advocated as an important factor in disease prevention from observational data, only to be later discredited when the evidence from randomized trials did not support a causal interpretation to the findings (16). For example, observational studies reported a strong inverse association between vitamin C and coronary heart disease (CHD), which did not attenuate on adjustment for a variety of risk factors (17). However, results of experimental data obtained from randomized trials showed a null association with a positive point estimate for the association (18). The confidence intervals for the observational and experimental associations did not overlap (3). Similar stories apply to the observational and experimental associations between β-carotene and smoking-related cancers (19; 20), and vitamin E and CHD (21). More worrying is the history of hormone-replacement therapy, which was previously advocated as beneficial for the reduction of breast cancer and cardiovascular mortality on the basis of observational data, but was subsequently shown to increase mortality in randomized trials (22; 23).

1.1.3 The need for an alternative

As the knowledge of the human genome developed, the search for genetic determinants of disease expanded from monogenetic disorders (that is, disorders which are due to a single mutated gene), such as sickle-cell anaemia (cited above), to polygenic and multifactorial disorders, where the burden of disease risk is not due to a single gene, but to multiple genes combined with lifestyle and environmental factors. These diseases, such as cancers, diabetes and CHD, tend to cluster within families, but also depend on other factors, such as diet or blood pressure. Several genetic factors have been found which relate to these diseases, especially through the increased use of whole-genome scans known as genome-wide association studies (GWAS). However, this is of limited interest from a clinical point-of-view, as an individual's genome cannot be changed. We here present an introduction to Mendelian randomization: a method for using genetic data to estimate causal associations of modifiable (non-genetic) risk factors using observational data.

1.2 What is Mendelian randomization?

Mendelian randomization is here defined as the use of non-experimental studies to determine the causal effect of a phenotype on an outcome by making use of genetic variation. We shall use the word "phenotype" to refer to the putative causal risk factor, which can be thought of as an exposure, a biomarker or any other risk factor which may affect the outcome (24). Usually the outcome is disease, although there is no methodological restriction as to what outcomes can be considered. Non-experimental studies encompass all observational studies, including cross-sectional, cohort and case-control designs, where there is no intervention instituted by the researcher. These are contrasted with clinical trials.

1.2.1 Motivation

A foundational aim of epidemiological enquiry is the estimation of the effect of changing one risk factor on an outcome (3). This is known as the causal effect of the phenotype on the outcome and typically differs from the observational association between phenotype and outcome (25), due to endogeneity of the phenotype (26). Endogeneity, literally "coming from within", of a variable in an equation means that there is a correlation between the variable and the error term, and occurs when the variable is predicted by the terms in the model in which it appears (27). For example, those who regularly take headache tablets are likely to have more headaches than those who do not, but taking headache tablets is unlikely to be a cause of the increased incidence of headaches. Taking tablets is an endogenous variable in this context, and so the causal effect of taking tablets on headaches cannot be estimated from this observational setting. The opposite of endogenous is exogenous; an exogenous variable comes from outside of the model and is not explained by the terms in the model.

The idea of Mendelian randomization is to find an exogenous genetic variant (or variants) which is associated with the phenotype, but is not associated with any other risk factor which affects the outcome, and is not directly associated with the outcome, in that any impact of the genetic variant on the outcome must come via its association with the phenotype (2). These assumptions define an instrumental variable (IV) (28; 29). As IVs were initially developed for use in the field of economics, a number of terms commonly used in the IV literature derive from this field and are not always well understood by statisticians or epidemiologists. Table 1.2 is a glossary of terms which are commonly used in each field.

Economics term	Statistics term	Notes
Endogenous / endogeneity	Confounded / confounding	Traditionally, confounding refers to a narrower circumstance than endogeneity. A 'confounder' (denoted U) has been defined as a variable which is associated with the risk factor of interest and the outcome. However, it has been shown that it is possible for a variable to be a 'confounder' without biasing causal effects. Endogeneity means that there is a correlation between the regressor and the error term in an equation. A better definition for confounding would be a bias in the estimation of a causal effect, which corresponds with the definition of endogeneity. This definition includes phenomena which are traditionally thought of as separate from confounding, such as measurement error and reverse causation.
Exogenous / exogeneity	Unconfounded / No confounding	
Regressor	Covariate	Any term in a regression equation
Outcome	Outcome	Denoted Y in this text
Endogenous/ exogenous regressor	Confounded/ unconfounded variable	Denoted X in this text; if endogenous, the causal effect of X on Y cannot be estimated by OLS of Y on X
Instrumental variable / Excluded instrument	Instrumental variable / Instrument	Denoted G in this text; the instrument is called 'excluded' as it is not included in the second-stage of the two-stage regression method often used for calculating IV estimates
Included regressor	Measured covariate	A covariate which is included in a model, such as a multivariate regression
OLS	Least-squares regression	OLS stands for Ordinary Least Squares. The OLS estimate is the observational association, as opposed to the IV estimate, which is the causal association.
Concentrate out	Profile out	To exclude a nuisance parameter from an equation by forming a profile likelihood by replacing with its maximum likelihood estimate given the other variables
Panel data	Longitudinal data	Data on multiple items at multiple timepoints. Panel data can include time-series (single item) and cross-sectional (single timepoint) data, neither of which is generally thought of as longitudinal data.

Table 1.2: A dictionary of instrumental variable terms used in the economics and statistics fields

1.2.2 Instrumental variables

An alternative definition of Mendelian randomization is "instrumental variable analysis using genetic instruments" (30; 31). While not all Mendelian randomization studies have used IV methodology (32), the use of genetic variants as IVs is at the core of Mendelian randomization (33).

An IV is an exogenous variable associated with an endogenous exposure which is used to estimate the causal effect of changing the exposure while keeping all other factors equal (25; 34). In the language of Mendelian randomization, the genetic variant(s) are considered as IVs for the causal association of phenotype on outcome (35). The fundamental conditions for an IV to satisfy are summarized as (2; 28; 33):

i. the IV is associated with the phenotype,

ii. the IV is not associated with any confounder,

iii. the IV is conditionally independent of the outcome given the phenotype and confounders.

The use of a particular genetic variant as an IV is controversial as these assumptions cannot be fully tested and may be violated for various epidemiological and biological reasons (2; 33; 36; 37; 38). A British study into the distribution of genetic markers and non-genetic factors (such as environmental exposures) in a group of blood donors and a representative sample from the population showed marked differences in the non-genetic factors, but no more difference than would be expected by chance in the genetic factors (37), indicating that genetic factors seem to be distributed independently of possible confounders in the population of the United Kingdom (39). This gives plausibility to the general suitability of genetic variants as IVs, but in each specific case, justification of the assumptions relies on biological knowledge about the genetic markers in question.

As a plausible example of a valid genetic IV, in the Japanese population, a common genetic mutation in the ALDH2 gene affects the processing of alcohol, causing excess production of a carcinogenic by-product, acetaldehyde, as well as nausea and headaches. We can use this genetic variant as an instrumental variable to assess the causal association between alcohol consumption and oesophageal cancer. Here, alcohol consumption is the phenotype and oesophageal cancer the outcome. Assessing the causal association here using observational data is complicated by the strong association between alcohol and tobacco smoking, another risk factor for oesophageal cancer (40). Individuals with two copies of the mutation tend to avoid alcohol, due to the severity of the short-term symptoms. Their risk of developing oesophageal cancer is one-third of the risk of those with no

copies of the mutation (41). Carriers of a single copy of this mutation exhibit only a mild intolerance to alcohol. They are still able to drink, but they cannot process the alcohol efficiently and have an increased exposure to acetaldehyde. Carriers of a single copy are at three times the risk of developing oesophageal cancer compared to those without the mutation, with up to 12 times the risk in studies of heavy drinkers (41). There is no link between having this genetic mutation and many other risk factors.

The genetic mutation provides a fair test to compare three populations who differ systematically only in their consumption of alcohol and exposure to acetaldehyde, and who have vastly differing risks. The evidence for a causal link between alcohol consumption, exposure to acetaldehyde and oesophageal cancer is compelling (42). In this example, a further natural experiment can be exploited: women in Japanese culture tend not to drink for social reasons. A similar study into alcohol and blood pressure showed a significant association between ALDH2 and blood pressure for men, but not for women (43). This provides further evidence that the change in outcome is not due to the genetic variant itself, but due to the effect of the phenotype. This strengthens our belief that the genetic variant is a valid IV, and the change in outcome is causally due to alcohol consumption via exposure to acetaldehyde, not due to the violation of the IV assumptions, such as the correlation of the IV with another risk factor.

In the above example, we used Mendelian randomization to assess the causal nature of the phenotype-outcome association. There are several reasons why it is desirable to go beyond testing for a causal effect and to estimate the size of the causal effect. Firstly, this is usually the parameter representing the answer to the question of interest (24). Secondly, with multiple genetic variants, greater power can be achieved. If several independent IVs all show a concordant causal effect, the overall estimate of causal effect using all the IVs may give statistical significance even if none of the individual IVs does (44; 45). Thirdly, often a null association is expected (40). By estimating a confidence interval for the causal effect, we obtain bounds on the plausible size of any causal association. Although it is not statistically possible to prove the null hypothesis, we can reduce the plausible causal effect to one which is of no clinical relevance.

1.2.3 Analogy with randomized controlled trials

Mendelian randomization is analogous to a randomized controlled trial (RCT) (46; 47; 48). A RCT, considered the "gold standard" of medical evidence (32), involves dividing a target population into two or more subgroups in a random way. These subgroups are each given different treatment programmes. Randomization is preferred over any other assignment to

subgroups as all possible confounders, known and unknown, are on average balanced (49). However in many situations, for ethical or practical reasons, it is not possible to intervene on the factor of interest to estimate the causal effect by direct experiment (40).

In Mendelian randomization, we use the IV to form subgroups analogous to those in a RCT, as shown in Figure 1.1. From the IV assumptions, these subgroups differ systematically in the phenotype, but not in any other factor (50). A difference in disease incidence between these subgroups would therefore indicate a true causal relationship between phenotype and outcome (51).

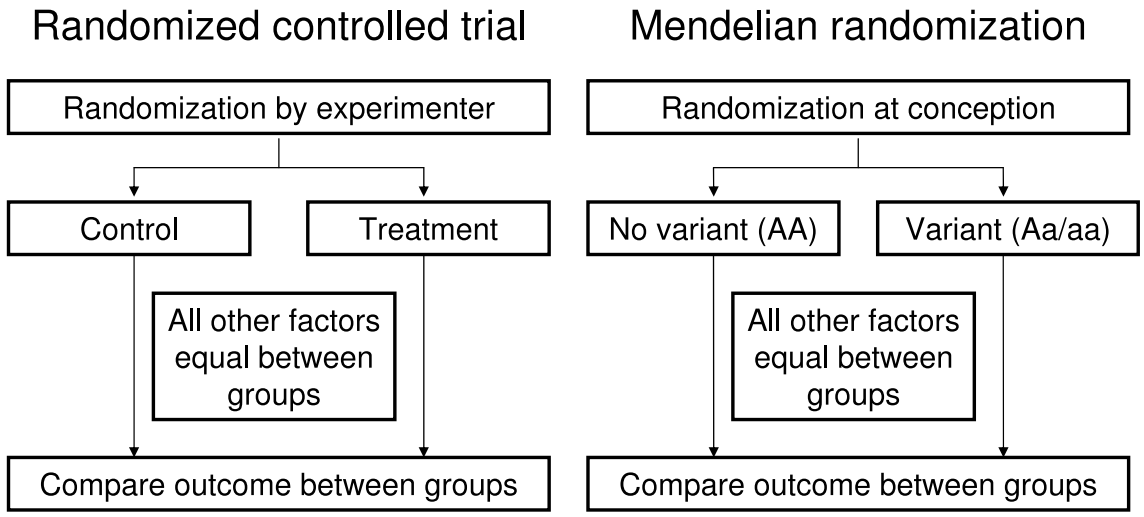


Figure 1.1: Comparison of randomized controlled trial and Mendelian randomization (adapted from (46))

However, Mendelian randomization is subtly different from a randomized trial. The aim of Mendelian randomization is not to estimate the size of a genetic effect, but the causal effect of the phenotype on the outcome. When the proportion of variation in the phenotype associated with the genetic variant is not large or is imprecisely estimated, studies will require large sample sizes (42), such as 10 000 or even 30 000 cases (3; 40), as the risk ratio from the difference in phenotype due to the genetic variant may be low (52). However, the population attributable risk of the phenotype is not necessarily low (40). Although the variation in phenotype attributable to the gene may be small, it can be similar to that attributable to treatment in a RCT (53).

1.2.4 Confounding

Mendelian randomization has also been named 'Mendelian deconfounding' (54) as it aims to give estimates of causal association free from bias associated with confounding.

The correlations between risk factors make it impossible in an observational study to look at the increase in one variable keeping all others equal, as changes in one factor will always be accompanied by changes in other factors (47). While we can measure individual confounders and adjust for them in our analysis, we can never be certain that all risk factors have been identified or measured precisely. This leads to what is known as unmeasured confounding (55). Additionally, if we adjust for a variable that lies on the true causal pathway between the phenotype of interest and outcome, this represents an over-adjustment and attenuates the causal association (56). By finding a genetic marker which satisfies the IV assumptions, we can estimate the unconfounded association between the genetic marker and outcome (33).

1.2.5 Reverse causation

Mendelian randomization also deals with problems of reverse causation (40). Reverse causation occurs when an association between the phenotype and outcome is not due to the phenotype causing a change in outcome, but outcome causing a change in phenotype. This could happen, for example, if the phenotype increases in response to pre-clinical disease (24). If the genetic variant is a valid IV, any difference in outcome between individuals in the genetically-defined subgroups is due to the genetic variant. As the genotype was determined at conception and cannot be changed, there is no possibility of reverse causation (50).

1.3 Genetic markers

Generally in Mendelian randomization, genetic markers used as IVs are in the form of single nucleotide polymorphisms (SNPs) (2; 57; 58; 59). As summarized in Table 1.1, a SNP is defined as a variation in the deoxyribonucleic acid (DNA) of an individual compared to the population at a single point (or locus), where one nucleotide, either A, C, G or T, has been replaced with another. These different variants in the genetic code are called alleles. Where there are two possible alleles at a particular locus (a diallelic SNP), we write the more common allele, the major allele or wildtype as A and the less common allele, the minor allele or variant as a. The proportion of minor alleles in a population is

called the 'minor allele frequency'. An arbitrary threshold of the minor allele frequency is set at 1%, below which a SNP is considered a mutation rather than a polymorphism.

As people have two copies of each DNA sequence, individuals can be categorized for each diallelic SNP into three possible subgroups corresponding to three combinations of alleles. These subgroups are named major homozygotes (AA), heterozygotes (Aa) and minor homozygotes (aa). We shall denote these subgroups as 0, 1 and 2, corresponding to the number of minor alleles for that SNP. For this reason, a diallelic SNP is usually considered to be a discrete random variable taking values from {0, 1, 2}. For a more complicated genetic instrument, such as a triallelic SNP where there are three possible alleles at one locus, there is no natural ordering of the six possible subgroups given by the SNP. A triallelic SNP can be considered as either an unordered categorical random variable or a discrete random variable using the average phenotype levels as an ordering.

Genetic sequences can be combined into haplotypes, which can then be used as IVs (2). A haplotype is a combination of alleles, one from each SNP measured, which are inherited together. Humans have two haplotypes at each locus, one from each parent. When SNPs are inherited together, usually due to physically proximity on the same chromosome, haplotypes can be inferred from SNP data using computer software as generally not all possible combinations of SNP alleles will be present in a population. In some cases, haplotypes can be determined uniquely from SNP data, whereas in others, there is uncertainty in this determination. If the SNPs satisfy the IV assumptions, then the haplotypes will also satisfy the IV assumptions.

1.4 Examples of Mendelian randomization

Mendelian randomization has been used in applied studies for a number of different contexts. A systematic review of applied Mendelian randomization studies was published by Bochud and Rousson in 2010 (60) and a list of the phenotypes and outcomes of some causal associations which have been assessed using Mendelian randomization is given in Table 1.3. The list includes the fields of epidemiology, nutrition, sociology, and economics. In summary, the only limitation in the use of Mendelian randomization to assess the causal effect of a phenotype on an outcome is the availability of a suitable genetic variant to use as the IV (2; 40).

Nature of phenotype	Phenotype	Outcome	Reference
Biomarker	CRP	insulin resistance	(61)
	CRP	CIMT	(62)
	CRP	CHD	(63; 64)
	CRP	cancer	(65)
	homocysteine	stroke	(66)
	SHBG	CHD	(67)
	lp(a)	MI	(68)
	HDL-C	MI	(69)
	APOE	cancer	(70)
	folate	blood pressure	(71)
Physical characteristic	BMI	CIMT	(72)
	BMI	blood pressure	(73)
	BMI	early menarche	(74)
	fat mass	academic achievement	(75)
Dietary factor	alcohol intake	oesophageal cancer	(76)
	alcohol intake	blood pressure	(43)
	milk intake	metabolic syndrome	(77)
	caffeine intake	stillbirth	(78)
Pathological behaviour	alcohol abuse	drug abuse	(79)
	ADHD	education	(80)
	depression	education	(80)
Inter-generational effects	interuterine folate	NTD	(24; 81)

Table 1.3: Examples of causal associations assessed by Mendelian randomization in applied research (a systematic list can be found in (60)). Acronyms: CRP = C-reactive protein, SHBG = sex-hormone binding globulin, lp(a) = lipoprotein(a), HDL-C = high-density lipoprotein cholesterol, APOE = apolipoprotein E, BMI = body mass index, ADHD = attention deficit hyperactivity disorder; CIMT = carotid intima-media thickness, CHD = coronary heart disease, MI = myocardial infarction, NTD = neural tube defects

1.5 The CRP CHD Genetic Collaboration dataset

This dissertation is motivated by data on C-reactive protein (CRP) and coronary heart disease (CHD) collected by the CRP CHD Genetics Collaboration (CCGC) (82).

The CCGC is a collaboration of 47 epidemiological studies seeking to ascertain the causal role of CRP on CHD using a Mendelian randomization approach. CRP is an acute-phase protein found in the blood which is associated with inflammation. It is known that CRP is observationally associated with CHD (83; 84), but it is not known whether this association is causal (85; 86; 87; 88). Studies from the collaboration measure CRP levels, genes relating to CRP, and CHD events. We use the term 'prevalent' to refer to a CHD event prior to blood draw for CRP measurement and 'incident' to refer to a CHD event subsequent to blood draw. Individual participant data (IPD) have been collected by the coordinating centre. In this dissertation, we restrict attention to participants of European descent, excluding the four studies with no European descent participants from analysis. This is to ensure greater homogeneity of the study populations and to prevent violations of the IV assumptions due to population stratification (40).

Table 1.4 lists the the major statistical features of the studies of the CCGC. Further epidemiological characterization of the studies can be found in Appendix 1 of the published paper from the collaboration (64), which is reproduced in this dissertation as Appendix G. General features of the studies can be found in Appendix G, Table B. Study acronyms are given in Appendix H. We discuss below issues relating to the phenotype, outcome, genetic instruments and study design which are relevant to the methods developed in this dissertation.

1.5.1 Study design

The collaboration includes prospective studies: cohort studies, case-cohort studies, nested case-control studies (both matched and unmatched); and retrospective studies: case-control studies (unmatched). In some prospective studies, CRP measurements have not been taken at recruitment, but rather at a later occasion, which we have defined as our baseline. Hence, some of the individuals who had incident events in the original study will have prevalent events in the baseline-transformed study. Four of the studies in the collaboration did not provide IPD but only summary data on numbers of individuals with and without CHD events for each genotype.

Study [2]	Study type	Total participants	Number of subjects with... Incident CHD	Prevalent CHD	CRP data [3]	SNP data [1] g1	g2	g3	g4
BRHS	Cohort with prevalent cases	3824	379	151	3516	✓	✓	✓	✓
BWHHS	Cohort with prevalent cases	3771	43	236	2970	✓	✓	✓	
CCHS	Cohort with prevalent cases	10259	680	241	9503	✓	✓		✓
CGPS	Cohort with prevalent cases	32038	188	899	30491	✓	✓		✓
CHS	Cohort with prevalent cases	4511	793	447	4051	✓	✓	✓	
EAS	Cohort with prevalent cases	907	61	28	644	✓	✓		✓
ELSA	Cohort with prevalent cases	5496	71	241	4504	✓	✓		✓
FRAMOFF	Cohort with prevalent cases	1680	46	81	1479	✓	✓	✓	✓
PROSPER	Cohort with prevalent cases	5777	476	768	4876	✓	✓	✓	
ROTT	Cohort with prevalent cases	5406	259	614	4524	✓	✓		✓
NPHSII	Cohort without prevalent cases	2282	99		2158	✓	✓		✓
WOSCOPS	Cohort without prevalent cases	1451	279		1334		✓		
EPICNOR	Nested matched case-control	3298	1074		2126	✓	✓		✓
HPFS	Nested matched case-control	737	200		403	✓	✓	✓	✓
NHS	Nested matched case-control	684	196		387	✓	✓	✓	✓
NSC	Nested matched case-control	1673	577		969	✓	✓	✓	✓
CAPS	Nested unmatched case-control	1157	198		783		✓	✓	✓
DDDD	Nested unmatched case-control	897	269		614	✓	✓	✓	✓
EPICNL	Nested unmatched case-control	3478	426		3215	✓	✓	✓	✓
WHIOS	Nested unmatched case-control	3756	1339		1725	✓	✓	✓	✓
MALMO	Nested unmatched case-control with prevalent cases	2148	530	398	139	✓	✓	✓	✓
SPEED	Nested unmatched case-control with prevalent cases	854	71	19	564	✓	✓	✓	✓
ARIC	Unmatched case-control [4]	2261	615	17	859	✓	✓		✓
CUDAS	Unmatched case-control	1107		56	983	✓	✓		
CUPID	Unmatched case-control	555		340	193	✓	✓		
HIFMECH	Unmatched case-control	1006		490	495		✓		
HIMS	Unmatched case-control	3946		522	3077	✓	✓	✓	
ISIS	Unmatched case-control	3618		2075	1258	(see Section 1.5.3)			
LURIC	Unmatched case-control	2747		1137	1599	✓	✓	✓	✓
PROCARDIS	Unmatched case-control	6464		3126	3302	✓	✓	✓	✓
SHEEP	Unmatched case-control	2671		1113	1083	✓	✓		✓
WHITE2	Unmatched case-control [5]	5515		31	4800	✓	✓		✓
CIHDS	Unmatched case-control (CRP in controls only)	6716		2236	4415	✓	✓		✓
CHAOS	Unmatched case-control (no CRP data)	2475		623	0	✓	✓		✓
FHSGRACE	Unmatched case-control (no CRP data)	4548		2146	0	✓	✓	✓	✓
GISSI	Unmatched case-control (no CRP data)	4034		3054	0	✓	✓	✓	✓
HVHS	Unmatched case-control (no CRP data)	4407		1040	0	✓	✓	✓	✓
INTHEART	Unmatched case-control (no CRP data)	4188		1883	0	✓	✓	✓	✓
UCP	Unmatched case-control (no CRP data)	2011		922	0	✓	✓	✓	✓
AGES	Tabular data	3219		800	0	✓	✓	✓	✓
HEALTHABC	Tabular data	1660		584	0	✓	✓	✓	✓
MONAKORA	Tabular data	1675		272	0	✓	✓	✓	✓
PENNCATH	Tabular data	1509		1022	0	✓	✓	✓	✓
Total		162416	8392	28089	103039				

Table 1.4: Summary of studies from the CRP CHD Genetics Collaboration with subjects of European descent

[1] g1 = rs1205, g2 = rs1130864, g3 = rs1800947, g4 = rs3093077 or equivalent proxies (see Section 1.5.3).

[2] A list of study abbreviations is given in Appendix H

[3] In retrospective case-control studies, CRP data is taken in controls only; in prospective studies, from subjects without prevalent CHD.

[4] Although ARIC was reported as a case-cohort study, one of the SNPs (rs2794521) was measured only on a subsample of the population containing a disproportionate number of individuals with CHD events, inducing an association between the SNP and CHD status. The study was analysed as a case-control study with sample restricted to those with a measured value of rs2794521.

[5] Although WHITE2 was reported as a cohort study, no incident events were reported and so it has been analysed as a case-control study.

1.5.2 Phenotype data

The phenotype CRP was measured throughout using a high-sensitivity assay. Some of the studies do not measure CRP level for all individuals, and others do not measure it for any individuals. In prospective cohort studies where individuals with a CHD event at baseline were not excluded from the study due to the study design, CRP measurements for individuals with prevalent CHD were excluded from analysis. In nested (prospective) case-control studies, blood was drawn and stored at baseline, to enable pre-CHD event measurement of CRP. In retrospective case-control studies, CRP measurements for cases were excluded from analysis, as they were measured after the CHD event, to prevent bias in the causal effect due to reverse causation. In both nested and retrospective case-control studies, preferential selection of diseased individuals into the study population induces an association between the IV and the outcome, known as selection bias, hence inference on CRP is taken only on the controls, as they form a more representative sample of the population as a whole (89). Table 1.4 lists the number of individuals in each study with a CRP measurement suitable for use in the IV analysis according to the criteria above. Further details on the measurement and storage of CRP can be found in Appendix G, Table C.

1.5.3 Genetic data

The 43 studies in the collaboration with European descent participants measure different genetic information in the form of SNPs in the CRP gene region. SNPs measured which lie outside the CRP gene region were discarded due to potential violation of the IV assumptions. This gene region is on chromosome 1 and is responsible for regulation of CRP. The number of SNPs measured in each study varied from 1 to 13. Over 20 SNPs in total were measured in at least one study. Four SNPs were pre-specified in the study protocol (82) as the instruments to be used in the analysis: rs1205, rs1130864, rs1800947 and rs3093077. These four SNPs show varying degrees of correlation and give rise to five haplotypes (Table 1.5) which comprise 99% of the variation exhibited in European descent populations (82). Over 99% of individuals in the CCGC had a genotype which was compatible with these haplotypes. Only 6 studies measure all four of the pre-specified SNPs. Some studies measure SNPs which are in complete linkage disequilibrium (LD) with one of the pre-specified SNPs, and which can be used as proxies for these SNPs (90). 20 measure all four SNPs or proxies thereof and an additional 17 measure some three out of these four. Five of the remaining studies considered measure fewer than this, and the final study ISIS measures no SNPs which correspond to any of these four.

Haplotype	rs1205 (g1)	rs1130864 (g2)	rs1800947 (g3)	rs3093077 (g4)
1	C	T	G	T
2	C	C	G	T
3	C	C	G	G
4	T	C	G	T
5	T	C	C	T

Table 1.5: Haplotypes in the CRP gene region tagged by four pre-specified SNPs

We use proxy rs1417938 which is in complete LD with rs1130864, and proxies rs3093068 and rs12068753 which are in complete LD with rs3093077. For studies FHSGRACE and INTERHEART, we use proxy rs2794521 in place of rs3093077, which alongside the other pre-specified SNPs tags the same 5 haplotypes as the pre-specified SNPs, as noted in the protocol paper (82). For study ARIC, we use SNPs rs2794521 in place of rs3093077 and the triallelic SNP rs3091244, which tags both SNPs rs1205 and rs1130864. For study ISIS, we used SNP rs2808628, which is in the CRP gene region but is not a proxy of any of the pre-specified SNPs. We were able to verify the stated LD relations in the SeattleSNP database (http://pga.gs.washington.edu [checked 01/12/09]), and in the SNAP database (http://www.broadinstitute.org/mpg/snap/ [checked 01/06/10]) (90), and to assess the correlation of these SNPs in studies from the collaboration measuring both the pre-specified and proxy SNP, where we saw almost complete LD. Throughout this dissertation in the text and in all graphs and tables, proxy SNPs are included as if they are the SNP of interest. We denote rs1205 (or proxies thereof) as g1, rs1130864 (or proxies thereof) as g2, rs1800947 (or proxies thereof) as g3, and rs3093077 (or proxies thereof) as g4.

There was some sporadic missingness in the genetic data in most of the studies, although this was rarely greater than 10% missingness per SNP and usually much less. Table 1.5 lists the pre-specified SNPs measured in each study. Further details on the measurement and storage of the genetic material can be found in Appendix G, Table D.

1.5.4 Outcome data

The outcome CHD was defined as fatal coronary heart disease (based on International Classification of Diseases codings) or nonfatal myocardial infarction (using World Health Organization criteria). In five studies, coronary stenosis (more than 50% narrowing of at least one coronary artery assessed by angiography) was also included as a disease outcome. Only the first CHD event was included in analysis; an individual could not contribute

more than one event to the analysis. We consider either a binary (all studies) or a survival outcome (cohort studies). Further details on the classification of disease in each study can be found in Appendix G, Table E.

1.5.5 Covariate data

Data on various covariates were measured by the individual studies, including physical variables such as body mass index (BMI), systolic and diastolic blood pressure; lipid measurements, such as total cholesterol, high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C), triglycerides, apolipoprotein A1, and apolipoprotein B; and inflammation markers, such as lipoprotein(a), interleukin 6, and fibrinogen. Graphs of the associations between the SNPs and covariates can be found in Figure 1 of the published paper from the collaboration (64), reproduced in Appendix F, and p-values for the correlation of haplotypes and SNPs with certain covariates can be found in Appendix G, Tables F and H. These show strong associations of CRP with each of the SNPs ($p < 10^{-30}$ for each of the four SNPs), but no more significant associations with any covariate than would be expected by chance (Figure 1, Appendix F: out of 84 tested associations between a covariate and SNP, one had $p < 0.01$ ($p = 0.003$ for association between height and rs1205), and three had $p \leq 0.05$). We conclude that the SNPs appear to be valid IVs for CRP.

Of particular interest is fibrinogen, a soluble blood plasma glycoprotein, which enables blood-clotting and is also associated with inflammation. In this dissertation in addition to the causal CRP-CHD association, we consider Mendelian randomization analysis of the causal association of CRP on fibrinogen. We use fibrinogen as an outcome for several reasons. Firstly, as a continuous variable, it is more convenient to use fibrinogen to demonstrate methods for IV analysis than CHD, a binary or survival outcome (91). There are specific difficulties in IV methods for analysis of binary outcomes which we shall discuss at length later, but which are avoided by the use of a continuous outcome. Secondly, the causal association of CRP on fibrinogen is of interest in its own right. The pathway of inflammation is not well understood, but is important as both CRP and fibrinogen are risk factors for coronary heart disease (CHD). Although CRP is associated with CHD risk, this association reduces on adjustment for various risk factors, and attenuates to near null on adjustment for fibrinogen (84). It is important therefore to assess whether CRP is causally associated with fibrinogen, since if so conditioning the CRP-CHD association on fibrinogen would represent an over-adjustment, which would attenuate a true causal association.

1.5.6 The need for Mendelian randomization

CRP is observationally associated with many known covariates which are also risk factors for CHD (Appendix G, Table G). Although adjustment for these covariates is possible and can be shown to reduce the association of CRP with CHD to be compatible with no association (Appendix G, Table I), such adjustment is controversial as the causal pathway is unknown, and so it is unclear which covariates should and should not be adjusted for in analysis. Mendelian randomization is able to answer the question of causal association without making assumptions about covariates, except that they are not associated with the SNPs used as IVs.

1.5.7 Statistical issues and difficulties in CCGC

The differences between the studies in the CCGC lead to difficulties in evidence synthesis and possible statistical heterogeneity in causal estimates from each of the studies.

1. **Study design:** The parameter usually estimated in a cohort study is typically a hazard ratio, which differs from the odds ratio estimated in a case-control study. In a matched case-control study, a conditional odds ratio is estimated, which differs from an unconditional odds ratio estimated in an unmatched case-control study.

2. **Phenotype data:** Where individuals in a study do not have a phenotype value due to sporadic missingness, the phenotype can be imputed from its conditional distribution in the analysis model. However, it is unclear how to include data from studies where no CRP data was measured.

3. **Genetic data:** Where studies have measured the same SNPs, it is possible to combine the information on the association between the genes and the phenotype across studies in addition to combining the information on the causal association of the phenotype on outcome. This should gain precision in estimation of the causal effect. However, when different studies measure different SNPs, some of which may be common, it is unclear how to combine the information on the genetic association.

4. **Outcome data:** Some studies include individuals with both prevalent and incident CHD. It is unclear how to include all of the CHD events from these studies without including CRP data on individuals twice. It is not clear how to include survival and binary outcomes in the same analysis model.

Additionally, there are the problems of weak instruments and missing data. A 'weak instrument' is defined as an IV for which the statistical evidence of association with the phenotype is not strong (2). An instrument can be weak if it explains a small amount of the variation of the phenotype, where the amount defined as 'small' depends on the sample size (44). Weak instruments give rise to estimates of causal association which may be biased (92).

Although missing data is a problem which is not unique to Mendelian randomization, missing genetic data represents a specific problem for such analyses. Mendelian randomization studies often have limited power, and so excluding participants due to the presence of missing data is not ideal if they provide information on the causal effect. Conversely, sample sizes are often large, and so a 10% gain in efficiency may correspond to a large absolute gain in sample size. Additionally, if there are multiple genetic variants which can be used as IVs, the aim would be to include all available genetic information, but not to exclude participants with missing data on some of the available IVs. Rather than compromising between maximizing genetic information or sample size, an efficient analysis would be able to include all participants regardless of which data were available.

Finally, the estimate from the Mendelian randomization analysis represents the answer the causal question: "for an intervention elevating CRP across their whole life, what would be the impact of an increase in CRP on CHD risk?". This raises issues due to the statistical difficulty of expression and interpretation of risk in a heterogenous population in the absence of knowledge of covariates, known as the problem of collapsibility. For certain measures of association, termed 'non-collapsible', the estimate of risk differs depending on whether it is considered for an individual or for the population as a whole.

1.6 Overview of dissertation

The structure of the dissertation is as follows. The central thesis is that the Bayesian framework presented provides a flexible framework for estimating causal effects using instrumental variables in a variety of circumstances. Following a review of the existing literature for statistical issues relating to Mendelian randomization, we highlight two specific problems of weak instrument bias and non-collapsibility. Weak instrument bias is a bias caused by failure of the assumption of no association between instrument and confounders being violated in finite samples. Non-collapsibility is the failure of an estimator to average correctly across a confounding distribution, causing the estimator to be different when considered conditionally on levels of the confounder, and when considered for the population as a whole. We investigate how the Bayesian framework introduced in the

thesis and other IV estimators behave in terms of bias and coverage in weak and strong instrument scenarios with continuous and binary outcomes. The problem of missing data is considered, with methods presented in a Bayesian framework to impute sporadic missing genetic data. The methods and observations of the previous chapters are used to analyse causal associations using data from the CCGC. Finally, we summarize our conclusions and make suggestions for future work.

1.6.1 Chapter structure

Chapter 2 comprises a literature review of statistical methodology for Mendelian randomization. The focus of the review is on methods for IV analysis and issues associated with estimating causal effects.

Chapter 3 illustrates, explains and estimates the impact of bias from weak instruments, and discusses how bias can be minimized in analysis and design of Mendelian randomization studies.

Chapter 4 shows how non-collapsibility of the odds ratio results in a difference between the marginal and conditional odds ratio. In instrumental variable analysis, where adjustment for confounders is not necessary to prevent bias by confounding, it is not clear what the target parameter for inference is. The findings of Chapters 3 and 4 are demonstrated by the use of simulation and real data.

Chapter 5 presents a Bayesian framework motivated by the issues of Chapters 3 and 4, as well as the research question posed by the data from the CCGC, involving the meta-analysis of individual patient data from several sources using different genetic instrumental variables and a variety of study designs.

Chapter 6 investigates the issues of bias and coverage in the analysis of continuous and binary outcomes. We show that a simple modification to the Bayesian method with continuous outcomes is analogous to a control variable approach with binary outcomes. We see how this reduces bias from weak instruments, changes the target causal parameter in a binary setting and avoids the need for asymptotic distributional assumptions on the causal parameter.

Chapter 7 introduces the problem of missing data. In a Bayesian setting, missing data can be naturally imputed where the distribution of the variable with missing data is defined in the model. However, it is not clear how to interpret the distribution of genetic data, which are often highly correlated due to the underlying biological processes of genetic inheritance. We present four methods to incorporate individuals with missing data into a Bayesian analysis.

Chapter 8 represents both the inspiration and culmination of the dissertation, as we show how the issues of the previous chapters are relevant to the research question of causal association of CRP on both fibrinogen and CHD. We analyse several different designs of study, showing how, under certain assumptions, the information on causal parameters from each of the studies in the collaboration can be combined using a single hierarchical model.

Chapter 9 comprises a discussion of the dissertation as a whole, giving conclusions, critical commentary on the limitations of the work presented, and possible directions for future work.

1.6.2 Novelty and publications

Although the issues of weak instrument bias and non-collapsibility (Chapters 3 and 4) are known in the contexts of econometrics and causal analysis, they have not received attention in the context of Mendelian randomization (2; 33). We provide insights into both issues with novel explanations of the phenomena and simulations to demonstrate how they relate to Mendelian randomization. We conclude each chapter with practical advice on the impact of the theoretical results on applied research. Papers published on the material presented in this dissertation on weak instrument bias are included in the dissertation as Appendices A and B. Although Bayesian estimation using IVs has been proposed elsewhere, the Bayesian framework of Chapter 5 is novel, as is the work in Chapter 6 on the properties of the Bayesian and other IV methods with continuous and binary outcomes. A paper published on the Bayesian framework is included as Appendix C, and a submitted paper on the properties of the IV methods as Appendix D. The work on missing data (Chapter 7) is novel (45); an accepted paper on the missing data methods is included as Appendix E. The methods developed for the CCGC applied analysis (Chapter 8) contain several novel components, such as use of haplotypes for studies measuring different numbers of SNPs and inclusion of studies without phenotype measurements. The applied CCGC paper is included as Appendix F, with detailed tables published as eAppendix 1 to the applied paper included as Appendix G, a list of study abbreviations and names published as eAppendix 2 included as Appendix H, and a précis of the statistical methods detailed in Chapter 8 published as eAppendix 5 included as Appendix I.

Chapter 2

Existing statistical methods for Mendelian randomization

This chapter comprises a review of the existing literature on statistical issues relating to Mendelian randomization. The scope of this literature review is to discuss methods for Mendelian randomization, with emphasis on statistical practice. Although specific issues in instrumental variable (IV) analysis which are relevant to Mendelian randomization will be discussed, IV analysis will not be reviewed exhaustively. Instrumental variables methods have been the subject of econometric research and practice for over 80 years (28; 93), and so a comprehensive treatment is impractical; here we focus on the issues of bias in finite samples (usually called "weak instrument bias") and estimation of causal effects with binary outcomes.

2.1 Review strategy

Papers have been searched for online using Google and Google Scholar search engines, the search databases PubMed and Web of Science, and the search facilities in the journals Statistics in Medicine, International Journal of Epidemiology, American Journal of Epidemiology, Statistical Methods in Medical Research and the Stata Journal. Terms searched for were: Mendelian randomiz(s)ation, instrumental variable(s), weak instrument. PubMed reported 127 hits for the search string "Mendelian randomization", Web of Science 352 and Google Scholar 1700. PubMed reported 335 hits for the string "instrumental variables", Web of Science 2237 and Google Scholar 74 900 (correct on 25/1/11). Papers were ranked by number of citations and date of publication, and the higher ranking and more epidemiologically relevant papers were read preferentially when the number of papers found was high. Relevant papers were found from the references of other papers

read. Abstracts were read to search for methodological papers preferentially over applied papers, although some applied papers have been included in the review.

2.2 Finding a valid instrumental variable

As has been stated in Section 1.2.2, in order for a genetic marker to be used to estimate a causal effect, it must satisfy the assumptions of an instrumental variable.

We assume that we have an outcome Y which is thought of as a function of a phenotype X and confounder U. We consider that the confounding factors can be summarized by a single random variable U (94), which satisfies the requirements of a sufficient covariate (95). A sufficient covariate is a covariate which, if known and conditioned on, would give an estimate of association equal to the causal association. As U is unlikely to be dominated by just a few confounding factors, ability to reduce the confounding factors to a univariate random variable seems a reasonable assumption. If we consider confounders $U_1, \dots U_p$ which are linearly related and normally distributed, then we can scale X and Y to replace these U_j with a single U with a standard normal distribution. We assume that the phenotype X can be expressed as a function of the confounder U and the genetic marker G. G may be a single genetic variant or a matrix corresponding to several independent genetic variants. G is assumed to satisfy the IV assumptions of Section 1.2.2, rewritten here in terms of random variables:

i. G is not independent of X ($G \not\perp\!\!\!\perp X$),

ii. G is independent of U ($G \perp\!\!\!\perp U$),

iii. G is independent of Y conditional on X and U ($G \perp\!\!\!\perp Y|X, U$).

This means that the joint distribution of Y, X, U, G, $p(y, x, u, g)$ factorizes as

$$p(y, x, u, g) = p(y|u, x)p(x|u, g)p(u)p(g) \tag{2.1}$$

which corresponds to the directed acyclic graph (DAG) Figure 2.1 (33; 95).

In the "potential outcomes" or counterfactual causal framework, a set of outcomes $Y(x), x \in X$ are considered to exist, where $Y(x)$ is the outcome which would be observed if the phenotype were set to $X = x$. At most one of these outcomes is ever observed (96). The causal assumptions encoded in the DAG (Figure 2.1) can be expressed in the language of potential outcomes as follows (25):

i'. $p(x|g)$ is a non-trivial function of g

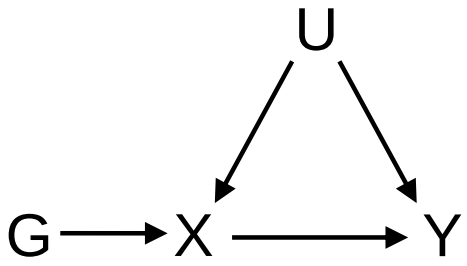


Figure 2.1: Directed acyclic graph (DAG) of Mendelian randomization assumptions

ii'. $\mathbb{E}(Y(x)|G) = \mathbb{E}(Y(x))$

iii'. $Y(x, g) = Y(x)$

where $Y(x, g)$ is the potential outcome which would be observed if X were set to x and G were set to g. Assumption ii'. is named 'conditional mean independence' and states that the mean value of the outcome for each phenotype value does not depend on the IV. This would not be true if, for example, the IV were associated with a confounder U. Assumption iii'. is named 'exclusion restriction' and states that the observed outcome for each value of the phenotype is the same for each possible value of the IV. This means that the IV can only affect the outcome through its association with the phenotype (97). We use the notation $do(X = x)$ to denote setting the value of X to x independent of confounders (98). We note that $\mathbb{E}(Y|X = x) \neq \mathbb{E}(Y|do(X = x))$ in general, for example due to confounding.

In order to interpret the unconfounded estimates produced by IV analysis as causal estimates, we require the additional structural assumption:

$$p(y, u, g, x|do(X = x_0)) = p(y|u, x_0)1(X = x_0)p(u)p(g) \tag{2.2}$$

where 1(.) is the indicator function. This ensures that intervening on X does not affect the distributions of any other variables except the conditional distribution of Y (99).

2.2.1 Parallel with non-compliance

An area in biostatistics where IVs are widely used is the adjustment of randomized trial results for non-compliance (25; 28). Non-compliance refers to the failure of participants in a clinical trial to adhere to a specified treatment regime. In this case, the IV is treatment assignment and the phenotype is treatment as received. Generally, treatment assignment is associated with treatment as received (assumption i.); treatment is assigned at random, so is independent of confounders (assumption ii.); and treatment assignment has

no direct effect on outcome and will be independent of outcome conditional on treatment received and confounders (assumption iii.). An intention to treat (ITT) analysis considers the difference in outcome between treatment groups as assigned. This answers the causal question: "how much does an individual benefit from being assigned to a treatment group?". An IV analysis considers the causal difference in outcome due to treatment, and answers the question: "how much does an individual benefit from receiving treatment?" (29).

Although there are important parallels between Mendelian randomization and non-compliance analyses, there are also several differences. The allocated and received treatment in a randomized trial are usually dichotomous, and there is usually a strong association between the two, with the majority of participants following their treatment regime. In Mendelian randomization, the genotype is discrete, but generally polychotomous, and phenotype is generally continuous. The proportion of variation in the phenotype explained by the IV may be as small as 1% or less (100). In adjustment for non-compliance, the IV is randomly allocated and so independence of the IV and confounders is automatic; in Mendelian randomization, this requires biological knowledge of the genetic variant.

2.2.2 Violations of the IV assumptions

The IV assumptions can be violated in several ways. We here distinguish between finite-sample violations and asymptotic violations. If the confounder is continuous, then the correlation between the genotype and confounder in any given dataset is almost surely different from zero, even when G and U are uncorrelated as random variables. We term this a "finite-sample violation" (101; 102) and do not regard this as invalidating an IV. However, there may be an underlying correlation structure in the random variables G, X, Y, U which is considered a violation of the IV assumptions. This may be due to biological factors, epidemiological factors, or genetic factors. These have been well-documented (33; 36; 38; 40; 103) and here we consider only statistical criteria for validity of the IV.

We note that the assumption of association between G and X does not preclude a non-causal interpretation to this association (33). Indeed, if G is not a functional variant of X, but is correlated with a functional variant, then it may still be a valid IV (51). Such correlation is known as linkage disequilibrium (LD). However, if there is any association between G and an alternative risk factor, either through pleiotropy (multiple function of one gene) (104), LD with another functional variant, population substructure (for example stratification due to ethnic heterogeneity) (35), developmental compensation (the genetic

effect on phenotype is dampened or buffered by another biological process) (3), or epigenetics (genetic effects other than those coded by DNA) (105), then the IV is not valid. The causal estimate based on this IV will be biased, although if the association with the phenotype is not strong, then the bias may not be large (34).

Typically, the IV assumptions cannot be tested, as the set of all confounders is unknown (33). Under certain assumptions, when multiple valid IVs are available, an overidentification test can help detect violations (Section 2.11.2). When a specific confounder U_p is known, the G-U_p association can be tested empirically (3). Throughout this dissertation, unless explicitly stated otherwise, we assume that the genetic instruments used are valid IVs.

2.3 Testing for a causal effect

Mendelian randomization studies address two related questions (33): whether there is a causal link between the phenotype and disease (4; 58), and what is the size of the causal effect (2; 54).

Under the assumption that the IV is valid, a valid test for the presence of a causal association of X on Y is to test for independence of G and Y, where a significant association between G and Y is indicative of a causal association (33; 51). However the converse is not true, as there may be zero correlation between G and Y without independence. This is known as the non-faithfulness of a DAG (106).

2.4 Estimating the causal effect

Although testing the causal effect is useful, it is more useful to estimate the magnitude of the causal effect. Issues relating to estimation of this causal effect will be the main focus of this literature review and dissertation as a whole. In this section, we list some of the general issues associated with parameter estimation: the assumptions necessary to estimate a causal effect, definitions of the causal parameters to be estimated, and collapsibility, which refers to the behaviour of a parameter when marginalized or averaged across a distribution. Having discussed these issues, we proceed to consider methods for constructing different IV estimators.

2.4.1 Additional IV assumptions

In order to estimate the causal effect, it is necessary to make further assumptions to the ones listed in Section 2.2. General assumptions often thought of as core assumptions include the ignorability of the selection mechanism of G (107), which means that G is assigned randomly, and the stable unit treatment value assumption (SUTVA) (96), which states that the outcome for one individual should be unaffected by variables in the model relating to the other individuals (108).

For several of our models, a specific structural form is assumed for the joint distribution of Y, X, U, G. Commonly assumed forms include the linear model, log-linear model, and the logistic model. In the linear model, we assume that, for each individual i where $i = 1, \ldots, N$, the phenotype x_i is a linear function of the instruments g_{ik} for $k = 1, \ldots, K$, the confounder u_i and an error term ϵ_{xi}. We generally assume that each instrument takes a fixed number of discrete values, usually either two or three ($g_{ik} \in \{0, 1\}$ or $g_{ik} \in \{0, 1, 2\}$). The instruments partition the population into genotypic subgroups indexed by j, with each subgroup containing all individuals with a particular genotype. The outcome y_i is assumed to be a linear function of the phenotype, confounder and an independent error term ϵ_{yi}:

$$x_i = \alpha_0 + \sum_k \alpha_{1k} g_{ik} + \alpha_2 u_i + \epsilon_{xi} \tag{2.3}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 u_i + \epsilon_{yi} \tag{2.4}$$

In the log-linear or logistic model, we assume that for each individual i the probability of event p_i is log-linear or logistic in the phenotype and confounder:

$$f(p_i) = \beta_0 + \beta_1 x_i + \beta_2 u_i$$
$$y_i \sim \text{Binomial}(1, p_i)$$

where $f(.)$ is the the logarithm function for a log relative risk model or the logistic function for a log odds ratio model. With a single instrument g_i, we omit the second subscript k. We identify β_1 as our causal effect of interest.

2.4.2 Causal parameters

Generally, the desired causal parameter of interest is that which corresponds to a population-based intervention, equivalent to a randomized controlled trial (RCT) (109).

The average causal effect (ACE) (33) under intervention in X is the expected difference in Y when the phenotype is set to two different values:

$$ACE(x_0, x_1) = \mathbb{E}(Y|do(X = x_1)) - \mathbb{E}(Y|do(X = x_0)) \tag{2.5}$$

The ACE is zero when there is conditional independence between Y and X given U, but the converse is not generally true, due to possible non-faithfulness (33).

With a binary outcome, the ACE is also called the causal risk difference. However, it is often more natural to consider a causal risk ratio (CRR) or causal odds ratio (COR):

$$CRR(x_0, x_1) = \frac{\mathbb{E}(Y|do(X = x_1))}{\mathbb{E}(Y|do(X = x_0))} \tag{2.6}$$

$$COR(x_0, x_1) = \frac{\mathbb{P}(Y = 1|do(X = x_1))\mathbb{P}(Y = 0|do(X = x_0))}{\mathbb{P}(Y = 1|do(X = x_0))\mathbb{P}(Y = 0|do(X = x_1))} \tag{2.7}$$

2.4.3 Collapsibility

A measure of association is said to be collapsible over a variable if it is constant across the strata of the variable, and if this constant value equals the value obtained from the marginal analyses (110). In a log-linear model, the relative risk is collapsible over a confounder U since

$$\begin{aligned} \mathbb{E}(Y|do(X = x)) &= \int \exp(\beta_0 + \beta_1\, x + \beta_2\, u)p(u)du \\ &= \exp(\beta_0^* + \beta_1\, x) \end{aligned} \tag{2.8}$$

with $\beta_0 \neq \beta_0^*$ but with the same relative risk β_1, where $p(u)$ is the marginal distribution of the confounder U.

In a logistic model, the odds ratio is non-collapsible, as it differs depending on the distribution of confounders (33). This is because, in general,

$$\begin{aligned} \mathbb{E}(Y|do(X = x)) &= \int \text{expit}(\alpha + \beta_1\, x + \beta_2\, u)p(u)du \\ &\neq \text{expit}(\alpha^* + \beta_1\, x) \end{aligned} \tag{2.9}$$

where expit is the inverse of the logistic function. This means that the COR will be different considered conditionally or marginally on U. Collapsibility is an important consideration in Mendelian randomization, as the set of confounders are typically unknown. The impact of the non-collapsibility of the COR will be discussed further in Chapter 4.

2.5 Ratio of coefficients method

Over the next sections, we discuss methods for IV estimation with both continuous and binary outcomes. We explain for each method how to estimate a causal association, and describe specific properties of the estimator. In turn, we consider the ratio of coefficients method, two-stage methods, likelihood-based methods, semi-parametric methods, and a method due to Greenland and Longnecker. We proceed to compare and contrast the estimators.

The ratio of coefficients method, or the Wald method (111), is the simplest way of estimating the causal association β_1 of X on Y. For a dichotomous IV $G = 0, 1$ and a continuous outcome, it is calculated as the ratio of the difference in the average outcomes to the difference in the average phenotype levels between the two IV groups (34; 112).

$$\hat{\beta}_1^R = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0} \tag{2.10}$$

where $\bar{y}_j$ for $j = 0, 1$ is the average value of outcome for all individuals with genotype $G = j$, and $\bar{x}_j$ is defined similarly for the phenotype. This estimator is valid under the assumption of monotonicity of G on X and linearity of the causal association with no (X, U) interaction (75; 99; 112). Monotonicity means that the average phenotype for each individual would be increased (or equivalently for each individual would be decreased) if that person had $G = 1$ compared to if they had $G = 0$.

With a binary outcome, the estimator is defined similarly, with $\bar{y}_j$ the log of the probability of an event in a log-linear model or the log odds of an event in a logistic model. This is also commonly quoted in its exponentiated form as $\exp(\hat{\beta}_1^R) = R^{1/\Delta x}$ where R is the relative risk or odds-ratio and $\Delta x = \bar{x}_1 - \bar{x}_0$ is the average difference in phenotype between the two groups (54; 71). This estimator is valid under the assumption of monotonicity of G on X and a log-linear or logistic model of disease on phenotype with no (X, U) interaction (99).

For a polytomous or continuous IV, the estimator is calculated as the ratio of the regression coefficient of outcome on IV (G-Y regression) to the regression coefficient of phenotype on IV (G-X regression) (2; 28).

$$\hat{\beta}_1^R = \hat{\beta}_{GY} / \hat{\beta}_{GX} \tag{2.11}$$

With a continuous outcome, the G-Y regression uses a linear model; with a binary outcome, a linear model may be used (34), although a log-linear or logistic regression is preferred.

For linear models, then this estimator is valid when

$$E(Y) = \beta_1 X + h(U). \tag{2.12}$$

For a log-linear model, where f is the log function, then this estimator is valid when

$$E(f(Y)) = \beta_1 X + h(U) \tag{2.13}$$

where $h(U)$ is an arbitrary function of U (99; 112). However, with a logistic model where f is the logit function, the ratio estimator $\hat{\beta}_1^R$ does not consistently estimate the coefficient β_1 (94; 99).

The ratio estimator can be intuitively motivated: the increase in Y for a unit increase in G ($\hat{\beta}_{GY}$) can be estimated as the product of the increase in X for a unit increase in G ($\hat{\beta}_{GX}$) and the increase in Y for a unit increase in X ($\hat{\beta}_1^R$) (2). For this reason, for continuous outcomes it has been called the linear IV average effect estimator (LIVAE) (99).

The ratio method uses a single IV. If more than one instrument is used then the causal estimates for each IV can be calculated separately. A bound on the size of the causal parameter may be calculated when the associations are non-linear (33; 113). The ratio estimator has no finite moments (101).

2.5.1 Confidence intervals

If the regression coefficients $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$ are assumed to be normal, critical values and confidence intervals for the estimator may be calculated using Fieller's Theorem (2; 114). For this, we need the correlation between $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$, which is generally assumed to be zero (89; 100). There are three possible forms of this confidence interval (115):

i. The interval may be a closed interval [a, b],

ii. The interval may be the complement of a closed interval $(-\infty, a] \cup [b, \infty)$,

iii. The interval may be unbounded.

The interpretation of the second interval is, for example, that the confidence interval for the ratio of the normal variables when viewed as a gradient on a graph of Y on X includes the vertical line (i.e. infinite ratio) but excludes the horizontal line (i.e. zero ratio). The interpretation of the third interval is that the confidence interval for the ratio of the normal variables when viewed as a gradient on a graph cannot exclude any interval of values. These unbounded confidence intervals occur because there is a non-zero

probability that the denominator term in the ratio may be close to zero. The confidence interval is more likely to be a closed interval if we have a "strong" instrument, that is an instrument with a large association with the phenotype.

Alternatively, asymptotically correct confidence intervals can be estimated using a Taylor expansion (116).

2.6 Two-stage methods

A two-stage method comprises two regression stages: the first-stage regression of the phenotype on the genetic IVs, and the second-stage regression of the outcome on the fitted values of the phenotype from the first stage. It is not a likelihood-based method, as the two stages are performed separately with no feedback from the second stage into the first.

2.6.1 Continuous outcome - two-stage least squares

With continuous outcomes and a linear model, the two-stage method is known as two-stage least squares (2SLS), or in some econometrics circles simply as the IV estimator (117). It can be used with multiple continuous or categorical IVs. The method is so called because it can be calculated using two regression stages (93). The first stage (G-X regression) regresses X on G to give fitted values $\hat{X}|G$. The second stage (X-Y regression) regresses Y on the fitted values $\hat{X}|G$ from the first stage regression. The causal estimate is this second-stage regression coefficient for the change in outcome caused by unit change in the phenotype.

Although estimation in two stages gives the correct point estimate, the standard error is not correct; the use of 2SLS software is recommended for estimation in practice (118). The estimated causal parameter is generally assumed to be normally distributed (119). The variance for the two-stage estimator with continuous outcomes is here calculated using a sandwich variance estimator to account for uncertainty in the first-stage regression (120; 121). Alternatively, uncertainty can be incorporated by the use of bootstrap confidence intervals (122; 123). The 2SLS estimator has a finite kth moment with at least $(k + 1)$ instruments when all the associations are linear and the error terms normally distributed (124). Estimates are consistent under the assumption of homoskedasticity and correct specification of the linear regressions (117).

With multiple instruments, the 2SLS estimator may be viewed as a weighted average of the ratio estimators using the instruments one at the time, where the weights are

determined by the relative strength of the instruments in the first-stage regression (112; 118).

2.6.2 Binary outcome

The analogue of 2SLS with binary outcomes is a two-stage estimator where the second-stage regression (X-Y regression) uses a log-linear or logistic regression model. This has been called the two-stage estimator (125), standard IV estimator (94), pseudo-2SLS (126), two-stage predictor substitution (2SPS) (127; 128) or Wald-type estimator (99).

However, such regression methods do not always yield 'consistent' estimators and have been called "forbidden regressions" (118; 129). For example, in the logistic case, the parameter β_1 in the logistic model is estimated with bias (99; 130). This is because the non-linear model does not guarantee that the residuals from the second-stage regression are uncorrelated with the instruments (126).

An alternative estimate has been proposed, using the residuals from the regression of phenotype on genotype in the regression of disease on genotype (94). This is known as a control function approach (131), or two-stage residual inclusion (2SRI) (127). If we have a first stage regression of X on G with fitted values $\hat{X}|G$ and residuals $\hat{R}|G = X - \hat{X}|G$, then the alternative IV estimator comes from a logistic regression additively on $\hat{X}|G$ and $\hat{R}|G$. The residual incorporates information from confounders in the first stage regression, for example with X defined as in equation (2.3), $\mathbb{E}(R|X = x, U = u) = \alpha_2 u$.

Sandwich variance estimators can be calculated, although coverage may be poor due to inconsistent estimation of the parameter β_1 (94).

2.7 Likelihood-based methods

We consider the likelihood-based limited information maximum likelihood method and a Bayesian framework which can use a similar model. These likelihood-based methods are parametric, in contrast to the semi-parametric methods of Section 2.8.

2.7.1 Limited information maximum likelihood method

If we have the linear model (2.3) and (2.4) but subsume the confounder into the error structure, such that for individual $i = 1, \ldots, N$:

$$\begin{aligned} x_i &= \alpha_0 + \sum_k \alpha_k g_{ik} + \epsilon_{xi} \\ y_i &= \beta_0 + \beta_1 x_i + \epsilon_{yi} \end{aligned} \tag{2.14}$$

then we can make assumptions of a bivariate normal distribution for $\boldsymbol{\epsilon} = (\epsilon_Y, \epsilon_X)^T \sim \mathcal{N}(0, \Sigma)$ and calculate the maximum likelihood estimate of β_1. This is known as limited information maximum likelihood (LIML) (27). We maximize the likelihood substituting for and profiling out (referred to by economists as 'concentrating out') Σ. If we rewrite the equations (2.14) as:

$$\begin{pmatrix} 1 & -\beta_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\alpha} \\ 0 \end{pmatrix} G + \begin{pmatrix} \epsilon_Y \\ \epsilon_X \end{pmatrix} \tag{2.15}$$

where $\boldsymbol{\alpha} = (\alpha_1 \ \ldots \ \alpha_K)$ and then define matrices $B = \begin{pmatrix} \boldsymbol{\alpha} \\ 0 \end{pmatrix}$ and $\Gamma = \begin{pmatrix} 1 & -\beta_1 \\ 0 & 1 \end{pmatrix}$ then we can write the profile likelihood as

$$N \log |\det \Gamma| - \frac{N}{2} \log \left| \frac{1}{n} \left(\begin{pmatrix} Y \\ X \end{pmatrix} \Gamma - BG \right)^T \left(\begin{pmatrix} Y \\ X \end{pmatrix} \Gamma - BG \right) \right| \tag{2.16}$$

We then maximize the profile likelihood to find the LIML estimate of β_1, noting that $\det |\Gamma| = 1$.

An alternative to LIML is full information maximum likelihood (FIML) (27). In FIML, each of the equations in the model are estimated simultaneously, whereas in LIML only a limited number of the equations are estimated and the other parameters are profiled out. For example, if there are measured covariates, then these can be incorporated into the model. If we seek to simultaneously model these covariates as functions of Y, X and G, in LIML the covariates are replaced by their unrestricted reduced form (i.e. written in terms of the parameters of equations 2.15), and only the parameters relevant to the equations of interest are estimated. Hence LIML is similar to FIML where there is a single phenotype of interest, but where there are multiple phenotypes, some of which are of interest, the estimates differ.

The LIML estimate ($\hat{\beta}_1^L$) minimizes the residual sum of squares from the regression of the component of Y not caused by X, $(y_i - \hat{\beta}_1^L x_i)$, on G. Informally, the LIML estimator is the one for which the component of Y due to confounding is as badly predicted by G as possible.

LIML has been called the "maximum likelihood counterpart of 2SLS" (132) and is equivalent to 2SLS with a single instrument and single phenotype. As with 2SLS, estimates are sensitive to heteroskedasticity and misspecification of the equations in the model. Use of the LIML estimator has been strongly discouraged, as it does not have defined moments for any number of instruments (133). However, use has also been encouraged especially with weak instruments, as the median of the distribution of the estimator is close to unbiased with even weak instruments (118).

2.7.2 Bayesian methods

Although Bayesian techniques for IV analysis do exist in the econometrics literature (134; 135) and the non-compliance literature (136), Bayesian methods for IVs are rare and have not received much attention from applied practitioners (137). In the context of genetic epidemiology, they have been used for meta-analysis of summary results from Mendelian randomization studies (71; 138) and modelling of gene-phenotype associations (139)

Bayesian methods have been recently proposed for IV analysis in the context of Mendelian randomization (140; 141). Models equivalent to equations (2.14) from LIML can be estimated in a Bayesian setting. Bayesian models are appealing due to the flexibility of the modelling assumptions, lack of reliance on conventional asymptotics for inference, correct propagation of uncertainty through the model, and natural extension to meta-analysis through the use of hierarchical modelling. A drawback is that prior distributions of the model parameters and error structures of the random variables must be fully specified. Posterior distributions can be estimated using Monte Carlo Markov chain (MCMC) methods. Bayesian methods will be discussed further in Chapter 5.

2.8 Semi-parametric methods

A semi-parametric model is a model with both parametric and non-parametric components. Typically semi-parametric estimators with IVs assume a parametric form assumed for the equations relating the outcome and phenotype, but make no assumption on the distribution of the errors. Semi-parametric models are designed to be more robust to model misspecification than fully parametric models (97).

2.8.1 Generalized method of moments

The generalized method of moments (GMM) is a semi-parametric estimator designed as a more flexible form of 2SLS to deal with problems of heteroskedasticity of error distributions and non-linearity in the two-stage structural equations (126; 142). With a single instrument, the estimator is chosen to give orthogonality between the instrument and the residuals from the second-stage regression. Using bold face to represent vectors, if we have

$$\mathbb{E}(y) = f(x; \boldsymbol{\beta}) \tag{2.17}$$

then the GMM estimate is the value of $\boldsymbol{\beta}$ such that

$$\sum_i (y_i - f(x_i; \boldsymbol{\beta})) = 0 \tag{2.18}$$
$$\text{and } \sum_i g_i (y_i - f(x_i; \boldsymbol{\beta})) = 0$$

where the summation is across i, which indexes study participants. In the linear case, $f(x_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_i$; in the log-linear case, $f(x_i; \boldsymbol{\beta}) = \log(\beta_0 + \beta_1 x_i)$; and in the logistic case, $f(x_i; \boldsymbol{\beta}) = \text{logit}(\beta_0 + \beta_1 x_i)$; where β_1 is our causal parameter of interest. We can solve these two equations numerically (142).

When there is more than one instrument, g_i becomes g_{ik} and we have a separate estimating equation for each instrument k for $k = 1, \ldots, K$. The orthogonality conditions for each instrument cannot generally be simultaneously satisfied. The estimate is taken as the minimizer of the objective function

$$(\mathbf{y} - f(\mathbf{x}; \boldsymbol{\beta}))^T \mathrm{G} (\mathrm{G}^T \Omega \mathrm{G})^{-1} \mathrm{G}^T (\mathbf{y} - f(\mathbf{x}; \boldsymbol{\beta})) \tag{2.19}$$

where $\mathrm{G} = (\mathbf{1}\, \mathbf{g_1} \ldots \mathbf{g_K})$ is the N by $K + 1$ matrix of instruments, including a column of 1s for the constant term in the G-X association.

Although this gives consistent estimation for general Ω, efficient estimation is achieved when $\Omega_{ij} = \text{cov}(\epsilon_i, \epsilon_j)$ $(i, j = 1, \ldots, N)$, where ϵ_i is the residual $y_i - f(x_i; \boldsymbol{\beta})$ (143). As the estimation of Ω requires knowledge of the unknown $\boldsymbol{\beta}$, we use the two-stage approach of Greene (144). We firstly estimate $\boldsymbol{\beta^*}$ using $(\mathrm{G}^T \Omega \mathrm{G}) = I$, where I is the identity matrix, which gives consistent but not efficient estimation of $\boldsymbol{\beta}$. We then use $e_i = y_i - f(x_i; \boldsymbol{\beta^*})$ to estimate $\mathrm{G}^T \Omega \mathrm{G} = \sum_i \mathbf{g_i} \mathbf{g_i}^T \epsilon_i^2$ as $\sum_i \mathbf{g_i} \mathbf{g_i}^T e_i^2$ in a second-stage estimation (142).

2.8.2 Structural mean models

The structural mean model (SMM) approach is another semi-parametric estimator designed in the context of randomized trials with incomplete compliance (145; 146). We recall that the potential outcome $Y(x)$ is the outcome which would have been observed if the phenotype X were set to x. This is also written as $Y|do(X = x)$ (147). In particular, the exposure-free outcome $Y(0)|X = x$ is the outcome which would have been observed if we had set $X = 0$ (97). Explicit conditioning is performed on $X = x$ to show that no other variable is changed from the value it would take if $X = x$ were true. We note that the expectation $\mathbb{E}(Y(0)|X = x)$ is typically different from the expected outcome if $X = 0$ had been observed, as intervening on X alone would not change the confounder distribution. An explicit parametric form is assumed for the expected difference in potential outcomes

between the outcome for the observed $X = x$ and the potential outcome for $X = 0$. In the continuous case, the linear or additive SMM is

$$\mathbb{E}(Y(x)) - \mathbb{E}(Y(0)|X = x) = \beta_1 x \tag{2.20}$$

and β_1 is taken as the causal parameter of interest. In the context of non-compliance, this is referred to as the "effect of treatment on the treated" (148).

As the expected exposure-free outcome $\mathbb{E}(Y(0)|X = x)$ is statistically independent of G, the causal effect is estimated as the value of β_1 which gives zero covariance between $\mathbb{E}(Y(0)|X = x) = \mathbb{E}(Y(x) - \beta_1 x)$ and G. This process is known as 'G-estimation' (149; 150). The estimating equations are

$$\sum_i (g_{ik} - \bar{g}_k)(y_i - \beta_1 x_i) = 0 \qquad k = 1, \ldots, K \tag{2.21}$$

where $\bar{g}_k = \frac{1}{N} \sum_i g_{ik}$ and the summation is across i, which indexes study participants.

Where the model for the expected outcomes is non-linear, this is known as a generalized structural mean model (GSMM). With a binary outcome, it is natural to use a log-linear or multiplicative GSMM:

$$\log \mathbb{E}(Y(x)) - \log \mathbb{E}(Y(0)|X = x) = \beta_1 x \tag{2.22}$$

Unfortunately, due to non-collapsibility, the logistic GSMM cannot be estimated in the same way, as the expectation $\operatorname{logit} \mathbb{E}(Y(x))$ depends on the distribution of the IV $p(g)$ (151). Vansteelandt and Goetghebeur address this problem by estimating $Y(x)$ assuming an observational model (152):

$$\operatorname{logit} \mathbb{E}(Y(x)) = \beta_{0a} + \beta_{1a} x \tag{2.23}$$

where the subscripts a indicate associational, as well as an GSMM model:

$$\operatorname{logit} \mathbb{E}(Y(x)) - \operatorname{logit} \mathbb{E}(Y(0)|X = x) = \beta_{1c} x \tag{2.24}$$

where the subscript c indicates causal. The associational parameters can be estimated by logistic regression, leading to estimating equations

$$\sum_i (g_{ik} - \bar{g}_k) \operatorname{expit}(\hat{Y}(x) - \beta_{1c} x_i) = 0 \qquad k = 1, \ldots, K \tag{2.25}$$

where $\operatorname{logit} \hat{Y}(x) = \hat{\beta}_{0a} + \hat{\beta}_{1a} x$ (153).

We note that the choice of estimating equations presented here are not the most efficient, but lead to consistent estimates (152).

2.9 Method of Greenland and Longnecker

A method of Greenland and Longnecker for meta-analysis of summarized data (154) has been proposed for Mendelian randomization analysis (155). The method for meta-analysis uses summary data in the form of log odds ratios for different exposure groups relative to a baseline group. These ratios are correlated, and so an estimate of the overall effect is calculated allowing for correlation using generalized least squares regression.

In adopting the method for IV analysis, we partition individuals into genotypic subgroups, with every individual in each subgroup having the same genotype. We estimate the difference in average phenotype and in log odds ratio of each subgroup compared to a baseline subgroup, and estimate a causal effect of increase in log odds ratio of disease for a unit increase in phenotype allowing for correlation between the subgroups using generalized least squares regression. The subgroups take the place of the exposure groups in the original method (65). This method is similar to one proposed for Bayesian analysis presented in Chapter 5. It does not require individual participant data, only numbers of diseased and healthy individuals and mean phenotype values in each subgroup. No allowance is made for the possible uncertainty in the mean phenotype values.

2.10 Comparison of methods

In several cases, estimates from different IV methods coincide. With a single instrument, the ratio and two-stage estimates are equal (99), and in the continuous setting the 2SLS, LIML, GMM and SMM point estimates coincide, although their estimates of uncertainty may not (97). For a general instrument, the linear (additive) and log-linear (multiplicative) GMM and GSMM models give rise to the same estimates (97). This is not true in the logistic case (97).

We will consider the following features when comparing IV methods: existence of finite moments, mean bias, median bias, coverage under the null, power, and robustness to model misspecification. Median bias refers to the difference between the median of the estimator over its distribution and the true parameter value. We generally prefer median bias as a criterion to mean bias, as mean bias is undefined when an estimator has no finite first moment. We are especially concerned about the behaviour of the estimators when the instruments are not strongly associated with the phenotype, so called weak instruments (see Section 2.13). Chapter 6 includes a theoretical discussion of the methods and comprehensive set of simulations for empirical comparison.

2.11 Efficiency and validity of instruments

2.11.1 Use of measured covariates

If we can find measured covariates which explain variation in the phenotype or outcome, and which are not on the causal pathway between phenotype and outcome, then we can incorporate such covariates in our model. In econometrics, such a variable is called an an exogenous regressor or included instrument, as opposed to an IV, which is called an excluded instrument (117). This is because the covariate is included in the model for the outcome. Incorporation of covariates increases efficiency and precision of the causal estimate (118). In a two-stage estimation, any covariate adjusted for in the first-stage regression should also be adjusted for in the second-stage regression; failure to do so can cause associations between the IV and confounders leading to bias. When adjusting for covariates, the correct measure of instrument strength is a partial R^2 statistic (156) (see Section 2.13).

2.11.2 Overidentification tests

When more than one instrument is used, an overidentification test, such as the Basmann test (157) or Sargan test (158), can be carried out to test whether the instruments have additional effects on the outcome beyond that mediated by the phenotype (30). Overidentification means that the number of instruments used in a GMM (or 2SLS) method is greater than the number of phenotypes measured. (The latter is usually one, although causal effects for additional phenotypes could be simultaneously estimated if the IV is valid for more than one phenotype.) This means that there is no unique solution to the GMM equations. The overidentification test is equivalent to testing whether the IVs have residual associations with the outcome once the main effect of the phenotype has been removed (30).

For example, the Sargan test statistic (117) is motivated as the average of the residual sum of squares in the regression of residuals from the IV regression on the instruments. It has a χ^2_{K-1} distribution under the null hypothesis of asymptotic orthogonality of the instruments to the IV residual errors, where K is the number of instruments.

$$\text{Sargan's statistic } = (y - \hat{\beta}_0 - \hat{\beta}_1 x)^T (\mathrm{I} - P_\mathrm{G})(y - \hat{\beta}_0 - \hat{\beta}_1 x)/N \tag{2.26}$$

where $P_\mathrm{G} = \mathrm{G}(\mathrm{G}^T\mathrm{G})^{-1}\mathrm{G}^T$ is the projection matrix of $\mathrm{G} = (\mathbf{1}\,\mathbf{g_1} \ldots \mathbf{g_K})$, the N by $K+1$ matrix of instruments. I is the identity matrix and N is the total number of individuals.

Overidentification tests are omnibus tests, where the alternative hypothesis includes failure of IV assumptions for one IV, failure for all IVs, and non-linear association between phenotype and outcome (117). They have limited power (30) and so may have limited practical use in detecting violations of the IV assumptions.

2.12 Meta-analysis

Having considered methods for the analysis of a single Mendelian randomization study, we turn our attention to the issue of meta-analysis. Meta-analysis of Mendelian randomization studies is of particular interest as it is generally necessary for precise estimation of the gene-phenotype and gene-disease associations (40), and hence for the estimation of the causal effect.

If it is possible to estimate the causal effect in each study, a meta-analysis can be performed directly on the estimated causal associations (89). However, due to imprecise or near-zero G-X association in some studies, some of the causal associations can have large or even infinite variance.

The simplest situation for meta-analysis is when a single dichotomous IV is used, which is the same in all studies. One difficulty is that when some studies are used in calculating both G-X and G-Y associations, these estimates will be correlated (2). If all studies measure both these associations, we can test the effect of phenotype on outcome by plotting a graph of the regression estimates of G-Y association against the regression estimates of G-X association (89). The points on this graph will have error in both directions and the gradient of the graph will show the causal X-Y association.

To include studies when either or both associations have been reported, a bivariate distribution of phenotype difference and outcome difference can be assumed, with variance-covariance matrix the sum of two components, for within and between study heterogeneity (71). For each study m measuring both G-X and G-Y associations, the estimated G-X association $\hat{\beta}_{GXm}$ is assumed to be normally distributed with mean μ_{xm} and variance v_{xm} and the estimated G-Y association $\hat{\beta}_{GYm}$ is normally distributed with mean μ_{xm} and variance v_{xm}. The correlation τ between β_{GXm} and β_{GYm} is assumed to be independent of m. The mean values μ_{xm} are assumed normally distributed across studies with mean μ_x and variance σ_x and the mean values μ_{ym} are normally distributed with mean μ_y and variance σ_y with correlation ψ between μ_{xm} and μ_{ym}.

$$\begin{pmatrix} \hat{\beta}_{GXm} \\ \hat{\beta}_{GYm} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \mu_{xm} \\ \mu_{ym} \end{pmatrix}, \begin{pmatrix} v_{xm} & \tau\sqrt{v_{xm}\, v_{ym}} \\ \tau\sqrt{v_{xm}\, v_{ym}} & v_{ym} \end{pmatrix} \right) \tag{2.27}$$
$$\begin{pmatrix} \mu_{xm} \\ \mu_{ym} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \mu_{x} \\ \mu_{y} \end{pmatrix}, \begin{pmatrix} \sigma_x & \psi\sqrt{\sigma_x\, \sigma_y} \\ \psi\sqrt{\sigma_x\, \sigma_y} & \sigma_y \end{pmatrix} \right)$$

To include studies where only one of the associations has been reported, we use the marginal distribution of $\hat{\beta}_{GXm}$ or $\hat{\beta}_{GYm}$ as appropriate. The correlation τ between associations within each study is usually assumed to be zero (138). A sensitivity analysis shows that this assumption is reasonable and robust (71). This method for the meta-analysis of the effect of one SNP on X and Y can be extended to treat G as a trichotomous random variable (138), corresponding to the three possible values of a SNP.

The parameters in the meta-analysis can be estimated either by maximization of the log-likelihood using numerical methods (71) or by using Bayesian methods with flat priors (71; 138). We note that this method only covers meta-analysis of causal associations when one SNP is measured in every study; Bayesian methods for meta-analysis to cover situations of multiple SNPs and different SNPs will be considered further in Chapter 5.

2.13 Weak instruments

Practical application of IV methods, especially in a Mendelian randomization context, is complicated by the issue of weak instruments. A 'weak instrument' is defined as an instrument for which the statistical evidence of association with the phenotype (X) is not strong (2). An instrument can be weak if it explains a small amount of the variation of the phenotype, where the amount defined as 'small' depends on the sample size (44). The F statistic in the first stage regression of X on G is usually quoted as a measure of the strength of an instrument (92). In this context, the F statistic is also known as the Cragg–Donald statistic (159). It is related to the proportion of variance in the phenotype explained by the genetic variants (R^2), sample size (N) and number of instruments (K) by the formula $F = (\frac{N-K-1}{K})\,(\frac{R^2}{1-R^2})$. As the F statistic depends on the sample size and number of instruments used, instrument strength is not a property of the genetic variant itself; the absolute strength of an instrument is only relevant in the context of a specific dataset. Weak instruments typically produce estimates of causal association with wide confidence intervals, but there is a further troublesome aspect to IV estimation with weak instruments.

Although IV methods are asymptotically unbiased, they demonstrate systematic finite sample bias. This bias, known as 'weak instrument bias', acts in the direction of the

confounded observational association between phenotype and outcome, and depends on the strength of the instrument (160). Weak instruments are also associated with underestimated confidence intervals and poor coverage properties (161). A generally quoted criterion is that an instrument is weak if the F statistic in the G-X regression is less than 10 (2; 102). However, using instruments with $F > 10$ only reduces bias to less than a certain level, and problems with weak instrument bias still occur (92).

A power-series expansion shows that the bias in the IV estimator is related to the F statistic (F) from the G-X regression (101). As F decreases, the bias of the IV estimator approaches the bias of the confounded association. If we consider "weak instrument asymptotics", where as the sample size increases, the coefficients in the G-X regression tend to zero and specifically are $O(N^{-\frac{1}{2}})$, where N is the sample size, then, as the sample size tends to infinity, the F statistic from the G-X regression tends to a finite limit (102). We consider the relative mean bias B, which is the ratio of the bias of the IV estimator to the bias of the confounded association $\hat{\beta}_{OBS}$ found by linear regression of Y on X:

$$B = \frac{\mathbb{E}(\hat{\beta}_{IV}) - \beta_1}{\mathbb{E}(\hat{\beta}_{OBS}) - \beta_1} \tag{2.28}$$

This measure has the advantage of invariance under change of units in Y. The relative mean bias in this case from the 2SLS method is asymptotically approximately equal to $1/F$ (102).

The accuracy of this approximation has been assessed by tabulating a series of critical values derived from simulations of the required F statistic to ensure, with 95% confidence, a relative mean bias in the 2SLS method of less than 5%, 10%, 20% and 30% with a given number of instruments (161). While this approximation is reasonable for a large number of instruments, it is less accurate when there are few instruments, as there typically are in an epidemiological context. Indeed, the relative mean bias cannot be estimated when there is only one instrument, since the 2SLS IV estimator and hence B has no finite kth moment when the number of instruments is less than or equal to k (162).

While the topic of weak instrument bias has been discussed for some years in econometrics (160; 163; 164), it is not well understood in relation to Mendelian randomization; this will be considered further in Chapters 2 and 6.

2.14 Computer implementation

Several commands are available in statistical software packages for IV estimation, such as R (165) and Stata (166). The commands in Stata *ivreg2*, *ivhettest*, *overid*, and *ivendog* (117)

have been written to implement the 2SLS method, with estimators and tests, including the Sargan statistic and the F (Cragg–Donald) statistic. The R command *tsls* in the library *sem* carries out a 2SLS procedure (167). The command in Stata *qvf* (123) has been written to implement a fast bootstrap estimation of standard errors for IV analysis. This can be used with non-linear models, such as with binary outcomes.

Linear GMM and SMM can be estimated in Stata using the *ivreg2* command with option *gmm* (117) or the *ivregress* command. Multiplicative GMM or GSMM can be estimated in Stata using the *ivpois* command (168). Generic estimating equations for GMM or GSMM can be solved in Stata using the *gmm* command (169) and in R using the *gmm* package (170).

The method of Greenland and Longnecker has been implemented in Stata as the *glst* command (171).

2.15 Mendelian randomization in practice

Having considered the methodological aspects of IV estimation for Mendelian randomization, we present some examples of the use of the methods and techniques listed above in epidemiological practice.

The majority of Mendelian randomization studies have used a single SNP as the genetic variant. Casas et al. (85) investigate the causal effect of C-reactive protein (CRP) on incident coronary events. They look at the effect of one gene on CRP levels, showing a significant association between the gene and CRP levels, but no association between the gene and disease, though with wide CIs on the G-Y association. Keavney et al. (172) assess the causal association of fibrinogen on coronary heart disease (CHD). Although there is a significant per allele effect on fibrinogen levels, there is no association between the genetic variant and CHD incidence, with fairly tight CIs. In each of these studies, tests of the association between the gene and known competing risk factors have been carried out, to assess the IV assumptions. No formal IV analysis is attempted and no estimate is made of the causal X-Y association.

The assessment of causal association has also been undertaken using multiple studies. Lewis et al. (76) show a null association between a genetic polymorphism associated with homocysteine levels and CHD in a random-effects meta-analysis comparing participants with two different genotypes. Lewis and Davey Smith (41) show a statistically significant result in a meta-analysis of the effect of alcohol on oesophageal cancer. Here, estimating the causal X-Y association was not possible, as the association of genotype with alcohol

intake was not linear in the three genotypes, and the alcohol intake in different studies showed considerable heterogeneity.

Causal estimation using the ratio method was employed by Kamstrup et al. (68) in assessing the causal association between lipoprotein(a) and risk of myocardial infarction. They demonstrate the crucial role played by the magnitude of the G-X association. In their study, known genetic variants explained 21-27% of the variation in lipoprotein(a), leading to a statistically significant estimate of the causal effect. In contrast, Lawlor et al. (31) show a null association between a genetic polymorphism associated with CRP levels and CHD in a random-effects meta-analysis comparing participants with TT versus CT or CC genotype, but the confidence interval for the causal estimate included the observational association estimate, despite a greater sample size and number of events. This is because the genetic marker used only explained less than 1% of the variance in CRP.

The 2SLS method has been used to synthesize evidence using haplotypes as an IV to test the effect of CRP on HOMA-R (2) (a measure of insulin resistance) and CRP on carotid intima-media thickness (62). Kivimäki et al. (62) measure three genetic variants which they combine as haplotypes, and use the four most common haplotypes as instruments. They note that the haplotypes are associated with CRP levels, but that there is no significant association between the haplotypes and CIMT. The 2SLS method gives a null causal association between CRP and CIMT, although with wide CIs. The confidence intervals given by this method are large compared to a standard multivariable regression technique adjusting for measured confounders. Lawlor et al. (2) take the most common pair of haplotypes (diplotype) for each participant as an IV to assess the causal association of CRP on HOMA-R. They exclude diplotypes with less than 10 participants, and plot CRP against HOMA-R for each of the 9 subgroups, using 2SLS to assess the association.

Timpson et al. (61) use 2SLS, but take the Durbin–Wu–Hausman test as the primary outcome of interest. This is a test of equality of the observational and IV associations, where a significant result indicates disagreement between the two estimates. However, this is not a test of no causal effect, as there may be a causal effect, but this may be different to the observational association. For this reason, it is more appropriate to consider the causal estimate as the outcome of interest (173).

The two-stage method has been used with binary outcomes to test the causal association of CRP on hypertension (174) and of sex-hormone binding globulin on type 2 diabetes (67). Confidence intervals were estimated by bootstrapping techniques using the *qvf* command in Stata.

Several variations on the two-stage method have been attempted with methods developed either heuristically or borrowed from other areas of research. Elliott et al. (63)

simply scale the coefficients in the G-Y regression by an estimate of the G-X association. Allin et al. (65) use the method of Greenland and Longnecker documented above. Neither of these allow for the uncertainty in the G-X association.

2.16 Conclusion

Although there is a wealth of IV methodology accumulated from many years of econometric research and practice, practical use of IV methodology in Mendelian randomization is limited and not well understood. This is for three main reasons. Firstly, there is a need for translational research to assess the implementation of IV research in the specific context of Mendelian randomization (30). This requires the search for a mutual language between medical statisticians and econometricians (31), as well as an investigation of the application of techniques and methods common in econometric practice in an epidemiological setting. An example is the use of measured covariates, which is common in econometric analysis but rare in Mendelian randomization practice, possibly due to the analogy of Mendelian randomization with an RCT, where adjustment for measured covariates is not uniformly practised. In areas such as weak instrument bias, where there is a growing body of research evidence, translational work is needed to see how the findings and practice of economics translates to the context of Mendelian randomization.

Secondly, there are still unanswered questions about the estimation of causal effects using IVs. In this dissertation, we focus on the issues of weak instruments and binary outcomes. The instruments used in Mendelian randomization typically have a small effect on the phenotype and show a high degree of correlation. Research is needed to investigate the effect of the use of weak instruments and multiple instruments on Mendelian randomization estimation, to find ways of minimizing bias and maintaining accurate coverage properties. We seek to form guidelines as to how to choose how many and which instruments to use in applied research. The majority of applications of Mendelian randomization involve binary outcomes, and so estimation of a causal effect which can be compared with an observational effect is of great practical importance. The bias of the ratio estimate in a logistic model and the status of "forbidden regressions" are highly relevant to applied analysis.

Thirdly, causal estimates from IV analysis tend to have wide confidence intervals compared to conventional epidemiological estimates, which deters applied researchers from reporting numerical results from IV analysis. We seek to expand methods for meta-analysis of Mendelian randomization to cover features exhibited in the CCGC, such as the availability of multiple genetic variants and individual participant data, to make efficient use of

data even with heterogeneous studies. We seek to exploit the structure of genetic data to find methods for imputation of missing data to maximize information from a given study.

Although the literature on IVs from econometrics and non-compliance provides methods for IV analysis which can be translated into a Mendelian randomization context, the specific nature of Mendelian randomization gives rise to issues which have not been adequately addressed elsewhere in the literature. This dissertation is intended to "bridge the gap", both to answer some of the open methodological questions concerning IV analysis and to communicate findings in existing research, hopefully leading to more principled analysis of Mendelian randomization studies.

Chapter 3

Weak instrument bias for continuous outcomes

3.1 Introduction

Although IV techniques can be used to give asymptotically unbiased estimates of causal association in the presence of confounding, these estimates suffer from a bias, known as weak instrument bias, when evaluated in finite samples (160; 163; 164). This bias acts in the direction of the observational confounded association, and its magnitude depends on the strength of association between genetic instrument and phenotype (34; 101). In this chapter, we consider the effect of this bias for continuous outcomes; we consider the biases affecting IV estimates with a binary outcome in Chapter 6.

We use data from the CRP CHD Genetics Collaboration (82) to estimate the causal association of C-reactive protein (CRP) on fibrinogen. Both CRP and fibrinogen are markers for inflammation. As the distribution of CRP is positively skewed, we take its logarithm and assume a linear association of log(CRP) on fibrinogen. Although log(CRP) and fibrinogen are highly positively correlated ($r = 0.45 - 0.55$ in the studies below), it is thought that long-term elevated levels of CRP are not causally associated with an increase in fibrinogen (64).

In this chapter, we demonstrate the direction and magnitude of weak instrument bias in IV estimation from simulated data, and show that it can be an important issue in practice (Section 3.2). We explain why this bias comes about, why it acts in the direction of the confounded observational association and why it is related to instrument strength (Section 3.3). We quantify the size of this bias for different strengths of instruments and different analysis methods, describing how important the bias may be expected to be in a given application (Section 3.4). When multiple genetic variants or models of genetic

association are available, we show how the choice of IV affects the variance and bias of the IV estimator (Section 3.5). We discuss methods of design and analysis of Mendelian randomization studies to minimize bias (Section 3.6). We conclude (Section 3.7) with a discussion of this bias from a theoretical and practical viewpoint, ending with a summary of recommendations aimed at applied researchers for how to design and analyse a Mendelian randomization study.

3.2 Demonstrating the bias from IV estimators

Firstly, we seek to demonstrate the bias in IV estimation using both real and simulated data.

3.2.1 Bias of IV estimates in small studies

As a motivating example, we consider the Copenhagen General Population Study (CGPS) (175), a cohort study from the CRP CHD Genetics Collaboration (CCGC) with complete cross-sectional baseline data on CRP, fibrinogen and three SNPs from the CRP gene region (rs1205, rs1130864 and rs3093077) for 35 679 participants. We calculate the observational estimate (simply regressing fibrinogen on log(CRP)) and IV estimate of association using all three SNPs as instrumental variables in a linear additive model. We then analyze the same data as if it came from multiple studies by dividing the study randomly into substudies of equal size, calculating estimates of association in each substudy and meta-analyzing the results using a fixed-effect model. We divide into 5, 10, 16, 40, 100 and 250 substudies.

We see from Table 3.1 that the observational estimate stays almost unchanged whether the data are analyzed as one study or as several studies. However, the IV estimate increases from near zero until it approaches the observational estimate and the standard error of the estimate decreases. We can see that even where the number of substudies is 16 and the average F statistic is around 10, there is a serious bias with a positive causal estimate ($p = 0.09$ using 2SLS) despite the causal estimate with the data analyzed as one study being near zero.

3.2.2 Simulation with one IV

As a simulation exercise, we take a simple example of a confounded association with a single IV, as considered previously in Section 2.4.1. Phenotype x_i for individual i is a linear combination of a genetic component g_i which can take values 0 or 1, normally distributed

No. of substudies	Observational estimate	2SLS estimate	LIML estimate	Mean F
1	1.6799 (0.0143)	-0.0468 (0.1510)	-0.0531 (0.1515)	152.0
5	1.6796 (0.0143)	-0.0092 (0.1478)	-0.0541 (0.1508)	31.44
10	1.6789 (0.0143)	0.0871 (0.1426)	-0.0068 (0.1485)	16.44
16	1.6781 (0.0143)	0.2300 (0.1372)	0.1641 (0.1426)	10.81
40	1.6761 (0.0143)	0.4562 (0.1266)	0.3093 (0.1385)	4.833
100	1.6713 (0.0142)	0.8279 (0.1078)	0.6575 (0.1279)	2.516
250	1.6695 (0.0141)	1.2711 (0.0826)	1.1796 (0.1022)	1.646

Table 3.1: Estimates of effect (standard error) of log(CRP) on fibrinogen (μmol/l) from Copenhagen General Population Study (N = 35 679) divided randomly into substudies of equal size and combined using fixed-effect meta-analysis: observational estimate using unadjusted linear regression, IV estimate from Mendelian randomization using 2SLS and LIML methods. F statistics from linear regression of log(CRP) on three genetic IVs averaged across substudies.

confounder u_i, and error ϵ_{xi} terms. Outcome y_i is a linear combination of x_i and u_i with normally distributed error ϵ_{yi}. The true causal association of X on Y is represented by β_1. To simplify, we have set the constant terms in the equations to be zero:

$$\begin{aligned} x_i &= \alpha_1\, g_i + \alpha_2\, u_i + \epsilon_{xi} \\ y_i &= \beta_1\, x_i + \beta_2\, u_i + \epsilon_{yi} \\ u_i &\sim \mathcal{N}(0, \sigma_u^2) \\ \epsilon_{xi} &\sim \mathcal{N}(0, \sigma_x^2); \epsilon_{yi} \sim \mathcal{N}(0, \sigma_y^2) \text{ independently} \end{aligned} \tag{3.1}$$

We simulated 50 000 datasets from this model, each with 200 individuals divided equally between the two genotypic subgroups, for a range of values of α_1. We set $\beta_1 = 0, \alpha_2 = 1, \beta_2 = 1, \sigma_u^2 = \sigma_x^2 = \sigma_y^2 = 1$, corresponding to a true null causal association, but simply regressing Y on X yields a strong positive confounded observational association of close to 0.5. We took 6 different values of α_1 from 0.05 to 0.55, thus varying the strength of the G-X association, corresponding to mean F statistic values between 1.07 and 8.65.

Causal estimates are calculated using the ratio method, although with a single linear instrument the estimates from the ratio, 2SLS and LIML methods are the same. The resulting distributions for the estimate of the causal parameter β_1 are shown in Figure 3.1 and Table 3.2. Because the IV estimate can be expressed as the ratio of two normally distributed random variables, it does not have a finite mean or variance; so we have expressed results using quantiles. For smaller values of α_1, there is a marked median bias in the positive direction and long tails in the distribution of the causal estimate. For the

smallest value $\alpha_1 = 0.05$, the mean F statistic is barely above its null expectation of 1 and the median IV estimate is close to the confounded observational estimate. For large values of α_1, the causal estimates have a skew distribution, with median close to zero but with more extreme causal estimates tending to take negative values. The F statistics vary greatly between simulations for each given α_1, with an interquartile range of similar size to the mean value of the statistic (Table 3.2). In practical applications therefore the F statistic from a single analysis is not necessarily a reliable guide to the underlying mean F statistic.

α_1	Mean F statistic (Observed IQ range)	Quantiles: 2.5%	25%	50%	75%	97.5%
0.05	1.07 (0.11 - 1.41)	-10.5393	-0.3859	0.4686	1.3159	10.8918
0.15	1.58 (0.18 - 2.17)	-9.2289	-0.4436	0.2870	0.9819	9.3405
0.25	2.59 (0.44 - 3.73)	-6.4495	-0.4672	0.1296	0.5983	5.8267
0.35	4.10 (1.17 - 5.94)	-4.0480	-0.4124	0.0456	0.3838	2.8776
0.45	6.12 (2.49 - 8.55)	-2.4233	-0.3423	0.0108	0.2806	0.9167
0.55	8.65 (4.27 - 11.81)	-1.5435	-0.2849	0.0002	0.2247	0.6417

Table 3.2: Quantiles of IV estimates of causal association $\beta_1 = 0$ using weak instruments with different mean F statistics (interquartile range (IQ)) from simulated data

3.3 Explaining the bias from IV estimators

We give three separate explanations for the existence of weak instrument bias, using the languages of algebra, random variables and graphs.

3.3.1 Correlation of associations

Firstly, there is a correlation between the numerator (G-Y association) and denominator (G-X association) in the ratio estimator. In the zero error case ($\sigma_x^2 = \sigma_y^2 = 0$) with true causal association of X on Y, and confounded association through U, model (3.1) reduces to

$$\begin{aligned} x_i &= \alpha_1 g_i + \alpha_2 u_i \\ y_i &= \beta_1 x_i + \beta_2 u_i \\ u_i &\sim \mathcal{N}(0, \sigma_u^2) \end{aligned} \tag{3.2}$$

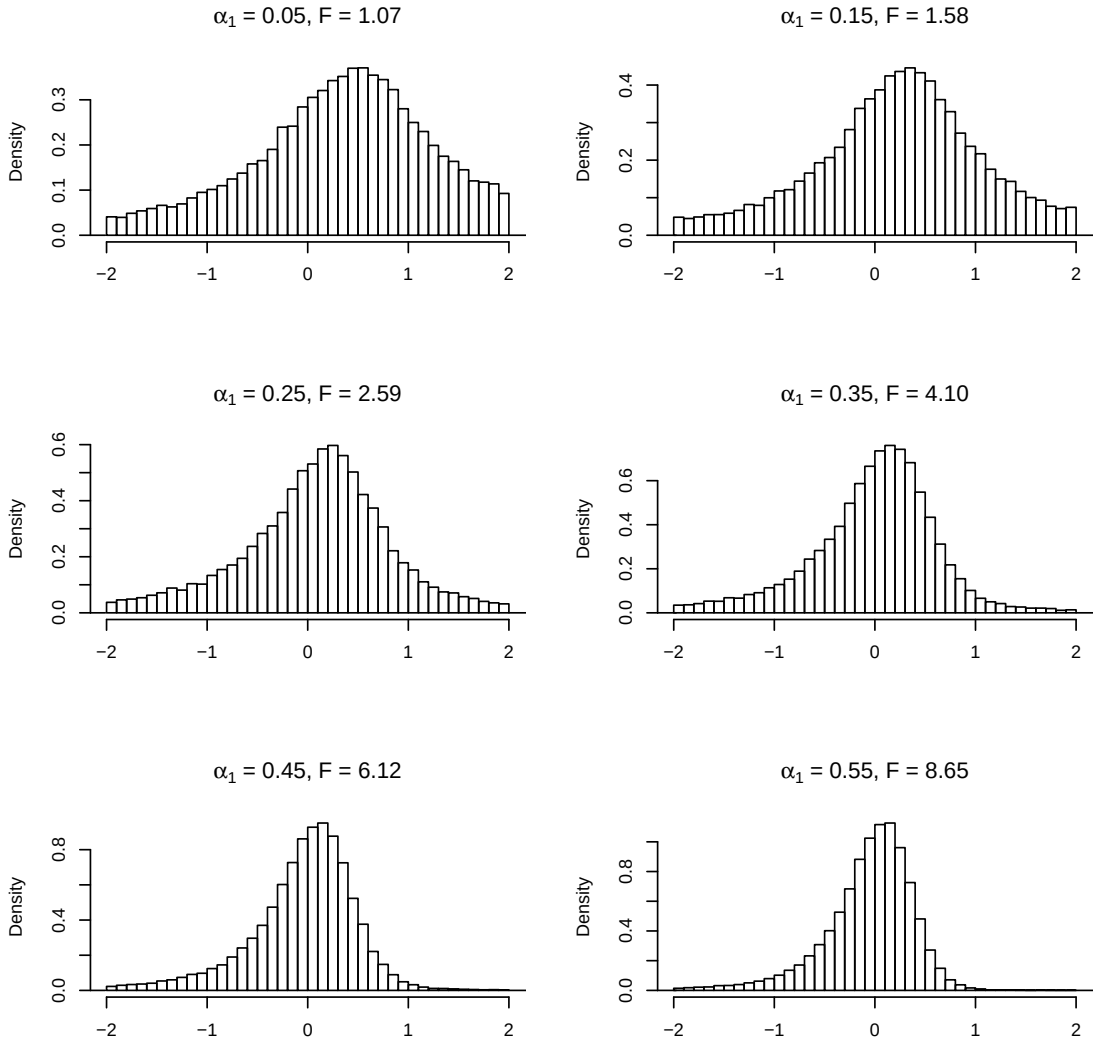


Figure 3.1: Histograms of IV estimates of causal association $\beta_1 = 0$ using weak instruments from simulated data. Average F statistics for each value of α_1 are shown

If $\bar{u}_j$ is the average confounder level for genotypic subgroup j, an expression for the causal association from the ratio method is

$$\beta_1^R = \beta_1 + \frac{\beta_2 \, \Delta u}{\alpha_1 + \alpha_2 \, \Delta u} \tag{3.3}$$

where $\Delta u = \bar{u}_1 - \bar{u}_0$ is normally distributed with expectation zero. When the instrument is strong, α_1 is large compared to $\alpha_2 \, \Delta u$, and the expression β_1^R will be close to β_1. When the instrument is weak, α_1 may be small compared to $\alpha_2 \, \Delta u$, and the bias $\beta_1^R - \beta_1$ is close to $\frac{\beta_2}{\alpha_2}$, which is the bias of the confounded observational association. This is true whether Δu is positive or negative. Figure 3.2 (left panel), reproduced from Nelson and Startz (160), shows how the IV estimate bias varies with Δu. Although for any non-zero α_1 the IV estimator will be an asymptotically consistent estimator as sample size increases and $\Delta u \to 0$, a bias in the direction of the confounded association will be present in finite samples. From Figure 3.2 (left panel), we can see that the median bias will be positive, as the bias is positive when $\Delta u > 0$ or $\Delta u < -\frac{\alpha_1}{\alpha_2}$, which happens with probability greater than 0.5. When the instrument is weak, the IV is measuring not the systematic genetic variation in the phenotype, but the chance variation in the confounders (101). If there is independent error in x and y, then the picture is similar, but more noisy, as seen in Figure 3.2 (right panel). Under model (3.1), the expression for the IV estimator is

$$\beta_1^R = \beta_1 + \frac{\beta_2 \, \Delta u + \Delta \epsilon_y}{\alpha_1 + \alpha_2 \, \Delta u + \Delta \epsilon_x}$$

where $\Delta \epsilon_x = \bar{\epsilon}_{x1} - \bar{\epsilon}_{x0}$ and $\Delta \epsilon_y = \bar{\epsilon}_{y1} - \bar{\epsilon}_{y0}$ defined analogously to Δu above.

This also explains the heavier negative tail in the histograms in Figure 3.1. The estimator takes extreme values when the denominator $\alpha_1 + \alpha_2 \, \Delta u$ is close to zero. Taking parameters α_1, α_2 and β_2 as positive, as in the example of Section 3.2, this is associated with a negative value of Δu, where the numerator of the ratio estimator will be negative. As Δu has expectation zero, the denominator is more likely to be small and positive than small and negative, giving more negative extreme values of β_R than positive ones.

3.3.2 Finite sample violation of IV assumptions

Alternatively, we can think of the bias as a violation of the first IV assumption in a finite sample. Although a valid instrument will be asymptotically independent from all confounders, in a finite sample there will be a non-zero correlation between the instrument and confounders. As before, this correlation biases the IV estimator towards the confounded association.

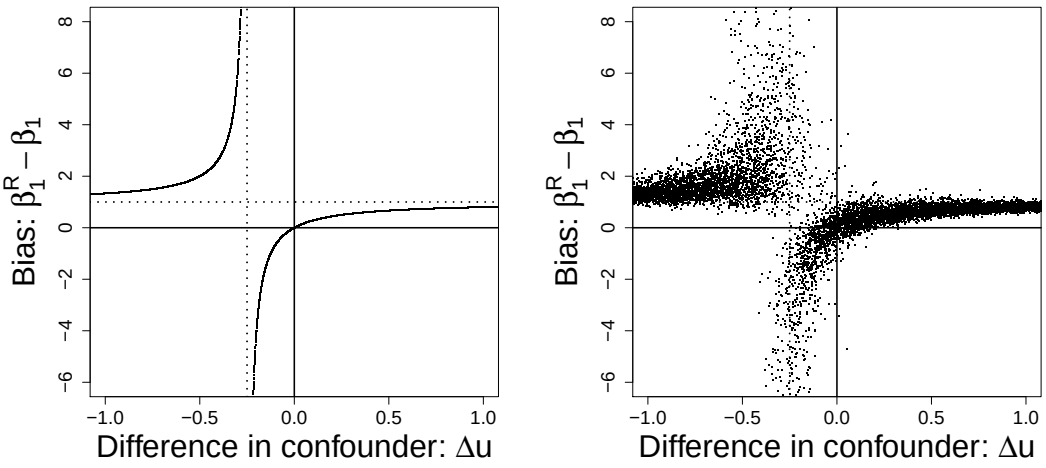


Figure 3.2: Bias in IV estimator as a function of the difference in mean confounder between groups ($\alpha_1 = 0.25$, $\alpha_2 = \beta_2 = 1$). Horizontal dotted line is at the confounded association $\frac{\beta_2}{\alpha_2}$, and the vertical dotted line at $\Delta u = -\frac{\alpha_1}{\alpha_2}$ where β_1^R is not defined. Left panel: no independent error in x or y, right panel: $\Delta\epsilon_x, \Delta\epsilon_y \sim \mathcal{N}(0, 0.1^2)$ independently.

If the instrument is strong, then the difference in phenotype between subgroups will be due to the genetic instrument, and the difference in outcome (if any) will be due to this difference in phenotype. However if the instrument is weak, that is it explains little variation in the phenotype, the chance difference in confounders may explain more of the difference in phenotype between subgroups than the instrument. If the effect of the instrument is near zero, then the estimate of the "causal association" approaches the association between phenotype and outcome caused by changes in the confounders, that is the observational confounded association (101). This shows that even stochastic (i.e. non-systematic) violation of the IV assumptions causes bias.

3.3.3 Sampling variation within genotypic subgroups

Finally, we can explain the bias graphically. We take model (3.1) with a negative causal association between phenotype and outcome ($\beta_1 = -0.4$), but with positive confounding ($\alpha_2 = 1, \beta_2 = 1, \sigma_x^2 = \sigma_y^2 = 0.2, \sigma_u^2 = 1$) giving a strong positive observational association between phenotype and outcome. We performed 1000 simulations with 600 subjects divided equally into 3 genotypic groups ($g_i \in \{0, 1, 2\}$). We took $\alpha_1 = (0.5, 0.2, 0.1, 0.05)$, corresponding to mean F values of (100, 16, 4.7, 2.0). The mean levels of phenotype and outcome for each genotypic group are plotted (Figure 3.3), giving simulated density functions for each group. In each simulation, we effectively draw one point at random from

each of these distributions; the gradient of the line through these three points is the 2SLS IV estimate. When the instrument is strong, the large phenotypic differences between the groups due to genotypic variation will generally lead to estimating a negative effect of phenotype on outcome, whereas when the instrument is weak the phenotypic differences between the groups due to genetic variation are small and the original confounded positive association is more likely to be recovered.

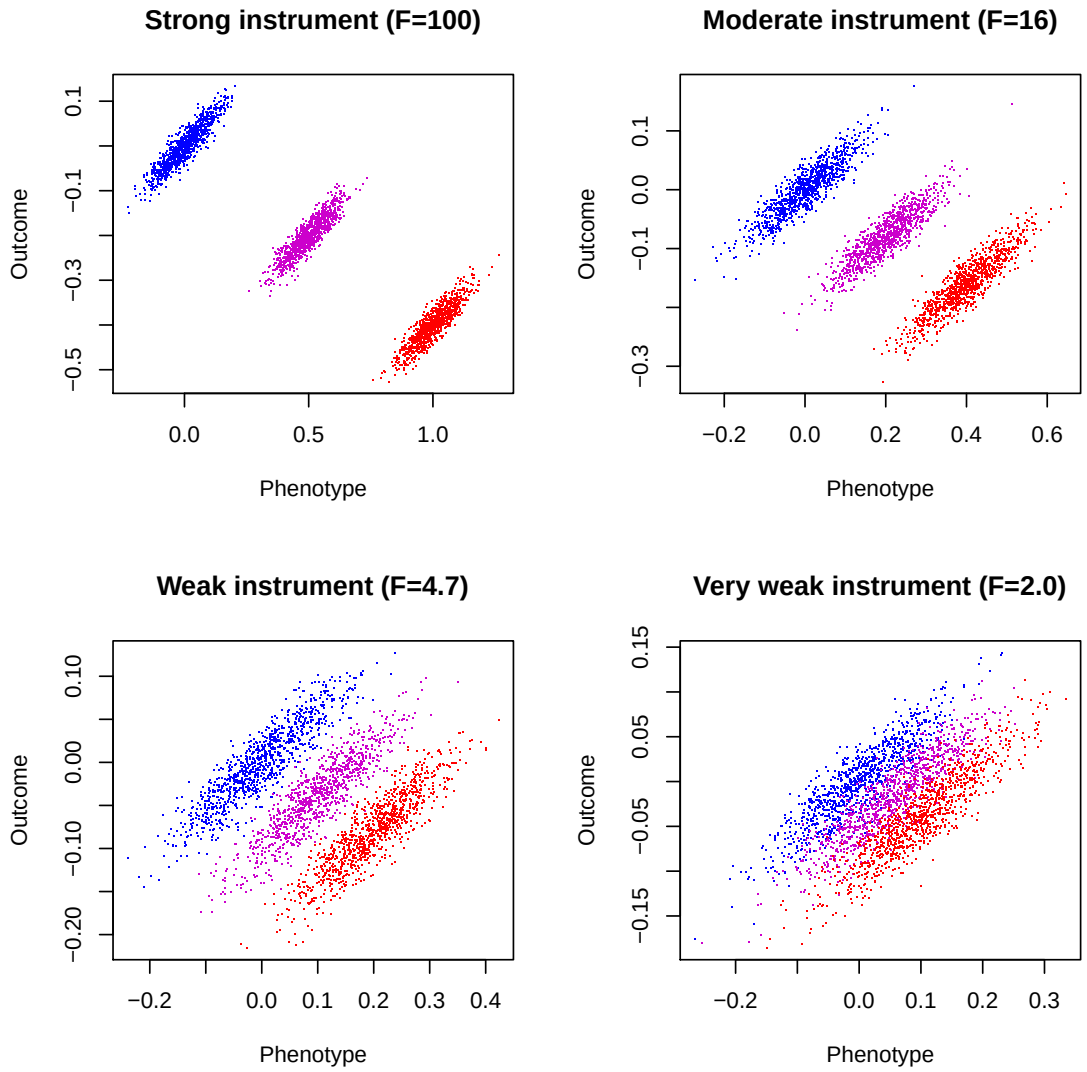


Figure 3.3: Distribution of mean outcome and mean phenotype level in three genotypic groups for various strengths of instrument

In summary, weak instrument bias reintroduces the problem that IVs were developed to solve. Weak instruments may convince a researcher that the observational association which they have measured is a causal association (101). The reason for the bias is that

the variation in the phenotype explained by the IV is too small compared to the variation in the phenotype caused by chance correlation between the IV and confounders.

3.4 Quantifying the bias from IV estimators

To get an idea as to whether the bias demonstrated and explained above is of sufficient magnitude to be a practical concern, we present simulations with parameters similar to what might be expected in a Mendelian randomization study, and examine the bias in the causal estimate. As discussed in Section 2.13, we consider the relative mean bias B, which is the ratio of the bias of the IV estimator ($\hat{\beta}_{IV}$) to the bias of the confounded association ($\hat{\beta}_{OBS}$) found by linear regression of Y on X:

$$B = \frac{\mathbb{E}(\hat{\beta}_{IV}) - \beta_1}{\mathbb{E}(\hat{\beta}_{OBS}) - \beta_1} \tag{3.4}$$

The relative mean bias from the 2SLS method is asymptotically approximately equal to $1/F$, where F is the expected F statistic in the regression of X on G (102). Both with a single instrument and with the LIML method, the mean of the IV estimator is not defined, so to compare bias in this setting, we instead consider the relative median bias. This is a novel measure formed by replacing the expectations in equation (3.4) with medians across simulations (118).

$$B^* = \frac{\text{median}(\hat{\beta}_{IV}) - \beta_1}{\text{median}(\hat{\beta}_{OBS}) - \beta_1} \tag{3.5}$$

3.4.1 Simulation of 2SLS bias with different strengths of 1 and 3 IVs

To investigate the size of the bias when there are few instruments, we take both model (3.1) with one genetic variable and a similar model except with three genetic variables g_1, g_2 and g_3:

$$\begin{aligned} x_i &= \sum_{k=1}^{3} \alpha_{1k}\, g_{ik} + \alpha_2\, u_i + \epsilon_{xi} \\ y_i &= \beta_1\, x_i + \beta_2\, u_i + \epsilon_{yi} \\ u_i &\sim \mathcal{N}(0, \sigma_u^2); \epsilon_{xi} \sim \mathcal{N}(0, \sigma_x^2); \epsilon_{yi} \sim \mathcal{N}(0, \sigma_y^2) \text{ independently} \end{aligned} \tag{3.6}$$

In model (3.6), each IV is taken as dichotomous, giving 8 possible genotype combinations. We simulated 100 000 datasets from this model for each set of parameters with 200 individuals divided equally between the 8 genotypic subgroups, meaning that the instruments

are uncorrelated. Model (3.1) was treated similarly, except the 200 individuals were divided into 2 genotypic subgroups. We considered four scenarios covering a range of typical situations, with $\sigma_x^2 = \sigma_y^2 = \sigma_u^2 = 1$ throughout:

a) null causal effect, moderate positive confounding ($\beta_1 = 0, \alpha_2 = 1, \beta_2 = 2$);

b) null causal effect, strong positive confounding ($\beta_1 = 0, \alpha_2 = 1, \beta_2 = 4$);

c) negative causal effect, moderate positive confounding ($\beta_1 = -1, \alpha_2 = 1, \beta_2 = 2$);

d) negative causal effect, moderate negative confounding ($\beta_1 = -1, \alpha_2 = 1, \beta_2 = -2$).

We took six values of $\alpha_1 = \alpha_{11} = \alpha_{12} = \alpha_{13}$ from 0.1 to 0.6, corresponding to different strengths of instrument with mean $F_{3,196}$ and $F_{1,198}$ values from 1.3 to 10.1. For each sample we calculated the IV estimator $\hat{\beta}_{IV}$ using the 2SLS method, and the confounded estimate $\hat{\beta}_{OBS}$ by linear regression.

Table 3.3 shows how the relative mean and median bias across simulations vary for different strengths of instrument. For three IVs, especially for stronger instruments, the relative median bias is larger than the relative mean bias. This is because the IV estimator has a negatively skewed distribution, as shown in Figure 3.1, and the skewness is more marked as the instrument becomes stronger. We can see that $1/F$ seems to be a good, if slightly conservative, estimate for the relative median bias, agreeing with Staiger and Stock (102). A mean F statistic of 10 would on average limit the IV estimator bias to 10% of the bias of the confounded association. For a single IV, the relative median bias is lower than for three IVs, and substantially so for stronger instruments. Although the distribution of the IV estimator for a single instrument is skew and heavy-tailed, the relative median bias is around 5% or less for even fairly weak instruments with $F = 5$ (118).

Although these results show that for a mean F value of 10 we have a relative median bias of less than 10%, there is no guarantee that if we have observed an F statistic of 10 or greater from data that the mean value is 10 or greater. From Table 3.2, for a mean F value of 4.10, we observe an F value greater than 10 in 8% of simulations, and for a mean F value of 6.12 in 18% of simulations.

3.4.2 Comparison of bias using different IV methods

There are several methods for calculating IV estimates, some of which are more robust to weak instruments than others. We here comment on the 2SLS, limited information

α_1	0.1	0.2	0.3	0.4	0.5	0.6
Mean F statistic	1.26	2.02	3.29	5.06	7.33	10.1
$1/F$	0.79	0.49	0.30	0.198	0.136	0.099
a) Null causal effect, moderate positive confounding						
Relative mean bias with 3 IVs	0.78	0.42	0.19	0.101	0.063	0.044
Relative median bias with 3 IVs	0.79	0.46	0.26	0.163	0.112	0.080
Relative median bias with 1 IV	0.78	0.38	0.15	0.038	0.010	0.002
b) Null causal effect, strong positive confounding						
Relative mean bias with 3 IVs	0.79	0.41	0.19	0.099	0.062	0.042
Relative median bias with 3 IVs	0.79	0.46	0.26	0.162	0.109	0.079
Relative median bias with 1 IV	0.77	0.38	0.14	0.041	0.011	0.002
c) Negative causal effect, moderate positive confounding						
Relative mean bias with 3 IVs	0.79	0.42	0.20	0.100	0.061	0.042
Relative median bias with 3 IVs	0.79	0.47	0.27	0.164	0.110	0.081
Relative median bias with 1 IV	0.77	0.40	0.15	0.040	0.011	0.002
d) Negative causal effect, moderate negative confounding						
Relative mean bias with 3 IVs	0.79	0.41	0.19	0.098	0.062	0.044
Relative median bias with 3 IVs	0.80	0.46	0.26	0.160	0.110	0.081
Relative median bias with 1 IV	0.80	0.39	0.15	0.035	0.011	-0.000

Table 3.3: Relative mean and median bias of the 2SLS IV estimator across 100 000 simulations for different strengths of instrument using three IVs and one IV. Mean $F_{3,196}$ and $F_{1,198}$ statistics are equal to 2 decimal places

maximum likelihood (LIML) (27) and the Fuller(1) methods (176) as they can be calculated using the *ivreg2* command in Stata and have different finite moments properties with various numbers of instruments (117).

The LIML estimator is close to median unbiased for all but the weakest instrument situations (102; 177). With one IV, the estimate from LIML coincides with that from the ratio and 2SLS methods. However, Hahn, Hausman and Kuersteiner (133) strongly discourage the use of the LIML estimator, as it does not have defined moments for any number of instruments, as opposed to the 2SLS estimator, which has a finite variance when there are three or more instruments. The 2SLS method, when all the associations are linear and the error terms normally distributed, has a finite kth moment when the number of instruments is at least $(k+1)$ (124). The Fuller(1) estimator is an adaption of the LIML method (133), which again has better weak instrument properties than 2SLS (92), and is designed to have all moments, even with a single instrument (92; 177).

To investigate the bias properties of these methods, we conduct a simulation using

the same parameters as in Section 3.4.1, analysing 100 000 simulations with 1 and 3 instruments using the 2SLS, LIML and Fuller(1) methods for instruments with $\alpha_1 = 0.2, 0.4, 0.6$. Table 3.4 shows how, with three IVs, the median bias is close to zero for LIML with instruments with mean F statistic greater than 5, whereas it is large for the 2SLS and Fuller(1) methods. For instruments with F close to 10, the mean bias of the Fuller(1) estimator is close to zero. With one IV, as before, the 2SLS / LIML estimator is approximately median unbiased with a mean F statistic of 10, whereas the Fuller(1) estimate still shows considerable median and mean bias with a mean F statistic of 10.

This simulation shows a trade-off amongst IV methods between asymptotic and finite sample properties. The LIML method performs best overall in terms of median bias, even though mean bias is always undefined. However, methods with finite mean bias perform badly in terms of median bias. Although absence of a finite mean presents serious theoretical problems in the comparison of bias, it would seem to be more of a mathematical curiosity than a practical problem. Extreme values of the causal estimate would generally be discarded due to implausibility and finite-sample near violation of the first IV assumption (non-zero G-X association) in the dataset.

3.5 Choosing a suitable IV estimator

Including more instruments, where each instrument explains extra variation in the phenotype, should give more information on the causal parameter. However as shown above, bias may increase, due to the weakening of the set of instruments. In this section, we consider the impact of choice of instrument on the bias of the IV estimator.

3.5.1 Multiple candidate IVs

In order to investigate how using more instruments affects bias in the IV estimator, we perform 100 000 simulations in a model where, for each participant indexed by i, the phenotype x_i depends linearly on six dichotomous genetic instruments ($g_{ik} = 0$ or $1, k = 1, \ldots, 6$), a normally distributed confounder u_i, and an independent normally distributed error term ϵ_{xi}. Outcome y_i is a linear combination of phenotype, confounder, and an independent error term ϵ_{yi}.

$$\begin{aligned} x_i &= \sum_{k=1}^{6} \alpha_{1k}\, g_{ik} + \alpha_2\, u_i + \epsilon_{xi} \\ y_i &= \beta_1\, x_i + \beta_2\, u_i + \epsilon_{yi} \\ u_i &\sim \mathcal{N}(0, \sigma_u^2); \epsilon_{xi} \sim \mathcal{N}(0, \sigma_x^2); \epsilon_{yi} \sim \mathcal{N}(0, \sigma_y^2) \text{ independently} \end{aligned} \tag{3.7}$$

		3 IVs			1 IV [1]		
		$\alpha_1 = 0.2$	$\alpha_1 = 0.4$	$\alpha_1 = 0.6$	$\alpha_1 = 0.2$	$\alpha_1 = 0.4$	$\alpha_1 = 0.6$
Mean F statistic		2.02	5.06	10.1	2.02	5.06	10.1
a) Null causal effect, moderate positive confounding							
Median bias	2SLS	0.4556	0.1526	0.0702	0.3872	0.0401	0.0022
	LIML	0.1617	0.0033	-0.0017			
	Fuller(1)	0.3888	0.0858	0.0353	0.7324	0.2863	0.1159
Mean bias [2]	2SLS	0.4129	0.0935	0.0374	-	-	-
	Fuller(1)	0.4248	0.0338	0.0006	0.7091	0.2661	0.0673
b) Null causal effect, strong positive confounding							
Median bias	2SLS	0.9081	0.3121	0.1414	0.7692	0.0850	0.0041
	LIML	0.2899	0.0127	-0.0004			
	Fuller(1)	0.7675	0.1757	0.0718	1.4530	0.5678	0.2260
Mean bias	2SLS	0.8217	0.1916	0.0761	-	-	-
	Fuller(1)	0.8376	0.0721	0.0034	1.4235	0.5347	0.1297
c) Negative causal effect, moderate positive confounding							
Median bias	2SLS	0.4571	0.1531	0.0715	0.3908	0.0399	0.0019
	LIML	0.1601	0.0028	0.0014			
	Fuller(1)	0.3915	0.0858	0.0376	0.7319	0.2860	0.1145
Mean bias	2SLS	0.4096	0.0927	0.0391	-	-	-
	Fuller(1)	0.4223	0.0339	0.0030	0.7083	0.2682	0.0643
d) Negative causal effect, moderate negative confounding							
Median bias	2SLS	-0.4555	-0.1545	-0.0706	-0.3842	-0.0413	-0.0020
	LIML	-0.1580	-0.0035	0.0004			
	Fuller(1)	-0.3858	-0.0862	-0.0360	-0.7297	-0.2882	-0.1158
Mean bias	2SLS	-0.4076	-0.0930	-0.0385	-	-	-
	Fuller(1)	-0.4214	-0.0339	-0.0019	-0.7086	-0.2694	-0.0659

Table 3.4: Median and mean bias across 100 000 simulations using 2SLS, LIML and Fuller(1) methods for a range of strength of three IVs and one IV

[1] With 1 IV, the estimates from 2SLS and LIML coincide.

[2] Mean bias is reported only when it is not theoretically infinite

We set $\beta_1 = 0, \alpha_2 = 1, \beta_2 = 1, \sigma_x^2 = \sigma_y^2 = \sigma_u^2 = 1$ so that X is observationally strongly positively associated with Y, but there is a null causal association. We take parameters for the genetic association $\alpha_{1k} = 0.4$ for each genetic instrument k, corresponding to a mean F value of 10.2 (quartiles 5.8, 9.3, 13.7). We used a sample size of 512 divided equally between the $2^6 = 64$ genotypic subgroups. The instruments are uncorrelated, so that variation explained by each of the instruments is independent, and the mean F values do not depend greatly on the number of IVs (mean 10.2 using 1 IV, 11.3 using 6 IVs).

Table 3.5 shows the median and 95% range of the estimates of bias from the 2SLS and LIML methods and the mean bias for the 2SLS method using all combinations of all numbers of IVs as the instrument, with the mean across simulations of the F statistic for all the instruments used. We also give results using the IV with the greatest and lowest observed F values in each simulation, as well as using all IVs with an F statistic greater than 10 in univariate regression of phenotype on each IV.

Using 2SLS, as the number of instruments increases, while the variance of estimates decreases the bias increases, despite the mean F value remaining fairly constant. This is because there is a greater risk of imbalances in confounders between the greater number of genotypic subgroups defined by the instruments. The data are being subdivided in more different ways, and so there is more chance of one of these divisions giving genotypic groups with different average levels of confounders. However, the more instruments that are used, the smaller the variability of the IV estimator. This is because a greater proportion of the variance in the phenotype is being modelled.

The greatest increase in median bias is from one instrument to two instruments, and coincides with the greatest increase in precision. With LIML, a similar increase in precision is observed, but no increase in bias. For 2SLS, the mean bias is similar to the median presented, except that mean bias is close to zero with two IVs, increasing steadily as the number of instruments increases. In the case of a single IV, the theoretical mean is infinite (101). For LIML, the mean bias is infinite for all numbers of IVs (133).

Using the single IV with the greatest observed F gives markedly biased results, despite a mean F value of 23.9. There is a similar bias only using IVs with $F > 10$. In the simulation, each IV in truth explains the same amount of variation in the phenotype. If however the IVs used are chosen because they explain a large proportion of the variation in the phenotype in the data under analysis, then the estimate using these IVs is additionally biased. This is because the IVs explaining the most variation will be overestimating the proportion of true variation explained, due to chance correlation with confounders overestimating the underlying difference between genotypic groups in both phenotype and outcome, leading to an overestimate of the causal association in the direction of the

confounded observational association. In the notation of Section 3.3.1, Δu is large and, having the same sign as α_1, leads to an estimate biased in the direction of $\frac{\beta_2}{\alpha_2}$. Similarly, if the IV with the least F statistic is used as an instrument, the IV estimator will be biased in the opposite direction to the observational association. These characteristics are evident in the simulations (Table 3.5). A commonly used rule for the validity of an IV is that the observed F statistic is greater than 10. However, if this rule is used to choose between instruments, this rule itself introduces a selection bias (178).

We therefore have a situation analogous to a bias–variance trade-off (26). As an alternative to the mean squared error, we suggest using the median absolute bias (MAB) (median $|\hat{\beta}_{IV} - \beta_1|$) as a criterion for how many instruments should be used. Table 3.5 shows that in this case, despite the increase in the bias, the 2SLS estimate using all six IVs is preferred. However, naive use of the MAB as a criterion for choosing between estimators would seem unwise, as the MAB is less for the estimator using the single SNP with the greatest observed F statistic than for choosing a single SNP at random, despite the increase in median bias from the selection effect.

	Median	2.5% to 97.5% quantiles			Mean bias [1]	MAB	Mean F
	2SLS		LIML		2SLS		statistic
Estimate using 1 IV	0.0001	-1.1151 to 0.5345			-	0.2130	10.2
2 IVs	0.0239	-0.5380 to 0.3947	-0.0003	-0.6383 to 0.3900	-0.0002	0.1472	10.4
3 IVs	0.0312	-0.3871 to 0.3342	-0.0004	-0.4801 to 0.3233	0.0165	0.1205	10.6
4 IVs	0.0346	-0.3109 to 0.2982	-0.0003	-0.3961 to 0.2833	0.0241	0.1051	10.8
5 IVs	0.0367	-0.2633 to 0.2731	-0.0004	-0.3430 to 0.2552	0.0284	0.0948	11.0
6 IVs	0.0378	-0.2294 to 0.2552	-0.0003	-0.3055 to 0.2344	0.0312	0.0875	11.3
IV with greatest F	0.1419	-0.2988 to 0.5206			-	0.1777	23.9
IV with least F	-0.3208	-2.5742 to 0.5795			-	0.3956	6.7
IVs with F > 10	0.1114	-0.2032 to 0.3919	0.0989	-0.2204 to 0.3895	0.1071	0.1304	16.4

Table 3.5: Median and 95% range of bias using 2SLS and LIML methods, mean bias and median absolute bias (MAB) using 2SLS method and mean F statistic across 100 000 simulations using combinations of six uncorrelated instruments

[1] Mean bias is reported only when it is not theoretically infinite

3.5.2 Overidentification

When multiple instruments are used, a common econometric tool is an overidentification test (117), such as the Sargan test (158). This is a test for incompatibility of estimates based on different instruments, and can be used to test validity of the IV assumptions in a dataset. While this can be useful in indicating possible bias from violation of the

underlying IV assumptions, it does not identify bias from the finite-sample violation of the IV assumptions due to weak instruments. For the data summarized in Table 3.5 using all six IVs, 7% of the simulations failed the Sargan test at $p < 0.05$, slightly more than would be expected with a valid instrument. While the median estimate from 2SLS using all six IVs in simulations which failed the Sargan test was 0.0789, the median estimate in simulations which passed the Sargan test was 0.0345, close to the overall median of 0.0378. Overidentification tests are omnibus tests, where the alternative hypothesis includes failure of IV assumptions for one IV, failure for all IVs, and non-linear association between phenotype and outcome. Hence, while the test can recognize problems with the model, it has limited use to combat weak instruments.

3.5.3 Multiple instruments in the Framingham Heart Study

As a further illustration, we consider the Framingham Heart Study (FHS), a cohort study measuring CRP and fibrinogen at baseline with complete data for nine SNPs on the CRP gene for 1500 participants. The observational estimate of the log(CRP)–fibrinogen (μmol/l) association is 1.134 (95% CI 1.052 to 1.217). We calculate the causal estimate of the association using the 2SLS method with different numbers of SNPs as an instrument. Figure 3.4 shows a plot of the 2SLS IV estimates against number of instruments, where each point represents the causal estimate calculated using the 2SLS method with a different combination of SNPs. The range of point estimates of the causal association reduces as we include more instruments, but the median causal estimate across the different combinations of IVs increases. The 2SLS estimate using all 9 SNPs in an additive per allele model is -0.005 (95% CI: -0.721 to 0.711, $p = 0.99$, $F_{9,1490} = 3.34$). If we relax the genetic assumptions of a per-allele model and additivity between SNPs to instead use a model with one coefficient for each of the 49 genotypes represented in the data, the 2SLS estimate is 0.792 (95% CI 0.423 to 1.161, $p = 0.00003$, $F_{48,1451} = 1.66$). Using LIML, the estimate is 0.052 (95% CI -0.706 to 0.809, $p = 0.89$). This illustrates the bias in the 2SLS method due to the use of multiple instruments, showing how an estimate close to the observational association (1.134, 95% CI: 1.052 to 1.217) can be recovered by injudicious choice of instrument. The LIML method with 48 genetic parameters shows signs of some bias, but gives a substantially different answer to the 2SLS method, suggesting its possible use as a sensitivity analysis to the 2SLS method. In the extreme case, if each of the individuals in a study were placed into separate genetic groups, then the IV estimate would be the observational association.

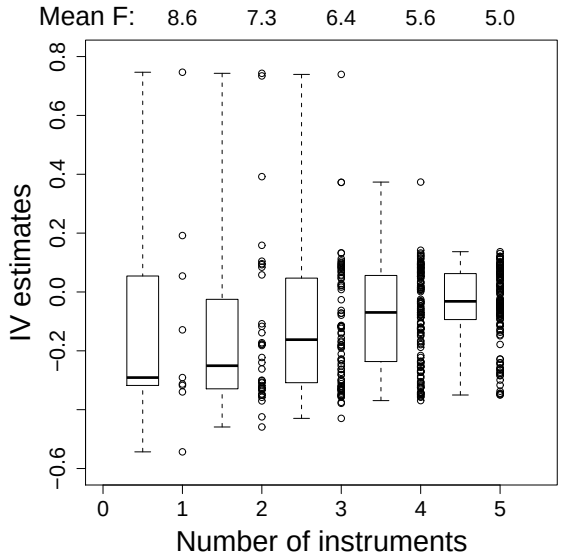


Figure 3.4: IV estimates for causal association in Framingham Heart Study of log(CRP) on fibrinogen (μmol/l) using all combinations of varying numbers of SNPs as instruments. Point estimates, associated box plots (median, inter-quartile range, range) and mean F statistics across combinations are displayed

3.5.4 Model of genetic association

As the magnitude of weak instrument bias depends on the F statistic, models for the G-X association which give larger F statistics would be preferred. A model of genetic association with one parameter per SNP (for example a dominant/recessive model or per-allele model) will typically have a greater F statistic than a model with a separate coefficient for each level of the SNP (here called a categorical model). However, if the simpler model does not represent the true model under which the data were generated, then bias due to model misspecification may be introduced.

To investigate this we modify model (3.6) with three instruments and model (3.1) with one instrument, so that the genetic association is not necessarily additive:

$$x_i = \sum_{k=1}^{K} (\alpha_{1k}\, g_{ik} + d_k\, 1(g_{ik} = 2)) + \alpha_2\, u_i + \epsilon_{xi} \tag{3.8}$$

where $1(.)$ is the indicator function, $g_{ik} \in \{0, 1, 2\}$ for all i, k and $K = 1$ or 3. We conducted 100 000 simulations using the same parameters as Section 3.5.1, with α_{1k} fixed at 0.5 for all k and the dominance parameter d_k taking values 0 (true additive model), +0.2 (major dominant model) and −0.2 (minor dominant model). With three instruments, the genetic instruments divide the chosen population of size 243 into 27 equally sized subgroups.

With one instrument, the population is divided into subgroups of size 108, 108 and 27, corresponding to a SNP with minor allele frequency $\frac{1}{3}$.

	3 IVs			1 IV		
	Median bias	MAB	Mean F	Median bias	MAB	Mean F
Additive: categorical	0.038	0.087	11.2	0.056	0.218	7.8
Per-allele	0.016	0.086	21.4	0.000	0.228	14.6
Major dominant: categorical	0.027	0.072	15.9	0.045	0.200	9.9
Per-allele	0.011	0.072	30.3	0.000	0.207	18.5
Minor dominant: categorical	0.056	0.109	7.7	0.068	0.240	6.3
Per-allele	0.024	0.108	14.0	0.002	0.253	11.3

Table 3.6: Median bias and median absolute bias (MAB) of 2SLS IV estimate of $\beta_1 = 0$ and mean F statistic across 100 000 simulations using per-allele and categorical modelling assumptions for true additive, major dominant and minor dominant models

We analysed the data generated assuming additivity between instruments and either a per-allele model (1 instrument per SNP) or a categorical model (2 instruments per SNP) for each IV. Table 3.6 shows that the per-allele model has lower median bias than the categorical model even when the underlying genetic model is misspecified. The median absolute bias (MAB) is similar in each model, with a slight preference for the categorical model with a single instrument. The categorical model suffers from greater weak instrument bias because the mean F statistic is smaller. This indicates that, where the genetic model is approximately additive, the more parsimonious per-allele model should be preferred over a categorical model, as the gain in precision would not seem to justify the increase in bias.

3.6 Minimizing the bias from IV estimators

We continue by listing specific ways to minimize bias from weak instruments in the design and analysis of Mendelian randomization studies.

3.6.1 Increasing the F statistic

The F statistic is related to the proportion of variance in the phenotype explained by the genetic variants (R^2), sample size (N) and number of instruments (K) by the formula $F = (\frac{N-K-1}{K})(\frac{R^2}{1-R^2})$. As the F statistic depends on the sample size, then bias can be reduced by increasing sample size. Similarly, if there are instruments that are not contributing much

to explaining the variation in the phenotype, then excluding these instruments will increase the F value. As demonstrated in Section 3.5, in general, employing fewer degrees of freedom to model the genetic association, that is using parsimonious models, will increase the F statistic and reduce weak instrument bias, provided that the model does not misrepresent the data (44; 45).

However, as shown above, it is not enough to simply rely on an F statistic measured from data to inform us about bias (178). Returning to the example from Section 3.2.1 where we divided the CGPS study into 16 equally sized substudies with mean F statistic 10.81, Figure 3.5 shows the forest plot of the estimates of these 16 substudies using the 2SLS method with their corresponding F values. We see that the substudies which have greater estimates are the ones with higher F values. The correlation between F values and point estimates is 0.83 ($p < 0.001$). The substudies with higher F values also have tighter CIs and so receive more weight in the meta-analysis. If we exclude from the meta-analysis substudies with an F statistic less than 10, then the pooled estimate increases from 0.2300 (SE 0.1372, $p = 0.09$) to 0.4322 (SE 0.1574, $p = 0.006$). Equally, if we only use as instruments in each substudy the IVs with an F statistic greater than 10 when regressed in a univariate regression on the phenotype, then the pooled estimate increases to 0.2782 (SE 0.1470, $p = 0.06$). So neither of these approaches are useful in reducing bias.

Although the expectation of the F statistic is a good indicator of bias, with low expected F statistics indicating greater bias, the observed F statistic shows considerable variation. In the 16 substudies of Figure 3.5, the F statistic ranges from 3.4 to 22.6. In more realistic examples, assuming similar instruments in each study, larger studies would have higher expected F statistics due to sample size which would correspond to truly stronger instruments and less bias. However, the sampling variation of causal effects and observed F statistics in each study would still tend to follow the pattern of Figure 3.5, with larger observed F statistics corresponding to more biased causal estimates.

So while it is desirable to use strong instruments, the measured strength of instruments in data is not a good guide to the true instrument strength. As also demonstrated in Section 3.5.1 for the choosing of IVs, any guidance that relies on providing a threshold (such as $F > 10$) for choosing which instruments to use or as an inclusion criterion for a meta-analysis, is flawed and may introduce more bias than it prevents.

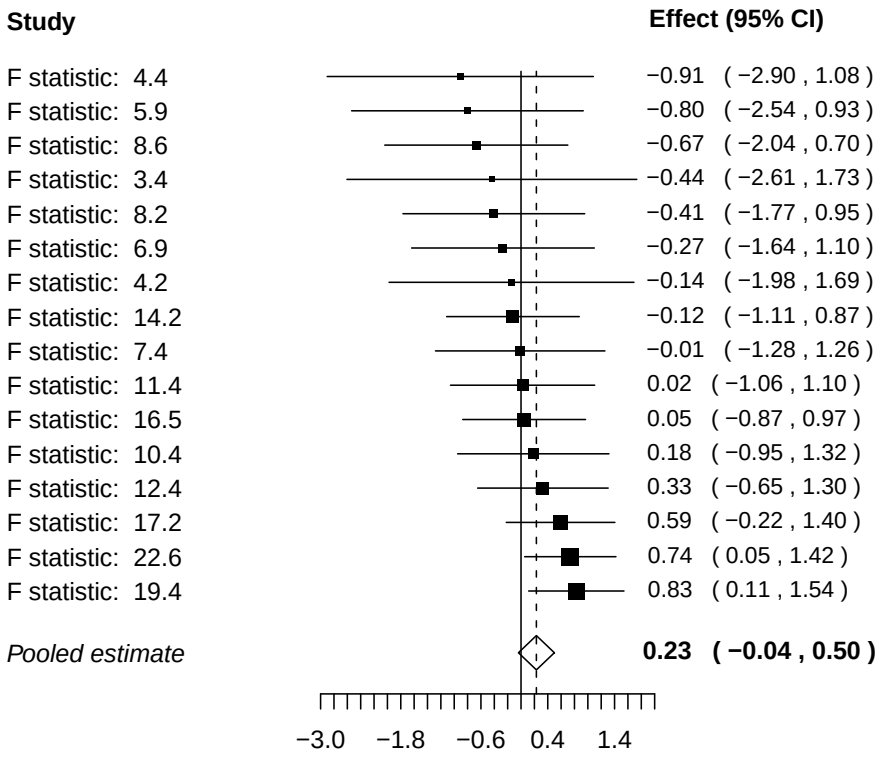


Figure 3.5: Forest plot of causal estimates of log(CRP) on fibrinogen (μmol/l) using data from Copenhagen General Population Study divided randomly into 16 equally sized sub-studies (each $N \simeq 2230$). Studies ordered by causal estimate. F statistic from regression of phenotype on three IV. Size of markers is proportional to weight in a fixed-effect meta-analysis

3.6.2 Adjustment for measured covariates

If we can find measured covariates which explain variation in the phenotype, and which are not on the causal pathway between phenotype and outcome, then we can incorporate such covariates in our model. This will increase precision and reduce weak instrument bias. Precision will be further increased if these covariates can be used to explain variation in the outcome.

To exemplify this, we perform 100 000 simulations in a model similar to (3.1) with a single IV, but with two separate terms accounting for confounding between X and Y, corresponding to measured (V) and unmeasured (U) confounders.

$$\begin{aligned} x_i &= \alpha_1 g_i + \alpha_2 u_i + \alpha_2 v_i + \epsilon_{xi} \\ y_i &= \beta_1 x_i + \beta_2 u_i + \beta_2 v_i + \epsilon_{yi} \\ u_i, v_i, \epsilon_{xi}, \epsilon_{yi} &\sim \mathcal{N}(0,1) \text{ independently} \end{aligned} \tag{3.9}$$

We again set $\beta_1 = 0, \alpha_2 = 1, \beta_2 = 1$ and vary the parameter for the genetic association α_1 from 0.05 to 0.55, corresponding to mean F values from 1.05 to 6.11. We use a sample size of 200 equally divided between two genotypic groups, $g_i = 0, 1$. We calculate an estimate of causal association from the 2SLS method, both with and without adjustment for V in the G-X and $\hat{X}$-Y regressions. R^2 in the regression of X on V is 33%. The relevant measure of instrument strength with a measured confounder is the partial F statistic for G in the regression of X on G and V (156). Table 3.7 shows that adjustment for measured covariates increases the F statistic and decreases the median bias of the IV estimator. For stronger instruments, we also see a reduction in the variability of the estimator.

	Not adjusted			Adjusted		
α_1	Mean F	Median bias	IQ range	Partial F	Median bias	IQ range
0.05	1.05	0.6418	-0.1026 to 1.3859	1.58	0.4659	-0.3830 to 1.3138
0.15	1.39	0.4573	-0.2408 to 1.1406	2.09	0.2916	-0.4442 to 0.9776
0.25	2.06	0.2478	-0.3819 to 0.7446	3.09	0.1290	-0.4535 to 0.5949
0.35	3.08	0.1110	-0.4282 to 0.4821	4.62	0.0460	-0.4104 to 0.3883
0.45	4.42	0.0412	-0.4122 to 0.3414	6.63	0.0115	-0.3468 to 0.2819
0.55	6.11	0.0138	-0.3620 to 0.2691	9.16	0.0030	-0.2822 to 0.2277

Table 3.7: Bias of the IV estimator, median and interquartile (IQ) range across simulations from model (3.9), for different strengths of instrument without and with adjustment for confounder

As an example, we consider data on interleukin-6 (IL6), a cytokine which is involved in the inflammation process upstream of CRP and fibrinogen (179). Elevated levels of IL6 lead to elevated levels of both CRP and fibrinogen, so IL6 is correlated with short-term variation in CRP (84), but is independent of underlying genetic variation in CRP (64). We assume that it is a confounder in the association of CRP with fibrinogen and not on the causal pathway (if such a pathway exists). As IL6 has a positively skewed distribution, we take its logarithm. The Cardiovascular Health Study (CHS) is a cohort study measuring CRP, IL6 and fibrinogen at baseline, as well as 3 SNPs on the CRP gene, with complete data for 4137 subjects. The proportion of variation in log(CRP) explained in the data by log(IL6) is 26%. We calculate the causal estimate of the CRP-fibrinogen association for each SNP separately and for all the SNPs together in an additive per allele model, both without and with adjustment for log(IL6) in the first and second stage regressions. Results are given in Table 3.8. We see that after adjusting for log(IL6) the causal estimate in each case has decreased, its standard error has reduced, and the F statistic has increased. This indicates both that weak instrument bias has been reduced, and that precision has been improved.

	Not adjusted		Adjusted	
IV estimate	Estimate (SE)	F statistic	Estimate (SE)	Partial F
Using rs1205	0.219 (0.201)	79.6	0.173 (0.196)	100.2
Using rs1417938	-0.457 (0.407)	27.6	-0.458 (0.362)	37.2
Using rs1800947	0.354 (0.325)	28.6	0.324 (0.316)	36.5
Using all SNPs	0.186 (0.194)	24.4	0.127 (0.188)	32.2

Table 3.8: Estimate and standard error (SE) of IV estimator for causal effect of log(CRP) on fibrinogen and F statistic for regression of log(CRP) on IVs calculated using each SNP separately and all SNPs together in additive per allele model, without and with adjustment for log(IL6) in Cardiovascular Health Study

3.6.3 Borrowing information across studies

The IV estimator would be unbiased if we knew the true values for the average phenotype in different genotypic groups. In a meta-analysis context (71), we can combine the estimates of genotype–phenotype association from different studies to give more precise estimates of phenotype levels in each genetic group. In the 2SLS method, an individual participant data (IPD) fixed-effect meta-analysis for data on individual i in study m with phenotype

x_{im}, outcome y_{im} and g_{ikm} for number of alleles of genetic variant k ($k = 1, 2, \ldots K$) is:

$$\begin{aligned} x_{im} &= \alpha_{0m} + \sum_{k=1}^{K} \alpha_{km}\, g_{ikm} + \epsilon_{xim} \\ y_{im} &= \beta_{0m} + \beta_1\, \hat{x}_{im} + \epsilon_{yim} \\ \epsilon_{xim} &\sim \mathcal{N}(0, \sigma_x^2); \epsilon_{yim} \sim \mathcal{N}(0, \sigma_y^2) \text{ independently} \end{aligned} \tag{3.10}$$

The phenotype levels are regressed on the instruments using a per allele additive linear model separately in each study, and then the outcome levels are regressed on the fitted values of phenotype ($\hat{x}_{im}$). The terms α_{0m} and β_{0m} are study-specific intercept terms. Here we assume homogeneity of variances across studies; we can use generalized method of moments (GMM) (117) or Bayesian methods (140) (see Chapter 5) to allow for possible heterogeneity.

If the same genetic variants are measured and assumed to have the same effect on the phenotype in each study, we can use common genetic effects (ie. $\alpha_{km} = \alpha_k$) across studies by replacing the first line in model (3.10) with

$$x_{im} = \alpha_{0m} + \sum_{k=1}^{K} \alpha_k\, g_{ikm} + \epsilon_{xim} \tag{3.11}$$

where the coefficients α_k are the same in each study. If the assumption of fixed genetic effects is correct, this will improve the precision of the $\hat{x}_{im}$ and reduce weak instrument bias. Model (3.11) can be used even if, for example, the phenotype is not measured in one study, under the assumption that the data are missing at random (MAR) (180).

To illustrate, we consider the Copenhagen City Heart Study (CCHS), Edinburgh Artery Study (EAS), Health Professionals Follow-up Study (HPFS), Nurses Health Study (NHS), and Stockholm Heart Epidemiology Program (SHEEP), which are cohort studies or case-control studies measuring CRP and fibrinogen levels at baseline (82). In case-control studies, we use the data from controls alone since these better represent cross-sectional population studies. These five studies measured three SNPs: rs1205, rs1130864 and rs3093077 (or rs3093064, which is in complete linkage disequilibrium with rs3093077). We estimate the causal association using the 2SLS method with different genetic effects (model 3.10), common genetic effects (model 3.11) and by a fixed-effect meta-analysis of summary estimates from each study.

Table 3.9 shows that the studies analyzed separately have apparently disparate causal estimates with wide CIs. The meta-analysis estimate assuming common genetic effects across studies is further from the confounded observational estimate and closer to the

Study	N	Causal estimate	95% CI	F statistic	df	Observational estimate (SE)
CCHS	7999	-0.286	-1.017 to 0.445	29.6	(3,7995)	1.998 (0.030)
EAS	650	0.745	0.113 to 1.396	6.9	(3,646)	1.115 (0.056)
HPFS	405	0.758	-0.071 to 1.587	5.3	(3,401)	1.048 (0.081)
NHS	385	-0.906	-2.154 to 0.341	6.1	(3,381)	0.562 (0.114)
SHEEP	1044	0.088	-0.588 to 0.763	10.5	(3,1040)	1.078 (0.051)
Different genetic effects		0.021	-0.362 to 0.403	14.4	(15, 10463)	
Common genetic effects		-0.093	-0.534 to 0.348	56.6	(3, 10475)	
Summary estimates		0.234	-0.107 to 0.575			

Table 3.9: Estimates of effect of log(CRP) on fibrinogen (μmol/l) from each of five studies separately and from meta-analysis of studies: studies included, number of participants (N), causal estimates using 2SLS with 95% confidence interval (CI), F statistic with degrees of freedom (df) from additive per allele regression of phenotype on SNPs used as IVs, observational estimate (standard error). Fixed-effect meta-analyses conducted using individual participant data (IPD) with different study-specific genetic effects, common pooled genetic effects and using summary estimates with inverse-variance weighting

estimate from the largest study with the strongest instruments (CCHS) than the model with different genetic effects, suggesting that the latter suffers bias from weak instruments.

The estimate from meta-analysis of study-specific causal estimates is greater than that from meta-analysis using the individual participant data. Although the CCHS study has about 8 times the number of participants as SHEEP and 12 times as many as EAS, its causal estimate has a larger standard error. The standard errors in the 2SLS method, calculated by sandwich variance estimators using strong asymptotic assumptions, are known to be underestimated, especially with weak instruments (161). Also, Figure 3.5 shows that causal estimates nearer to the observational association have lower variance. So a meta-analysis of summary outcomes may be biased due to overestimated weights in the studies with more biased estimates.

In the example at the beginning of the chapter (Section 3.2.1), if we use the IPD data to combine the substudies in the meta-analysis rather than combining summary estimates, then comparing Table 3.10 to Table 3.1 shows that the pooled estimates are somewhat less biased. If we additionally assume common genetic effects across studies, then we recover close to the original estimate based on analyzing the full dataset as one study and weak instrument bias has been eliminated.

Substudies	Summary	p-value	IPD different genetic effects	p-value	IPD common genetic effects	p-value
1	-0.0468 (0.1510)	0.76				
5	-0.0092 (0.1478)	0.95	-0.0273 (0.1479)	0.85	-0.0473 (0.1511)	0.75
10	0.0871 (0.1426)	0.54	0.0370 (0.1430)	0.80	-0.0457 (0.1510)	0.76
16	0.2300 (0.1372)	0.09	0.1530 (0.1372)	0.26	-0.0482 (0.1512)	0.75
40	0.4562 (0.1266)	< 0.001	0.2986 (0.1272)	0.02	-0.0433 (0.1511)	0.77
100	0.8279 (0.1078)	< 0.001	0.6782 (0.1056)	< 0.001	-0.0450 (0.1506)	0.77
250	1.2711 (0.0826)	< 0.001	1.1499 (0.0793)	< 0.001	-0.0413 (0.1505)	0.78

Table 3.10: Estimates of causal effect (SE) of log(CRP) on fibrinogen from Copenhagen General Population Study divided randomly into substudies and combined: using 2SLS summary study estimates by fixed-effect meta-analysis, using individual patient data (IPD) with different and common genetic effects across substudies

3.7 Discussion

This chapter demonstrates the effect of weak instrument bias on causal estimates in real and simulated data. We have shown by simulation and using a variety of explanations that the magnitude of this bias depends on the statistical strength of the association between instrument and phenotype. Using 2SLS, when multiple instruments were used, we found in our simulations that the median size of the bias of the IV estimator was approximately $1/F$ of the bias in the observational association, where F is the mean F statistic from the regression of phenotype on instrument. So a mean F statistic of 10 limits the median relative bias to less than 10%. When a single instrument was used, a mean F statistic of 5 seemed to be sufficient to ensure median relative bias was about 5%, and a mean F statistic greater than 10 ensured negligible bias from weak instruments. A limitation of this conclusion is that, unlike for the relative mean bias (181), there is no theoretical basis for this approximation and we have undertaken only a simulation exercise. Using LIML, the median bias was close to zero throughout, even in a real data example using a large number of correlated instruments.

While the magnitude of the bias depends on the instrument strength through the mean or expected F statistic, for a study of fixed size and underlying instrument strength, an observed F statistic greater than the expected F value corresponds to an estimate closer to the observational association with greater precision; conversely an observed F statistic less than the expected F value corresponds with an estimate further from the observational association with less precision. So simply relying on an F statistic from an individual

study is over-simplistic and simple threshold rules such as ensuring $F > 10$ may cause more bias than they prevent.

Using the 2SLS method, we demonstrated a bias–variance trade-off for number of instruments used in IV estimation. For a fixed mean F statistic, as the number of instruments increases, the precision of the IV estimator increases, but the bias also increases. Using the LIML method, bias did not increase appreciably with the number of instruments. Nevertheless, we seek parsimonious models of genetic association, for example using additive per allele effects and including only IVs with a known association with the phenotype, based on biological knowledge and external information. Provided the data are not severely misrepresented, these should provide the best estimates of causal association. It is also possible to summarize multiple SNPs using a gene score (44). If this is done using pre-specified weights, this makes strong assumptions about the effects of different SNPs which may itself introduce bias. The use of a data-derived weighted gene score is equivalent to 2SLS (182). Again, post-hoc use of observed F statistics to choose between instruments may cause more bias than it prevents.

Ideally, issues of weak instrument bias should be addressed prior to data collection, by specifying sample sizes, instruments, and genetic models using the best prior evidence available to ensure that the expected value of F statistics are large. Where this is not possible, our advice would be to conduct sensitivity analyses using different IV methods, numbers of instruments and genetic models to investigate the impact of different assumptions on the causal estimate.

Generally, the LIML estimate is less biased than the 2SLS estimate. Difference between the 2SLS and LIML IV estimates is evidence of possible bias from weak instruments. When a single instrument is used, the 2SLS and LIML estimates coincide, and the IV estimate is close to median unbiased. The LIML estimate with any number of instruments and the 2SLS estimate with one instrument do not have finite moments, and so do not have a defined mean bias; however this would not generally be a problem in applied research. The Fuller(1) estimator does have a finite mean for any number of instruments, but shows considerable median and mean bias with one instrument.

Another technique which helps reduce weak instrument bias is adjustment for covariates. Including predictors of the phenotype in the first stage regression, or predictors of the outcome in the second stage regression, increases precision of the causal estimate. The former will also increase the F statistic for the genetic IVs, and thus reduce weak instrument bias.

In a meta-analysis context, bias is a more serious issue, as it arises not only from the bias in the individual studies, but also from the correlation between causal effect size and

variance which results in studies with effects closer to the observational estimate being over-weighted. By using a single IPD model, we reduce the second source of bias. Additionally, we can pool information on the genetic association across studies to strengthen the instruments. The assumptions of homogeneity of variances and common genetic effects across studies will often be overly restrictive. Allowing for heterogeneity across studies in phenotype variance, genetic effects, and in the causal effects themselves, is possible in a Bayesian framework (140), and is discussed in Chapter 5.

Finally, we emphasize that the use of a genetic instrument in Mendelian randomization relies on certain assumptions. In this chapter we have assumed, although these may fail in finite samples, that they hold asymptotically. If these assumptions do not hold, for example if there were a true correlation between the instrument and a confounder, then IV estimates can be entirely misleading (183) and "the cure can be worse than the disease" (184).

3.7.1 Key points from chapter

- Bias from weak instruments can result in seriously misleading estimates of causal effects. Studies with instruments having high mean F statistics are less biased on average. However, if a study by chance has a higher F statistic than expected, then the causal estimate will be more biased.

- Data-driven choice of instruments or analysis can exacerbate bias. In particular, any guideline such as $F > 10$ is misleading. Methods, instruments, and data to be used should be specified prior to data analysis. Meta-analysis based on summary study-specific estimates of causal association are susceptible to bias.

- Bias can be alleviated in a single study by using the LIML rather than 2SLS method and by adjusting for measured confounders, and in a meta-analysis by using IPD modelling. We advocate parsimonious modelling of the genetic association (e.g. per allele additive SNP model rather than one coefficient per genotype). This should be accompanied by sensitivity analyses to assess potential bias.

Chapter 4

Collapsibility for IV analyses of binary outcomes

4.1 Introduction

When an estimate of association between a phenotype and outcome from an observational study is compared to that from a randomized controlled trial (RCT), there is often disagreement between the estimates (3). As previously stated, this may be due to confounding or reverse causation in observational studies, or non-compliance in trials (185). However, even when there is no confounding, reverse causation or non-compliance and the model is correctly specified, there may be a difference between the estimates, as the observational study estimate will typically be conditional on covariates, while the RCT estimate is typically marginal across these covariates (109). This is known as non-collapsibility, and it affects estimates of odds ratios (33).

A second, related issue is that of whether an effect estimate represents a subject-specific or a population-based effect (186). If individuals in a population have heterogeneous levels of risk, a non-collapsible measure of association differs depending on whether it is considered for an individual within the population or for the population as a whole. Covariates for the outcome represent one source of such heterogeneity for risk.

As we have seen in previous chapters, instrumental variables (IV) can be used to estimate causal effects which are free of bias from confounding and reverse causation. However, when the measure of association is not collapsible across variation in risk, it is not clear which quantity is being estimated. For this reason, regression analyses of non-linear problems using IV techniques have been labelled "forbidden regressions" by econometricians (118; 129; 187). We explore the reasons for this prohibition in this chapter.

The use of instrumental variables in epidemiological research has been advocated in randomized trials to adjust for non-compliance, and in observational studies to adjust for unmeasured covariates. Difference between the estimates of an association in a randomized trial with and without adjustment for compliance is taken as evidence of bias due to treatment contamination. Similarly, difference between an association using observational data estimated by conventional regression methods with adjustment for known covariates and by IV analysis is taken as evidence of unmeasured confounding or reverse causation. For this reason, it is important to know whether the estimates compared are targeting the same quantity or not. Although the general context of this chapter will be that of Mendelian randomization, there is no restriction of the mathematical findings to the use of genetic IVs.

In this chapter, we define non-collapsibility, and illustrate it for the odds ratio parameter (Section 4.2). We define odds ratios which are marginal and conditional on the phenotype, which reflect the effect of a population intervention in the phenotype (marginal) or an individual intervention (conditional). Odds ratio also differ depending on the choice of covariates conditioned on. The difference between various odds ratios is demonstrated using simulated and real data (Section 4.3). We show how the ratio or two-stage IV estimate in a logistic model is consistent for the odds ratio corresponding to an increase in the risk factor across its population distribution, conditional within strata of the instrument and marginal across all other covariates. This is similar to the odds ratio from a randomized controlled trial without adjustment for any covariates, where the intervention in the risk factor corresponds to a unit change across the population. Under certain specific conditions, when adjustment in the IV regression is made for an estimate of the unmeasured covariates, an individual odds ratio can be estimated which is conditional on covariates (Section 4.4). Finally, we discuss how the issue of non-collapsibility affects the interpretation of analyses of observational data, RCTs and instrumental variable situations (Section 4.5).

4.2 Collapsibility

We introduce the concept of collapsibility by telling two short stories about odds ratios which represent the answer to different causal questions about interventions in a risk factor.

4.2.1 Collapsibility across a covariate

A person approaches a statistician in a dark alleyway and says in a low and indeterminate voice: "What's the odds ratio for heart disease of smoking?". The statistician replies, "1.89". The stranger comes closer: "Thank you, kind sir, for helping a lady with her problem". The statistician replies, "Oh, you are female. In that case, your odds ratio is actually 2." The lady exclaims, "So the odds ratio for men is less than 1.89?". The statistician replies, "No, for men it is also 2.".

Paradoxically, this story can be true. The numbers chosen to tell the story are given in the left half of Table 4.1. We see that the odds ratio changes depending on whether the ratio is conditional on sex or not. While the statistician is being obtuse, as in this toy example the stratum-specific or individual odds ratio is the same for men and women and each individual is a member of exactly one of those categories, this story illustrates the non-collapsibility of the odds ratio. For simplicity, we assume that the populations of smokers and non-smokers contains men and women in equal proportions, meaning that sex is not a confounding factor in the association of smoking with heart disease. In contrast, as the right half of Table 4.1 shows, a relative risk is the same whether conditional or marginal on sex.

	Probability of event			Probability of event		
	Non-smoker	Smoker	Odds ratio	Non-smoker	Smoker	Relative risk
Men	$\frac{3}{13}$	$\frac{3}{8}$	2	0.3	0.6	2
Women	$\frac{1}{21}$	$\frac{1}{11}$	2	0.05	0.1	2
Overall	0.168	0.318	1.89	0.175	0.35	2

Table 4.1: Illustrative example of collapsing an effect estimate over a covariate: non-equality of conditional and marginal odds ratios and equality of relative risks

A measure of association is collapsible over a covariate, as defined by Greenland et al. (110), if, when it is constant across the strata of the covariate, this constant value equals the value obtained from the overall (marginal) analysis. Non-collapsibility is the violation of this property. The relative risk and absolute risk difference are collapsible across strata measures of association (188; 189). Odds ratios are generally not collapsible unless both risk factor and outcome are independent of the covariate, or risk factor and covariate are conditionally independent given the outcome, or outcome and covariate are conditionally independent given the risk factor (190). Hazard ratios from survival analyses are also not generally collapsible (191).

4.2.2 Collapsibility across the risk factor distribution

The lady continues: "My cardiovascular risk score is 1.8. What is the odds ratio for heart disease of increasing the score by one?". The statistician replies: "2". "And for my husband, who has a risk score of 1.4?". "2". "And for my children, who have risk scores of 0.4 and 0.2?". "The odds ratio for an individual is 2". "So the odds ratio for our family if everyone's risk score increased by one is ...". "1.94".

If the true probability of event (π) is related to the risk score (X) by the risk model $\operatorname{logit}\pi = -2 + X\log(2)$, then the odds ratio for any individual for a unit increase in X is 2. However, for a group of heterogeneous individuals, the odds ratio is different to 2. As above, if the true risk model is $\log\pi = -2 + X\log(2)$, then the relative risk for any individual is 2 and the population relative risk is also 2.

Logistic-linear model: $\operatorname{logit}\pi = -2 + X\log(2)$			
Risk score (x)	Probability given $X = x$	Probability given $X = x + 1$	Odds ratio
0.2	0.135	0.237	2
0.4	0.152	0.263	2
1.4	0.263	0.417	2
1.8	0.320	0.485	2
Average	0.217	0.351	1.94
Log-linear model: $\log\pi = -2 + X\log(2)$			
Risk score (x)	Probability given $X = x$	Probability given $X = x + 1$	Relative risk
0.2	0.155	0.311	2
0.4	0.179	0.357	2
1.4	0.357	0.714	2
1.8	0.471	0.943	2
Average	0.291	0.581	2

Table 4.2: Illustrative example of collapsing an effect estimate across the risk factor distribution: non-equality of individual and population odds ratios and equality of relative risks

These two examples both demonstrate the attenuation of the odds ratio when the probability of an event is averaged across a distribution. In the first example, the variation can be explained by a covariate, and the different odds ratios represent the measure of association for a change from non-smoker to smoker conditional or marginal on the covariate, sex. In the second example, the risk model is constructed so that there is no omitted covariate, simply individuals with different levels of the risk factor, and the odds

ratio represents the measure of association for a unit increase in the risk score. In both cases, the odds ratio for an individual in the population is different to the odds ratio for the population as a whole.

4.3 Exploring differences in odds ratios

Before considering how issues of collapsibility affect IV estimation, we firstly consider different definitions of odds ratios, and then see how these odds ratios have different numerical values in simulated and real data.

4.3.1 Individual and population odds ratios

We consider the association between a phenotype (X) and an outcome (Y). We assume the covariates for the outcome can be summarized by a single random variable V (94). If V were known and conditioned on, the estimate of association of X on Y would be equal to the causal association. We note that as V contains all information on the covariates for Y, any sufficient covariate U is a function of V. As the distribution of Y is unlikely to be dominated by just a few factors, ability to reduce the covariates to a single univariate random variable seems a reasonable assumption. For example, if all the covariates $V_1, \dots V_p$ are linearly associated and normally distributed, then we could replace these V_j with a single normally distributed V.

An individual effect is the change in outcome due to an intervention in the phenotype conditional on the phenotype, and a population effect is the change in outcome due to an intervention in the phenotype averaged across the distribution of the phenotype. For a binary outcome $Y = 0$ or 1, the conditional individual odds ratio (CIOR) is defined as the odds ratio for unit increase in the phenotype from x to $x + 1$ for a given value of v:

$$\text{CIOR}(x, v) = \frac{\text{odds}(Y(x + 1, v))}{\text{odds}(Y(x, v))} \tag{4.1}$$

where $\text{odds}(Y) = \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}$ and $Y(x, v) = Y|(X = x, V = v)$ is the outcome random variable with phenotype level x and covariate level v.

The conditional population odds ratio (CPOR) is defined as the odds ratio for unit increase in the distribution of the phenotype from X to $X + 1$. This is an increase from x to $x + 1$ marginalized over the phenotype distribution for a given value of v:

$$\text{CPOR}(v) = \frac{\text{odds}(Y(X + 1, v))}{\text{odds}(Y(X, v))} \tag{4.2}$$

where the probabilities in the odds function are averaged across (or integrated over) the distribution of X.

In general, the CIOR may be a function of x and v, although in a logistic-linear model of association, where the logit of the probability of outcome (π) is a linear function in X and V with no interaction term:

$$\begin{aligned} Y &\sim \text{Binomial}(1, \pi) \\ \text{logit}(\pi) &= \beta_0 + \beta_1 X + \beta_2 V \end{aligned} \tag{4.3}$$

the CIOR is independent of x and v:

$$\begin{aligned} \text{CIOR}(x, v) &= \frac{\mathbb{P}(Y(x+1, v) = 1)}{\mathbb{P}(Y(x+1, v) = 0)} \Big/ \frac{\mathbb{P}(Y(x, v) = 1)}{\mathbb{P}(Y(x, v) = 0)} \\ &= \frac{\left(\frac{\exp(\beta_0+\beta_1(x+1)+\beta_2 v)}{1+\exp(\beta_0+\beta_1(x+1)+\beta_2 v)}\right)\left(\frac{1}{1+\exp(\beta_0+\beta_1 x+\beta_2 v)}\right)}{\left(\frac{\exp(\beta_0+\beta_1 x+\beta_2 v)}{1+\exp(\beta_0+\beta_1 x+\beta_2 v)}\right)\left(\frac{1}{1+\exp(\beta_0+\beta_1(x+1)+\beta_2 v)}\right)} \\ &= \exp(\beta_1) \end{aligned} \tag{4.4}$$

This is the odds ratio estimated by a logistic regression of Y on X and V.

Unless X is constant, the CPOR is a non-trivial function of the variable v even in the case of model (4.3), and so we remove the dependence on V by integrating over the joint distribution of X and V to obtain a marginal population odds ratio (MPOR):

$$\text{MPOR} = \frac{\text{odds}(Y(X+1, V))}{\text{odds}(Y(X, V))} \tag{4.5}$$

$$\begin{aligned} &= \frac{\mathbb{P}_{X,V}(Y(X+1, V) = 1)}{\mathbb{P}(Y(X+1, V) = 0)} \Big/ \frac{\mathbb{P}(Y(X, V) = 1)}{\mathbb{P}(Y(X, V) = 0)} \\ &= \frac{\mathbb{E}_{X,V}[Y(X+1, V)]}{1 - \mathbb{E}_{X,V}[Y(X+1, V)]} \Big/ \frac{\mathbb{E}_{X,V}[Y(X, V)]}{1 - \mathbb{E}_{X,V}[Y(X, V)]} \end{aligned} \tag{4.6}$$

This represents the ratio of the odds for a population with the whole distribution of the phenotype shifted up by one to the odds for a population with the original distribution of the phenotype. From here on, we assume that the model of association is logistic-linear, and drop the dependence on the value of x and v, referring to the CIOR(x, v) as simply the individual odds ratio (IOR) and the MPOR as the population odds ratio (POR). The POR depends on the (usually unknown) distributions of the phenotype and covariate, and is generally attenuated compared to the $\exp(\beta_1)$ due to the convexity of the logit function (Jensen's inequality). In Model (4.3), we can write the population log odds ratio

(PLOR = log POR) explicitly as:

$$\begin{aligned} \text{PLOR} = \quad & \text{logit} \int\int \text{expit}(\beta_0 + \beta_1(x+1) + \beta_2 v) f(x,v) dx dv \\ - \quad & \text{logit} \int\int \text{expit}(\beta_0 + \beta_1 x + \beta_2 v) f(x,v) dx dv \end{aligned} \tag{4.7}$$

where $f(x, v)$ is the joint distribution of X and V and $\text{expit}(x) = (1 - \exp(-x))^{-1}$ is the inverse of $\text{logit}(x)$.

We can think of the PLOR as the estimate of association from a simulated RCT where the intervention is a unit increase in the phenotype. In the context of a randomized trial, the ratio between the odds of two randomized groups is known as an incident odds ratio (109). In a simulated example, we can calculate the incident odds ratio in our simulated population. For each individual $i = 1, \ldots, N$, we consider a counterfactual individual, identical to the first, except with phenotype x_i increased by one. We separately draw two independent sets of outcomes y_{1i}, y_{2i} for the original and counterfactual populations.

$$\begin{aligned} \text{logit}(\pi_{1i}) &= \beta_0 + \beta_1 x_i + \beta_2 v_i \\ y_{1i} &\sim \text{Binomial}(1, \pi_{1i}) \\ \text{logit}(\pi_{2i}) &= \beta_0 + \beta_1 (x_i + 1) + \beta_2 v_i \\ y_{2i} &\sim \text{Binomial}(1, \pi_{2i}) \end{aligned} \tag{4.8}$$

The incident log odds ratio (InLOR) is calculated as the log odds ratio for a unit intervention on phenotype, which is the difference in log odds between the real and counterfactual populations.

$$\text{InLOR} = \log\left(\frac{\widehat{\text{odds}}(Y_2)}{\widehat{\text{odds}}(Y_1)}\right) \tag{4.9}$$

$$= \log\left(\frac{\sum y_{2i}}{N - \sum y_{2i}}\right) - \log\left(\frac{\sum y_{1i}}{N - \sum y_{1i}}\right) \tag{4.10}$$

This is a Monte Carlo approximation to the integrals in (4.7), meaning that InLOR $\to$ PLOR as $N \to \infty$. In our calculations, we sum the probabilities $\hat{\pi}_{1i}, \hat{\pi}_{2i}$ rather than summing over the events y_{1i}, y_{2i} to reduce sampling variation in equation (4.10).

Both the individual and population effects are *ceteris paribus* (Latin: "with all other things equal") estimates; they estimate the effect on the outcome of an intervention on the risk factor with all other factors (such as covariates) kept equal (192). For this reason, both can be thought of as causal effects. The population estimate is averaged across levels of the phenotype and other covariates, whereas the individual estimate is conditional on the value of phenotype and other covariates (186).

4.3.2 Marginal and conditional estimates

If there are multiple covariates, then a causal effect can be conditional on some covariates and marginal across others, depending on which covariates are conditioned on. Although odds ratios typically differ depending on covariate adjustment, a null causal association of X on Y leads to an odds ratio of one no matter which covariates the odds ratio is considered to be marginal and conditional across. For this reason, distinction between unconfounded odds ratios is not an issue for hypothesis testing, but for parameter estimation (see Section 2.3); conditional and marginal odds ratios test the same null hypothesis.

4.3.3 Population and individual odds ratios in simulated data

We consider a confounded model of association between a phenotype and outcome, simulating data for N participants indexed by i. We aim to show how the individual and population odds ratios differ in a simple setting. The phenotype (X) is a linear combination of a covariate G which takes two values, a normally distributed covariate V and an error term. The outcome (Y) is a binary variable, taking value 1 with probability π_1, which is a logistic function of the phenotype and covariate V. Although G will be thought of later as an IV, it could here be any covariate dividing the population independently of V into strata with different mean phenotype levels.

$$\begin{aligned} x_i &= \alpha_0 + \alpha_1 g_i + \alpha_2 v_i + \epsilon_i \\ \text{logit}(\pi_{1i}) &= \beta_0 + \beta_1 x_i + \beta_2 v_i \\ y_i &\sim \text{Binomial}(1, \pi_{1i}) \\ v_i \sim \mathcal{N}(0,1), \epsilon_i &\sim \mathcal{N}(0, \sigma_x^2) \text{ independently} \end{aligned} \tag{4.11}$$

The individual log odds ratio (ILOR) conditional on V is β_1 as in equation (4.4).

To illustrate the difference between the population and individual log odds ratios, we set $\beta_0 = -2$, $\alpha_0 = 0$ throughout and consider two different sizes of ILOR, $\beta_1 = 0.4, -0.8$ (corresponding to IORs 1.49 and 0.45), and seven different values for the covariate effect ($\beta_2 = -1.0, -0.6, -0.2, 0, 0.2, 0.6, 1.0$). We assume that G divides the population into two strata of equal size ($g_i = 0, 1$). We consider the PLOR in five scenarios:

1. X is constant ($\alpha_1 = 0, \alpha_2 = 0, \sigma_x^2 = 0$)
2. X varies independently of the covariate V ($\alpha_1 = 0, \alpha_2 = 0, \sigma_x^2 = 2$)
3. X is correlated with the covariate V ($\alpha_1 = 0, \alpha_2 = 1, \sigma_x^2 = 1$)

4. X has constant levels depending on G ($\alpha_1 = 1, \alpha_2 = 0, \sigma_x^2 = 0$)

5. X varies with V and G ($\alpha_1 = 1, \alpha_2 = 1, \sigma_x^2 = 1$)

Results were calculated using the Monte Carlo method (equation (4.10)) for a large sample ($N > 1000000$) and checked by numerical integration using the *adapt* package in R (193). The numerical integration algorithm was quite sensitive to the parameters used, as integrating over too large a range induced numerical overflow and integrating over too small a range lost accuracy by clipping the tails of the distribution. In contrast, the Monte Carlo estimates were very stable across iterations.

		$\beta_2 = -1.0$	$\beta_2 = -0.6$	$\beta_2 = -0.2$	$\beta_2 = 0$	$\beta_2 = 0.2$	$\beta_2 = 0.6$	$\beta_2 = 1.0$
Scenario 1	$\beta_1 = 0.4$	0.3491	0.3814	0.3980	0.4000	0.3980	0.3814	0.3491
	$\beta_1 = -0.8$	-0.7202	-0.7742	-0.7975	-0.8000	-0.7975	-0.7742	-0.7202
Scenario 2	$\beta_1 = 0.4$	0.3347	0.3648	0.3814	0.3835	0.3814	0.3648	0.3347
	$\beta_1 = -0.8$	-0.6220	-0.6678	-0.6933	-0.6967	-0.6933	-0.6678	-0.6220
Scenario 3	$\beta_1 = 0.4$	0.3364	0.3683	0.3863	0.3886	0.3863	0.3683	0.3364
	$\beta_1 = -0.8$	-0.6739	-0.7227	-0.7475	-0.7506	-0.7475	-0.7227	-0.6739
Scenario 4	$\beta_1 = 0.4$	0.3437	0.3772	0.3955	0.3978	0.3955	0.3772	0.3437
	$\beta_1 = -0.8$	-0.7227	-0.7709	-0.7910	-0.7931	-0.7910	-0.7709	-0.7227
Scenario 5	$\beta_1 = 0.4$	0.3683	0.3863	0.3863	0.3794	0.3683	0.3364	0.2994
	$\beta_1 = -0.8$	-0.5429	-0.6097	-0.6738	-0.7010	-0.7227	-0.7475	-0.7475

Table 4.3: Population log odds ratio (PLOR) for unit increase in phenotype from five example models

Table 4.3 shows that even in this simple model, the PLOR is only equal to the ILOR when X is constant and there is no other covariate which is a competing risk factor for Y. A competing risk factor (even if it is not a confounder), variation in X, and stratification of X all result in an attenuation of the PLOR. The maximal attenuation in the examples considered here is 27% (-0.5429 from -0.8). If we had instead considered a log-linear model of Y on X and examined the population relative risk, Table 4.3 would have contained only the two values 0.4 and -0.8, as the population relative risk is equal to the individual relative risk throughout.

This example illustrates the non-collapsibility of the odds ratio. The odds ratio for a risk factor does not average correctly, attenuating when averaged across a population with any variation or heterogeneity in the risk factor, or when there is an alternative risk factor. The relative risk does average correctly. This means that an odds ratio for a risk factor estimated from observational data by logistic regression conditional on covariates will be an overestimation of the expected effect of the same intervention on the population.

4.3.4 Population and individual odds ratios in five studies

To show a similar difference between the population and the individual odds ratios in real data, we consider data from five studies which investigate heart disease, of which three are retrospective case-control studies: Precocious Coronary Artery Disease Study (PROCARDIS), Ludwigshafen Risk and Cardiovascular Health Study (LURIC), Stockholm Heart Epidemiology Program (SHEEP); and two are cohort studies: Cardiovascular Health Study (CHS) and Rotterdam Study (ROTT). We take cross-sectional data from 21 090 individuals including 6218 with a previous history of myocardial infarction (MI) (defined using World Health Organization criteria) to investigate the effect of C-reactive protein (CRP) on MI. Logistic models of disease outcome on log-transformed CRP were constructed with various levels of adjustment for confounding. In this section, the goal is not the estimation of causal association, but rather to investigate the magnitude of the attenuation of the population from the individual odds ratio.

We compare the ILOR of a unit increase in log(CRP), estimated by logistic regression, with the PLOR of a unit increase in log(CRP). The PLOR is estimated by increasing the predictor in the logistic model, which represents the probability of an event, by $\hat{\beta}_1$, the coefficient for a unit increase in log(CRP) from the logistic regression model, and summing over the new probabilities to obtain the mean number of cases for a counterfactual population with log(CRP) increased by one.

For individual i, if we have the linear predictor (η_i) for our regression model of probability of MI event (π_i) on log(CRP) (x_i) and confounders (v_{ij}):

$$\eta_i = \text{logit}(\pi_i) = \beta_0 + \beta_1 x_i + \sum_j \beta_{2j} v_{ij} \tag{4.12}$$

Then our population log odds ratio is estimated as:

$$\begin{aligned} \widehat{\text{PLOR}} = \quad & \text{logit}\left(\tfrac{1}{N}\sum_i \text{expit}(\hat{\beta}_0 + \hat{\beta}_1(x_i+1) + \sum_j \hat{\beta}_{2j} v_{ij})\right) \\ - \quad & \text{logit}\left(\tfrac{1}{N}\sum_i \text{expit}(\hat{\beta}_0 + \hat{\beta}_1(x_i) + \sum_j \hat{\beta}_{2j} v_{ij})\right) \end{aligned} \tag{4.13}$$

This is similar to the Monte Carlo approach of equation (4.10), except that summation of the event probabilities is across the empirical distribution of the phenotype and confounders from the data.

This calculation assumes that the regression model in use is correct, and specifically that all covariates which represent competing risk factors have been accounted for. Although this is an unrealistic assumption, it is made here for purpose of illustration. In

case-control studies, as the probabilities of an event cannot be estimated directly, we have adjusted the model intercept to give a 7% incidence rate in the population from which the case-control sample was ascertained (194).

Table 4.4 shows how the individual odds ratios represent an over-estimation of the true effect of a population unit intervention in CRP levels on MI. While the estimates of association in Table 4.4 should not be regarded as causal effects, due to the unrealistic assumptions of no unmeasured confounders or competing risk factors, the estimates illustrate that, in real data, the individual and population odds ratios can be somewhat different. The linear predictor, the logit of the probability of an event, has an approximate normal distribution. In PROCARDIS, with no adjustment, the standard deviation of the linear predictor for the cohort is 0.41, increasing to 0.92 on adjustment for sex, diabetes status and age, and to 1.38 on further adjustment for total cholesterol, high-density lipid cholesterol and log(triglycerides). This indicates that individuals in the population have heterogeneous levels of risk of developing MI. In CHS, the standard deviation of the linear predictor for the fully adjusted model considered here is 0.89, and there is less attenuation of the individual odds ratio compared with PROCARDIS. Even assuming the effect of CRP is no longer confounded, further adjustment for unmeasured covariates would lead to greater attenuation of the POR. This is because the logistic function is less well approximated by a linear function as the domain and range of the function considered widens. In the maximally-adjusted models considered here, there is a 5–14% attenuation of the PLOR compared to the ILOR.

4.3.5 Summary

An odds ratio changes when marginalized across heterogeneity in risk, whether the heterogeneity is explainable by covariates or represents different levels of the phenotype. These two issues of marginalization across a covariate and phenotype distribution are related, but separate. Marginalizing over covariates is necessary when considering a population odds ratio, as otherwise the population odds ratio is a function of the covariate and so takes different values across strata of the covariate. With an individual odds ratio, marginalizing over or conditioning on a covariate is a choice to be made in terms of interpretation of the coefficients in the model. An odds ratio from a RCT usually targets a odds ratio marginal across covariates, as adjustment for covariates is not necessary. Observational epidemiological analysis using logistic regression targets a conditional individual odds ratio, as adjustment for covariates is necessary to avoid confounding. Once a choice of covariates has been made for the model, a population or an individual odds ratio can

Model [1]	Individual (log) odds ratio	Population (log) odds ratio
PROCARDIS ($N = 6464, n = 3135$)		
No adjustment	1.4408 (0.3652)	1.4330 (0.3598)
Adjustment for sex, diabetes status and age	1.4371 (0.3626)	1.3911 (0.3301)
Further adjustment for tchol, hdl, log(tg)	1.3048 (0.2661)	1.2570 (0.2287)
LURIC ($N = 3236, n = 1335$)		
No adjustment	1.2801 (0.2470)	1.2775 (0.2449)
Adjustment for sex, diabetes status and age	1.2690 (0.2382)	1.2633 (0.2337)
Further adjustment for sbp, tchol, hdl, bmi, log(tg)	1.1927 (0.1762)	1.1852 (0.1699)
SHEEP ($N = 1994, n = 858$)		
No adjustment	1.4312 (0.3585)	1.4241 (0.3535)
Adjustment for sex, diabetes status and age	1.4057 (0.3405)	1.3881 (0.3280)
Further adjustment for tchol, hdl, bmi, log(tg)	1.2872 (0.2525)	1.2637 (0.2341)
CHS ($N = 4506, n = 449$)		
No adjustment	1.2554 (0.2275)	1.2538 (0.2262)
Adjustment for sex, diabetes status and age	1.2284 (0.2057)	1.2186 (0.1977)
Further adjustment for sbp, tchol, hdl	1.1854 (0.1701)	1.1758 (0.1619)
ROTT ($N = 5402, n = 647$)		
No adjustment	1.3525 (0.3020)	1.3476 (0.2983)
Adjustment for sex, diabetes status and age	1.2327 (0.2092)	1.2200 (0.1988)
Further adjustment for tchol, hdl	1.1849 (0.1697)	1.1732 (0.1597)

Table 4.4: Individual and population odds ratios (log odds ratios) for a unit increase in log(CRP) on myocardial infarction (MI) odds from logistic regression in five studies (N = number of participants, n = number of events)

[1] tchol = total cholesterol, hdl = high-density lipid cholesterol, bmi = body mass index, sbp = systolic blood pressure, tg=triglycerides

be estimated. The difference in interpretation between the two odds ratios is between a population-averaged and an individual-specific effect. Neither of the estimates is 'correct' or 'incorrect'; they simply represent the answer to different questions.

4.4 Instrumental variables

In this section, we consider how the difference between individual and population odds ratios is relevant to IV estimation. We show this firstly analytically, considering a simple model of association between an instrument, phenotype and outcome. We then show this by simulation in a more realistic setting.

4.4.1 Relation of the two-stage IV estimator and population odds ratio

We aim to show through analytic results and careful simulation how the quantity estimated by the two-stage method is a population odds ratio.

With a single instrument, the two-stage estimator equals the ratio of the coefficient from the logistic regression of outcome on the IV to the coefficient from the linear regression of phenotype on the IV.

$$\hat{\beta}_1^R = \hat{\beta}_{GY}/\hat{\beta}_{GX} \tag{4.14}$$

We assume here that G takes values 0 and 1, and that the outcome Y has a Bernouilli distribution with probability of event π and linear predictor $\eta = \text{logit}(\pi)$.

$$\begin{aligned} X &= \alpha_0 + \alpha_1 G + g(U) + \epsilon_X \\ \eta = \text{logit}(\pi) &= X + h(V) \\ Y &\sim \text{Bernouilli}(\pi) \end{aligned} \tag{4.15}$$

where $g(.)$ is an arbitrary function of the covariates U for X, $h(.)$ is an arbitrary function of the covariates V for Y, and ϵ_X is an independent error term for X. We consider the logistic regression of Y on G using the model:

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 g_i \tag{4.16}$$

We have the likelihood L and log-likelihood ℓ such that

$$L = \prod_i \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{4.17}$$

$$\begin{aligned}\ell &= \sum_i y_i \log \pi_i + (1-y_i)\log(1-\pi_i) \\ &= \sum_i y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_i \log(1-\pi_i) \\ &= \sum_i y_i(\gamma_0 + \gamma_1 g_i) - \sum_i \log(1+\exp(\gamma_0+\gamma_1 g_i))\end{aligned} \tag{4.18}$$

Differentiating, we obtain

$$\frac{\partial \ell}{\partial \gamma_0} = \sum_i y_i - \sum_i \operatorname{expit}(\gamma_0 + \gamma_1 g_i) \tag{4.19}$$

$$\frac{\partial \ell}{\partial \gamma_1} = \sum_i g_i y_i - \sum_i g_i \operatorname{expit}(\gamma_0 + \gamma_1 g_i) \tag{4.20}$$

Whence,

$$\hat{\gamma}_0 = \operatorname{logit}\left(\frac{\sum_i y_i(1-g_i)}{\sum_i (1-g_i)}\right) \tag{4.21}$$

$$\hat{\gamma}_1 = \operatorname{logit}\left(\frac{\sum_i y_i g_i}{\sum_i g_i}\right) - \operatorname{logit}\left(\frac{\sum_i y_i(1-g_i)}{\sum_i (1-g_i)}\right) \tag{4.22}$$

As the sample size N tends to infinity, by the law of large numbers, $\sum_i y_i g_i \to \mathbb{E}[YG] = \mathbb{P}(Y=1, G=1)$. Thus

$$\begin{aligned}\hat{\gamma}_1 &\to \operatorname{logit}\left(\frac{\mathbb{P}(Y=1,G=1)}{\mathbb{P}(G=1)}\right) - \operatorname{logit}\left(\frac{\mathbb{P}(Y=1,G=0)}{\mathbb{P}(G=0)}\right) \\ &= \operatorname{logit}(\mathbb{P}(Y=1|G=1)) - \operatorname{logit}(\mathbb{P}(Y=1|G=0)) \\ &= \operatorname{logit}(\mathbb{E}[Y|G=1]) - \operatorname{logit}(\mathbb{E}[Y|G=0]) \\ &= \operatorname{logit}(\mathbb{E}[Y(X(1))] - \operatorname{logit}(\mathbb{E}[Y(X(0))]\end{aligned} \tag{4.23}$$

where here $Y(x) = Y|(X=x)$ and $X(g) = X|(G=g)$ (note that $Y \perp\!\!\!\perp G|X$ in this example) and the probabilities and expectations are averaged across the distribution of X and V. Hence we see that the coefficient $\hat{\gamma}_1 = \hat{\beta}_{GY}$ is the log odds ratio corresponding to an increase of α_1 across the distribution of X conditional on G. As we see, this log odds ratio is a population odds ratio conditional on G but marginal in all other covariates. As the sample size increases, the denominator of the IV estimate converges in probability to the constant α_1, so the IV estimator converges to the ratio $\frac{1}{\alpha_1}\operatorname{plim}_{N\to\infty}\hat{\beta}_{GY}$ by Slutsky's theorem. We write this quantity as $\operatorname{plim}\hat{\beta}_1^R$ as we shall refer to it as the IV estimand.

4.4.2 IV estimation in simplistic simulated scenarios

We take a series of scenarios, starting with a simple model for the joint distribution of G, U, V, X and Y and adding complexity step-by-step. U represents a covariate for X and V a independent covariate for Y. For simplicity of calculation, both U and V take values 0 or 1 with equal probability. Neither covariate is regarded as known and so both are omitted from the models. In each case, the coefficient of X (the ILOR) is 1. We calculate the PLOR (which is marginal in all covariates) and IV estimand $\operatorname{plim} \hat{\beta}_1^R = \frac{1}{\alpha_1} \operatorname{plim} \hat{\beta}_1^{GY}$ for five scenarios.

1. No variation in phenotype or linear predictor.

$$\begin{aligned} X &= G \\ \eta = \operatorname{logit}(\pi) &= X \end{aligned} \tag{4.24}$$

2. No variation in phenotype or linear predictor, smaller IV effect.

$$\begin{aligned} X &= 0.3G \\ \eta = \operatorname{logit}(\pi) &= X \end{aligned} \tag{4.25}$$

3. No variation in phenotype, variation in linear predictor.

$$\begin{aligned} X &= G \\ \eta = \operatorname{logit}(\pi) &= X + V \\ V &\sim \text{Bernouilli}(0.5) \end{aligned} \tag{4.26}$$

4. No variation in phenotype, variation in linear predictor, smaller IV effect.

$$\begin{aligned} X &= 0.3G \\ \eta = \operatorname{logit}(\pi) &= X + V \\ V &\sim \text{Bernouilli}(0.5) \end{aligned} \tag{4.27}$$

5. Variation in phenotype, variation in linear predictor.

$$\begin{aligned} X &= G + U \\ \eta = \operatorname{logit}(\pi) &= X + V \\ U, V &\sim \text{Bernouilli}(0.5) \text{ independently} \end{aligned} \tag{4.28}$$

Results are given in Table 4.5. In each of the first four examples, there is no random variation in X. In examples 1 and 2, there is no variation in the linear predictor except

	PLOR	
Example 1	$\text{logit}(\frac{1}{2}\{\text{expit}(2)+\text{expit}(1)\})$ $-\text{logit}(\frac{1}{2}\{\text{expit}(1)+\text{expit}(0)\})$	$=0.953$
Example 2	$\text{logit}(\frac{1}{2}\{\text{expit}(1.3)+\text{expit}(1)\})$ $-\text{logit}(\frac{1}{2}\{\text{expit}(0.3)+\text{expit}(0)\})$	$=0.995$
Example 3	$\text{logit}(\frac{1}{4}\{\text{expit}(3)+2\times\text{expit}(2)+\text{expit}(1)\})$ $-\text{logit}(\frac{1}{4}\{\text{expit}(2)+2\times\text{expit}(1)+\text{expit}(0)\})$	$=0.927$
Example 4	$\text{logit}(\frac{1}{4}\{\text{expit}(2.3)+\text{expit}(2)+\text{expit}(1.3)+\text{expit}(1)\})$ $-\text{logit}(\frac{1}{4}\{\text{expit}(1.3)+\text{expit}(1)+\text{expit}(0.3)+\text{expit}(0)\})$	$=0.952$
Example 5	$\text{logit}(\frac{1}{8}\,\text{expit}(1)+\frac{3}{8}\,\text{expit}(2)+\frac{3}{8}\,\text{expit}(3)+\frac{1}{8}\,\text{expit}(4))$ $-\text{logit}(\frac{1}{8}\,\text{expit}(0)+\frac{3}{8}\,\text{expit}(1)+\frac{3}{8}\,\text{expit}(2)+\frac{1}{8}\,\text{expit}(3))$	$=0.915$
	IV estimand $=\frac{1}{\alpha_1}\,\text{plim}_{N\to\infty}\,\hat{\beta}_{GY}$	
Example 1	1	$=1.000$
Example 2	$\frac{0.3}{0.3}$	$=1.000$
Example 3	$\text{logit}(\frac{1}{2}\{\text{expit}(2)+\text{expit}(1)\})$ $-\text{logit}(\frac{1}{2}\{\text{expit}(1)+\text{expit}(0)\})$	$=0.953$
Example 4	$\frac{1}{0.3}\{\text{logit}(\frac{1}{2}\{\text{expit}(1.3)+\text{expit}(0.3)\})$ $-\text{logit}(\frac{1}{2}\{\text{expit}(1)+\text{expit}(0)\})\}$	$=0.946$
Example 5	$\text{logit}(\frac{1}{4}\{\text{expit}(3)+2\times\text{expit}(2)+\text{expit}(1)\})$ $-\text{logit}(\frac{1}{4}\{\text{expit}(2)+2\times\text{expit}(1)+\text{expit}(0)\})$	$=0.927$

Table 4.5: Population log odds ratio (PLOR) and scaled limit of regression coefficient for IV in logistic regression of outcome on IV in infinite sample (IV estimand) for five example scenarios of IV estimation

due to the IV. Hence, the PLOR is different from 1, but $\text{plim}_{N\to\infty}\hat{\beta}_{GY} = 1$. In the first example, the IV causes a 5% attenuation, whereas in the second case with a weaker instrument, the attenuation is ten times smaller. In examples 3 and 4, the PLOR and $\text{plim}_{N\to\infty}\hat{\beta}_{GY}$ are both attenuated from 1. In example 3, there is an appreciable difference between the two, whereas in example 4 with less difference in the phenotype due to the IV, they are close. In example 3, the IV estimand is 0.953, the same as the PLOR in example 1; the heterogeneity in both cases is due to a single random variable with the same distribution: in example 1 the variable G for the PLOR, and in example 3 the variable V for $\frac{1}{\alpha_1}\text{plim}_{N\to\infty}\hat{\beta}_{GY}$.

In example 5, we note that $\mathbb{E}\left[\frac{\hat{\beta}_{GY}}{\hat{\beta}_{GX}}\right] \neq \frac{\mathbb{E}[\hat{\beta}_{GY}]}{\mathbb{E}[\hat{\beta}_{GX}]}$, and so we cannot make any conclusion about the expected value of the IV estimator in a finite sample without considering the joint distribution of $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$. Running the model of example 5 across 100 000 simulations with a sample size of 100, we obtained a mean two-stage estimate of 0.9488 (Monte Carlo error: 0.0012); with a sample size of 1000, mean estimate 0.9296 (0.0004); with a sample size of 10 000, mean estimate 0.9275 (0.0001). This compares with the true value of $\text{plim}_{N\to\infty}\hat{\beta}_{GY}$ of 0.9273. As the sample size increases, the impact of the correlation between the numerator and denominator on the IV estimate reduces, and the IV estimate is closer to the ratio of probability limits of the two regression coefficients, the IV estimand.

We conclude that the PLOR and IV estimand are not the same, as the IV estimand is conditional on the IV and the PLOR is not. However, when the variation in the phenotype is small, the difference between the estimands may be small.

4.4.3 IV estimation in more realistic simulated scenarios

To investigate how the IV estimator behaves in more realistic situations, we simulate data from a logistic model (4.29) (same model as (4.11) in Section 4.3.3) for confounded association with a single instrument.

$$\begin{aligned} x_i &= \alpha_0 + \alpha_1 g_i + \alpha_2 v_i + \epsilon_i \\ \text{logit}(\pi_{1i}) &= \beta_0 + \beta_1 x_i + \beta_2 v_i \\ y_{1i} &\sim \text{Binomial}(1, \pi_{1i}) \\ v_i &\sim \mathcal{N}(0,1), \epsilon_i \sim \mathcal{N}(0, \sigma_x^2) \text{ independently} \end{aligned} \tag{4.29}$$

We take a large sample size of 4000 divided equally into two groups ($g_i = 0, 1$). The parameter $\alpha_1 = 0.3$ with $\sigma_x^2 = 1$ corresponds to a strong instrument with mean F statistic in the regression of X on G of around 45. We set $\alpha_0 = 0, \alpha_2 = 1, \beta_0 = -2$ and consider

three values for β_1 of 0.4, −0.8 and 1.2 and seven values for β_2 of −1.0, −0.6, −0.2, 0, 0.2, 0.6, 1.0 corresponding to different levels and directions of confounding. We perform 2 500 000 simulations for each set of parameter values.

We estimate the observational log odds ratio by logistic regression of outcome on the phenotype X with no adjustment for confounding. The PLOR and IV estimand (plim $\hat{\beta}_1^R$) are calculated using both numerical integration as per equation (4.7) and the Monte Carlo approach of equation (4.10); identical answers are produced by both approaches. Using IVs, we calculate the two-stage estimate and the adjusted two-stage estimate. The adjusted two-stage estimate is calculated by regressing the outcome Y on both the fitted values $\hat{X}|G$ and the residuals from the first stage regression $R = X - \hat{X}|G$. These residuals are unbiased scaled estimators of the covariate V, which is considered unknown, and so including these in the second-stage regression is thought to give a better estimate of the ILOR (which is β_1) (94; 131).

Confounded association		$\beta_2 = -1.0$	$\beta_2 = -0.6$	$\beta_2 = -0.2$	$\beta_2 = 0$	$\beta_2 = 0.2$	$\beta_2 = 0.6$	$\beta_2 = 1.0$
$\beta_1 = 0.4$	Observational	-0.0887	0.1012	0.3005	0.4003	0.4978	0.6780	0.8279
	PLOR	0.3721	0.3893	0.3893	0.3828	0.3721	0.3405	0.3031
	IV estimand	0.3749	0.3907	0.3907	0.3848	0.3748	0.3443	0.3068
	Two-stage method	0.3751	0.3911	0.3907	0.3852	0.3751	0.3447	0.3066
	Adjusted two-stage	0.3760	0.3921	0.3992	0.4005	0.3994	0.3899	0.3703
$\beta_1 = -0.8$	Observational	-1.1977	-1.0662	-0.8967	-0.8004	-0.6995	-0.4919	-0.2876
	PLOR	-0.5387	-0.6062	-0.6721	-0.7004	-0.7234	-0.7500	-0.7500
	IV estimand	-0.5248	-0.5903	-0.6557	-0.6852	-0.7098	-0.7394	-0.7394
	Two-stage method	-0.5256	-0.5919	-0.6567	-0.6848	-0.7103	-0.7396	-0.7403
	Adjusted two-stage	-0.7419	-0.7794	-0.7991	-0.8005	-0.7988	-0.7823	-0.7542
$\beta_1 = 1.2$	Observational	0.6531	0.8773	1.0981	1.2009	1.2953	1.4529	1.5651
	PLOR	0.9527	0.9163	0.8544	0.8185	0.7813	0.7080	0.6403
	IV estimand	0.9851	0.9477	0.8831	0.8451	0.8056	0.7276	0.6558
	Two-stage method	0.9859	0.9482	0.8832	0.8456	0.8059	0.7276	0.6558
	Adjusted two-stage	1.1124	1.1664	1.1968	1.2012	1.1970	1.1650	1.1094

Table 4.6: Observational log odds ratio, population log odds ratio (PLOR) and IV estimand compared to two-stage and adjusted two-stage estimates of log odds ratio for unit increase in phenotype from model of confounded association. Median estimates across 2 500 000 simulations

Table 4.6 shows the observational log odds ratio, PLOR and IV estimand, and median estimates across simulations of the two-stage and adjusted two-stage methods. We see that the observational estimate is biased in the direction of the confounded association (β_2). The two-stage method estimates are attenuated compared to the conditional causal

effect, but close to the IV estimand and PLOR throughout. The difference between the two-stage estimate and the PLOR is due to the conditioning on G; the IV estimand, which is marginal in V and conditional on G is closer to the average two-stage estimate. The difference between the PLOR and IV estimand is however not large compared to that between the PLOR and ILOR. The adjusted two-stage method estimates are closer to the ILOR, with some attenuation when there is strong confounding, as the residuals measure variation in X not explained by G, which is the confounders plus error ($\alpha_2 v_i + \epsilon_i$).

A further set of simulations was conducted with the same parameters using Model (4.30), which is identical to the above model except with independent covariates U and V for the phenotype and outcome. This means that the association between X and Y is no longer confounded. The residual R is no longer related to the relevant covariate V in the second-stage logistic regression, but instead the variation in X not explained by G ($\alpha_2 u_i + \epsilon_i$).

$$\begin{aligned} x_i &= \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i \\ \text{logit}(\pi_{1i}) &= \beta_0 + \beta_1 x_i + \beta_2 v_i \\ y_{1i} &\sim \text{Binomial}(1, \pi_{1i}) \\ u_i, v_i &\sim \mathcal{N}(0,1), \epsilon_i \sim \mathcal{N}(0, \sigma_x^2) \text{ independently} \end{aligned} \tag{4.30}$$

Results are given in Table 4.7. We see that the PLOR and IV estimand are close throughout, and the median two-stage method is closest to the IV estimand as before. The observational estimate is an individual odds ratio, so conditional on X, but marginal in the unmeasured V as the model is misspecified when $\beta_2 \neq 0$, and so the observational estimate is attenuated compared to the ILOR even though there is no confounding (195) (see Section 4.3.2). The median adjusted two-stage estimate is more attenuated than in the previous example (128), and is not different to the observational estimate. This is because adjustment is made for the error term $\alpha_2 u_i + \epsilon_i$ in X, meaning that the odds ratio is conditional on all variation in X except that caused by G. Except for this variation in G, this is an individual odds ratio marginal in V, which is the same as the observational estimate.

4.4.4 Interpretation of the adjusted two-stage estimand

In an idealized setting, where the first-stage residual is precisely the correct term to adjust for in the second-stage regression, the adjusted two-stage approach is consistent for the ILOR (127). In Model (4.29), this would occur if $\sigma_x^2 = 0$. However, when this is not true, the adjusted two-stage estimate is attenuated (128). In the situation where none of the covariates for Y are associated with variation in X (i.e. there is no confounding), the

Unconfounded association		$\beta_2 = -1.0$	$\beta_2 = -0.6$	$\beta_2 = -0.2$	$\beta_2 = 0$	$\beta_2 = 0.2$	$\beta_2 = 0.6$	$\beta_2 = 1.0$
$\beta_1 = 0.4$	Observational	0.3494	0.3811	0.3980	0.4001	0.3981	0.3811	0.3493
	PLOR	0.3335	0.3637	0.3806	0.3828	0.3807	0.3637	0.3335
	IV estimand	0.3373	0.3669	0.3828	0.3848	0.3828	0.3669	0.3374
	Two-stage method	0.3381	0.3672	0.3834	0.3852	0.3832	0.3673	0.3375
	Adjusted two-stage	0.3499	0.3812	0.3985	0.4008	0.3984	0.3814	0.3496
$\beta_1 = -0.8$	Observational	-0.6961	-0.7592	-0.7958	-0.8006	-0.7958	-0.7593	-0.6960
	PLOR	-0.6266	-0.6721	-0.6972	-0.7004	-0.6972	-0.6721	-0.6265
	IV estimand	-0.6102	-0.6559	-0.6818	-0.6852	-0.6818	-0.6558	-0.6102
	Two-stage method	-0.6107	-0.6562	-0.6824	-0.6855	-0.6823	-0.6564	-0.6102
	Adjusted two-stage	-0.6964	-0.7595	-0.7963	-0.8008	-0.7962	-0.7598	-0.6960
$\beta_1 = 1.2$	Observational	1.0333	1.1326	1.1928	1.2009	1.1926	1.1323	1.0334
	PLOR	0.7513	0.7922	0.8155	0.8185	0.8154	0.7923	0.7513
	IV estimand	0.7737	0.8172	0.8419	0.8451	0.8419	0.8173	0.7737
	Two-stage method	0.7746	0.8172	0.8419	0.8453	0.8424	0.8177	0.7741
	Adjusted two-stage	1.0348	1.1322	1.1932	1.2003	1.1929	1.1326	1.0334

Table 4.7: Observational log odds ratio, population log odds ratio (PLOR) and IV estimand compared to two-stage and adjusted two-stage estimates of log odds ratio for unit increase in phenotype from model of unconfounded association. Median estimates across 2 500 000 simulations

residual in the adjusted two-stage method adjusts for the variation in X independent of that explained by the IV, leading to an estimate close to a marginal individual odds ratio. However, in such a scenario, the same estimate could be obtained by direct regression of Y on X. A more realistic situation is where some of the variation in X is due to covariates associated with Y, but not all. This corresponds to Model (4.29) with $\sigma_x^2 \neq 0$. Here, the residual is a combination of the independent variation in X and the covariate V, meaning that the adjusted two-stage analysis estimates an effect which is an odds ratio, but conditional on some unknown combination of variation in X and V. If there were additional covariates in Y not associated with X, as in Model (4.30), the odds ratio would be marginal in these covariates. When the covariates are unknown, as is usual in a Mendelian randomization study, it is not clear what odds ratio is being estimated by an adjusted two-stage approach. We return to this question of interpretation of IV estimates in the discussion.

4.4.5 IV estimation in five studies

We use a Mendelian randomization approach for the five studies from Section 4.3.4 viewed as cross-sectional studies using three or four SNPs in the CRP gene region as IVs to

estimate the causal association of log(CRP) on prevalent MI. We estimate the causal effect using the two-stage and adjusted two-stage methods, as well as a two-stage analysis adjusting for covariates in the first- and second-stage regressions. The covariates adjusted for in each study were the same as in the maximally adjusted model for each study in Table 4.4. If adjustment is made for a particular covariate in one stage of an IV analysis, it should be made in both stages (118). As the CRP levels were measured after the event, there is a possibility of bias in this analysis due to reverse causation. We therefore also perform a two-stage analysis using the CRP values only in non-cases, using the G-X association to give fitted values for cases. An adjusted two-stage method is here not possible, as residuals cannot be defined for cases except using CRP levels measured post event.

	SNPs used [1]	Two-stage method	Adjusted two-stage	Two-stage with covariate adjustment [2]
			CRP in all participants	
PROCARDIS	g1, g2, g4, g6	0.044 (0.172)	0.043 (0.175)	0.204 (0.194)
LURIC	g1, g2, g4, g6	-0.011 (0.251)	-0.011 (0.255)	-0.049 (0.254)
SHEEP	g1, g2, g7	0.231 (0.277)	0.240 (0.282)	0.188 (0.340)
CHS	g1, g3, g4, g5	0.352 (0.322)	0.352 (0.322)	0.214 (0.323)
ROTT	g1, g2, g6	0.299 (0.383)	0.306 (0.385)	0.326 (0.396)
			CRP in non-cases only	
PROCARDIS	g1, g2, g4, g6	0.038 (0.181)	-	0.205 (0.206)
LURIC	g1, g2, g4, g6	-0.042 (0.213)	-	-0.058 (0.207)
SHEEP	g1, g2, g7	0.139 (0.249)	-	0.058 (0.299)
CHS	g1, g3, g4, g5	0.303 (0.316)	-	0.170 (0.315)
ROTT	g1, g2, g6	0.270 (0.388)	-	0.303 (0.403)

Table 4.8: Estimates (SE) of causal association of log(CRP) on myocardial infarction (MI) from two-stage, adjusted two-stage methods, and two-stage method with adjustment for measured covariates in five studies

[1] g1 = rs1205, g2 = rs1130864, g3 = rs1417938, g4 = rs1800947, g5 = rs2808630, g6 = rs3093068, g7 = rs3093077

[2] Adjustment is made in each study for covariates as per the maximally adjusted model in Table 4.4

We note that these estimates of causal association (Table 4.8) are somewhat different to the observational associations estimated in Table 4.4. This indicates that the association between CRP and CHD may not be causal, although there are wide confidence intervals. In each study, the causal estimate decreases (that is becomes more negative) when CRP values are taken in non-cases only, indicating there may be some reverse causation, but

that confounding seems to be the main cause of the observational association. In some studies, there is a decrease in standard error of the causal effect despite the omission of half the data on CRP, indicating that the model of genetic association may be better estimated on the non-diseased subset of the population. Estimates of both individual and population causal effects test the same null hypothesis, and so assuming a model of null association, the two-stage and adjusted two-stage estimates should be similar with the adjusted estimate slightly larger in magnitude, as is the case here.

4.5 Discussion

In this chapter, we have seen how odds ratios differ depending on their exact definition. The magnitude of an odds ratio corresponding to an intervention depends on the choice of adjustment for competing risk factors, even if these are not confounders (i.e. not associated with the phenotype), and on whether the estimate is for an individual or population change in the phenotype. This is due to non-collapsibility of the odds ratio. This effect is especially severe when there is considerable between-individual heterogeneity for risk of event. When there is confounding, instrumental variable methods can be used to target a quantity close to the population odds ratio in a two-stage approach. The population odds ratio is similar to the incident odds ratio from an idealized RCT with intervention corresponding to a unit population intervention on the phenotype. By including the residuals from the first-stage regression in the second-stage analysis, an adjusted two-stage approach targets an odds ratio which is closer to the target parameter from a traditional multivariate regression analysis, the individual odds ratio conditional on all covariates. However there is attenuation from the individual odds ratio when there is variation in X not explained by covariates for Y or variation in the probability of Y not associated with variation in X. It is not clear for a general specification of the model what odds ratio is estimated by an adjusted two-stage approach.

4.5.1 Connection to existing literature and novelty

The appropriateness of the two-stage and adjusted two-stage methods have been the subject of recent discussion. Terza et al. (127) advocated adjusted two-stage methods as unbiased under certain circumstances as discussed in Section 4.4.4, as opposed to unadjusted two-stage methods, which are biased under all circumstances. Cai et al. (128) question the unbiasedness of the adjusted two-stage method, and provide independently the same derivation of the two-stage estimate as presented here in equation (4.23). This chapter

adds to the debate by interpreting the estimate from the two-stage method as a population effect, interpreting the estimate from the adjusted two-stage method as marginal in a certain combination of covariates, and by separating the issues of collapsibility into those due to unmeasured confounding and those due to intervention in the entire phenotype distribution. This is an important issue in Mendelian randomization, where the intervention is usually on a continuous phenotype, as opposed to in clinical trials, the context of the Terza and Cai papers, where the phenotype tends to be dichotomous.

4.5.2 Choice of target effect estimate

Generally, a population causal effect marginal across all covariates is the estimate of interest for a policy-maker as it represents the effect of intervention on the phenotype at a population level (153; 196). This is the effect estimated by a RCT without adjustment for covariates (197). However, the mathematical properties of population and marginal odds ratios are not as nice as those of the individual odds ratio conditional on all covariates, in that their attenuation from the coefficient β_1 in the underlying model depends on the size of the intervention, the amount of variation in the phenotype and the distribution of the covariates for the outcome. As the IV estimate corresponds to a change in the phenotype scaled by the effect of the instrument on the phenotype, it is advisable in IV analyses to quote odds ratios scaled for an increase (or decrease) in phenotype of comparable size to the size of the effect of the instrument on the phenotype.

In order to estimate an individual odds ratio conditional on all covariates in a logistic regression or two-stage IV analysis, it is necessary to measure and adjust for all covariates. An adjusted two-stage approach targets an odds ratio conditional on the phenotype and on some combination of covariates which are associated with the phenotype. In most applications of Mendelian randomization, there will be some correlation between the covariates for the phenotype and outcome, as otherwise, a causal association could be estimated using conventional regression methods. However, it is unlikely that all the covariates for the outcome constitute all of the variation in the phenotype, and so it is unclear what effect is estimated by an adjusted two-stage analysis. For this reason, although there is mathematical interest in the adjusted method, an untestable and usually implausible assumption of a specific form of the error structure is required for interpretation of the adjusted two-stage estimator, and so its use should not be recommended in applied practice. A better alternative would be to use the same covariates in the first- and second-stage IV regressions as in the observational analysis, so that under the hypothesis of no unmeasured confounding, the same conditional association is estimated in both analyses.

In a logistic regression, adjustment for covariates does not necessarily increase precision of the regression coefficients (198) and the decision of which covariates to adjust for should be guided by both understanding of the underlying model and desired interpretation of the effect estimate (195; 199). If we desire to estimate a population effect marginal across covariates then the two-stage method would seem appropriate. If estimation of a conditional parameter is desired, adjustment can be made for specific covariates.

4.5.3 "Forbidden" regressions

Much of the criticism of two-stage methods for IV estimation with non-linear models in econometric circles centres around the question of consistency of the estimator (187). Although consistency is a desirable property, it would seem to be a less important property than, say, coverage under the null. The work in this chapter suggests that the problem of consistency is one of interpretation of the IV estimate, rather than one of intrinsic bias of the estimate. As all odds ratios test the same null hypothesis, while caution should be expressed in comparing the magnitude of odds ratios estimating different quantities, it seems that there is no justification in labelling all such regressions as "forbidden" for reasons of consistency. This is especially true as some non-linear functions are collapsible, and so do not suffer from the problems highlighted in this chapter.

4.5.4 Different designs, different parameters

Table 4.9 summarizes how an odds ratio depends on the design and analysis of the study. We note that not all sources of bias have been included in this table (eg. non-compliance or treatment contamination in a RCT, canalization in an IV analysis). Nevertheless, it provides a useful summary of odds ratios estimated in different study designs and analyses.

A RCT and an instrumental variable approach target similar population-based parameters. For example, a study into effectiveness of invasive cardiac management on MI survival showed that an IV analysis gave results which were most similar to results from a RCT, compared to analyses using multivariable adjustment, propensity score adjustment, and propensity-based matching (197). However the estimands of the population effect of an intervention in phenotype of equal size in a RCT and an IV analysis may not be the same. This is because the RCT estimate is based on the difference in outcome caused by a short-term intervention, whereas, in the example of Mendelian randomization, the estimate is based on the difference in outcome caused by a life-long intervention due to the genetic variant. It has been argued that the Mendelian randomization estimate will be larger in magnitude than the RCT estimate (47), although this may be affected by

developmental compensation (also known as canalization) (3). This is compensation for the effect of the genetic variation on the phenotype by developmental processes which damp or buffer the genetic effect (2). For example, Mendelian randomization analysis of the effect of cholesterol on CHD have shown greater effects than RCTs (200).

Another reason why different answers may be obtained from analysis of a RCT and an IV approach is measurement error. IVs were initially conceived to deal with measurement error rather than confounding (111). Ratio IV estimates are not attenuated by measurement error, as the ratio IV method is symmetric in X and Y, and the G-X association is estimated. Estimates from conventional regression analysis are attenuated by measurement error, and correction for regression dilution bias would be necessary to ensure that the two estimands were the same (201).

It is tempting in Mendelian randomization studies to "claim the null hypothesis" of no causal effect by demonstrating that the causal effect of a phenotype on an outcome as estimated by Mendelian randomization is not compatible with the expected effect based on the observational effect (69; 85). Not only is this not valid as there may be a true causal effect smaller in magnitude than the observational association, but the two odds ratios may be estimating different quantities, making a test of equality of effects invalid.

In summary, the two-stage method has been criticized for a lack of theoretical basis and for giving inconsistent estimates even under the true model (99; 153). We have shown that this inconsistency is a property not of the two-stage approach, but of logistic regression in general, and can be partially rectified under certain assumptions by use of the adjusted method, or better, can be properly explained by correct interpretation of the causal effect.

4.5.5 Key points from chapter

- Odds ratio estimates for a binary outcome depend on the choice of covariates conditioned on and whether the odds ratio is for the change in phenotype for an individual or across a population.

- The two-stage IV analysis targets a parameter termed the 'IV estimand', a population odds ratio marginal across all covariates except the IV, which represents the population-averaged effect of an intervention in the phenotype averaged across covariate strata. It can be thought of as the estimate from an idealized RCT.

- The IV estimand and the estimate from the idealized RCT are similar in magnitude, and both attenuated compared to the individual odds ratio conditional on all covariates.

Method and analysis	Parameter of interest	Bias
Observational study, - no adjustment	Crude odds ratio	Biased due to confounding and reverse causation,
Observational study, - adjusted for all covariates	Individual odds ratio	None (assuming no measurement error, model correctly specified, etc.)
Observational study, - adjusted for known covariates	Individual odds ratio	Biased if there is residual confounding or reverse causation, OR is conditional on covariates included in model, marginal in others
Randomized controlled trial - no adjustment for confounders	Population odds ratio	None, effect corresponds to short-term intervention
Instrumental variable analysis - two-stage method	Population odds ratio	None, OR is conditional on IV, marginal in other covariates, effect may correspond to longer-term intervention
Instrumental variable analysis - adjusted two-stage method	Marginal individual odds ratio	Consistent for the individual OR under very specific assumptions. OR is conditional on variation in X not explained by G; hence conditional on some combination of covariates associated with X and independent error in X

Table 4.9: Summary of odds ratios (ORs) estimated by different study designs and analysis methods and possible sources of bias

- Adjustment can be made for specific covariates to estimate an odds ratio conditional on those covariates, and an adjusted method can be used to estimate an odds ratio which is generally closer to the individual odds ratio, but only interpretable based on a specific assumption about the error structure. The adjusted two-stage method is not recommended for use in practice.

Chapter 5

A Bayesian framework for instrumental variable analysis

5.1 Introduction

Our purpose in this chapter is to extend existing methods for instrumental variable (IV) analysis of Mendelian randomization studies to the context of multiple genetic markers measured in multiple studies, based on analysis of individual participant data (IPD).

We consider first the case where the outcome is continuous, and then consider binary outcomes. Several methods are available to estimate the causal association of a phenotype (X) on an outcome (Y) by use of an IV (G) in the presence of arbitrary confounding by a confounder (U) (see Chapter 2 for a review).

We seek to add to these established methods by introducing a Bayesian method. The main motivation for the method is to gain power by using data from multiple studies. We seek to use multiple, potentially different, SNPs simultaneously in each of these studies to obtain the most precise estimate possible of causal association by using all the available genetic data, while avoiding the problems of weak instruments. We recall from Chapter 3 that IV estimates using a weak instrument, where the association between phenotype and the IV is not statistically strong, suffer bias in the direction of the original observational association and deviation from a normal to a more heavy-tailed distribution.

We describe a Bayesian approach to the estimation of causal effects using genetic IVs. We present the simple case of a single genetic marker in one study (Section 5.2), and extend this to an analysis of multiple genetic markers in one study (Section 5.3). A hierarchical model for meta-analysis is then developed (Section 5.4) which efficiently deals with different genetic markers measured in different studies, and with heterogeneity between studies. The methods are exemplified by data on the causal association of C-reactive

protein (CRP) on fibrinogen from the CRP CHD Genetics Collaboration (CCGC). We continue to consider a similar model for binary outcomes (Section 5.5). Specific extensions associated with evidence synthesis and efficient analysis for the CCGC data are proposed (Section 5.6). The applied focus in this chapter is on the continuous outcome case for single studies and meta-analysis; the main analysis for the CRP-CHD causal association is presented in Chapter 8. We conclude by briefly discussing some of the features of the Bayesian framework for IV analysis (Section 5.7); this will be considered further with extensive simulation and comparison to alternative methods in Chapter 6.

5.2 Continuous outcome — A single genetic marker in one study

We consider in turn methods appropriate for use with a continuous outcome, and then for use with a binary outcome.

5.2.1 Conventional methods

We first consider the case of a single SNP in one study, where confounding causes the observational estimate of the association of phenotype and outcome to be different from the causal relationship. Let individual i have phenotype level x_i, outcome y_i, genotype g_i taking values 0,1,2, and unmeasured confounder u_i. We assume that all the confounders can be summarized by a single value u_i. Similarly to Palmer et al. (94), we consider the model represented in Figure 5.1:

$$\begin{aligned} x_i &= \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_{xi} \\ y_i &= \beta_0 + \beta_1 x_i + \beta_2 u_i + \epsilon_{yi} \\ u_i &\sim \mathcal{N}(0, \sigma_u^2), \epsilon_{xi} \sim \mathcal{N}(0, \sigma_1^2), \epsilon_{yi} \sim \mathcal{N}(0, \sigma_2^2) \text{ independently} \end{aligned} \tag{5.1}$$

As an example, we simulate data for a sample of size 300, containing 12 individuals with $g_i = 2$, 96 with $g_i = 1$ and 192 with $g_i = 0$, corresponding to Hardy-Weinberg equilibrium for a minor allele frequency of 20%. We set the parameters $(\alpha_0, \alpha_2, \beta_0, \beta_1, \beta_2, \sigma_u^2, \sigma_1^2, \sigma_2^2)$ = (0, 1, 0, 2, −3, 1, 0.25, 0.25), and consider the cases of a weak instrument ($\alpha_1 = 0.3$, giving an expected F statistic for the regression of X on G of 7), a moderate instrument

($\alpha_1 = 0.5$, F statistic 20) and a strong instrument ($\alpha_1 = 1$, F statistic 75):

$$\begin{aligned} x_i &= \alpha_1 g_i + 1u_i + \epsilon_{xi} \\ y_i &= 2x_i - 3u_i + \epsilon_{yi} \\ u_i &\sim \mathcal{N}(0,1), \epsilon_{xi} \sim \mathcal{N}(0, 0.25), \epsilon_{yi} \sim \mathcal{N}(0, 0.25) \text{ independently} \end{aligned} \tag{5.2}$$

Figure 5.2 shows the simulated data grouped by genotype graphically. For each of the three genotypic groups, the mean of the phenotype and outcome with 95% confidence intervals (CIs) are plotted. This shows how the genotypic groups differ on average in phenotype, and how the mean outcome differs as a result of the phenotype differences.

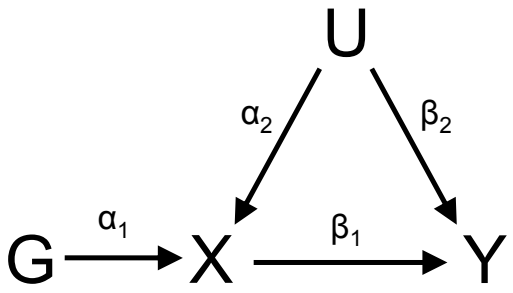


Figure 5.1: Directed acyclic graph (DAG) of Mendelian randomization assumptions

The observational estimates obtained by regressing Y on X (Table 5.1) are far from the true causal association ($\beta_1 = 2$) as expected because of the strong negative confounding (U is positively related to X but negatively to Y). The ratio method (assuming zero correlation between coefficients) gives estimates compatible with $\beta_1 = 2$, but with a wide confidence interval in the case of the weak or moderate instrument. Sensitivity analyses taking values for correlation of $\pm 0.1, \pm 0.2$ gave similar wide asymmetric confidence intervals.

5.2.2 A Bayesian method

Estimating the causal parameter by the ratio method is equivalent to determining the gradients in Figure 5.2 (2). We can reformulate the problem as one of linear regression with heterogeneous error in X. For each genotype value $j = 0, 1, 2$ we calculate the mean level of the phenotype $\bar{x}_j$ with its variance σ^2_{xj} and mean outcome $\bar{y}_j$ with its variance σ^2_{yj}. The model is

$$\begin{aligned} \bar{X}_j &\sim \mathcal{N}(\xi_j, \sigma^2_{xj}) \\ \bar{Y}_j &\sim \mathcal{N}(\eta_j, \sigma^2_{yj}) \\ \eta_j &= \beta_0 + \beta_1 \xi_j \end{aligned} \tag{5.3}$$

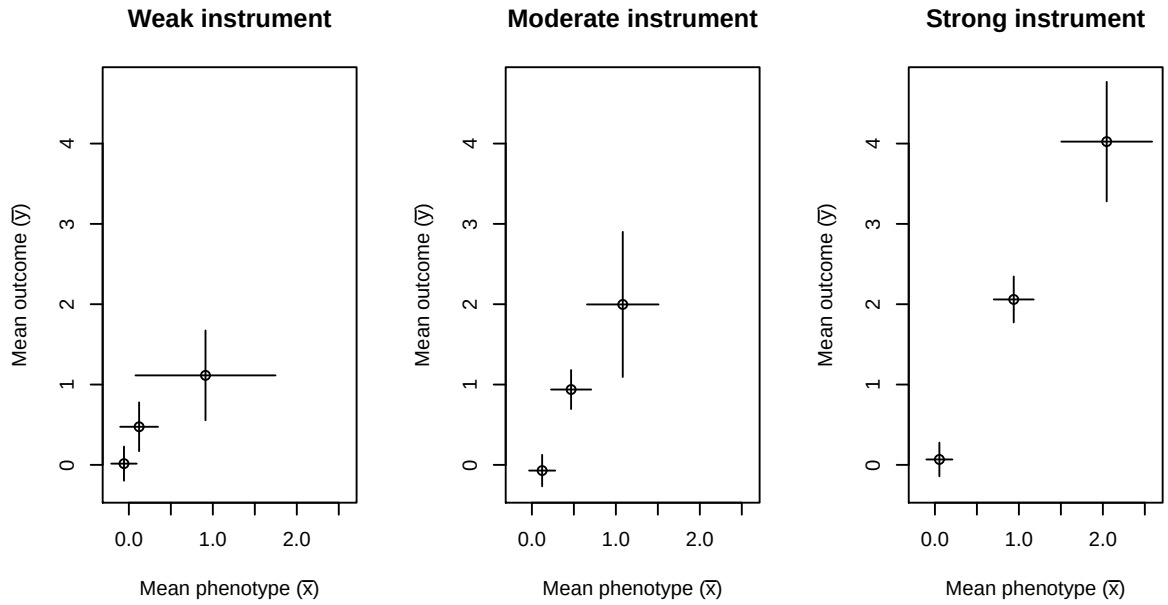


Figure 5.2: Graphs of mean outcome ($\bar{y}$) against mean phenotype ($\bar{x}$) in three genetic groups for the weak, moderate and strong instrument simulated examples of Section 5.2.1. Error bars are 95% CIs for the means

Thus we assume that each observed mean phenotype $\bar{x}_j$ is from a normal distribution with unknown true mean ξ_j and known variance σ^2_{xj}, each observed mean outcome $\bar{y}_j$ is from a normal distribution with unknown true mean η_j and known variance σ^2_{yj}, and there is a linear relationship between η and ξ. β_1 represents the increase in outcome for unit increase in true phenotype and is the parameter of interest.

To implement this model, we employ Bayesian analysis and Markov Chain Monte Carlo (MCMC) methods with Gibbs sampling. This allows extension to more complicated situations, as in the next sections. We used vague priors (independent normals with zero mean and large variance of 100^2) for the regression parameters and each ξ_j. We performed this analysis in WinBUGS (202) using 150 000 iterations, discarding the first 1000 as "burn-in", employing different starting values to assess convergence of the posterior distribution and sensitivity analyses to show lack of dependence on the prior distributions. The posterior distributions shown in Figure 5.3 are non-normal, with a heavier tail towards larger values especially for the weaker instruments. For this reason, the posterior median of the distribution of β_1 is taken as the estimate of the causal association. Table 5.1 shows that the estimates and intervals from this Bayesian group-based method are similar to those from the ratio method. Other simulated examples (not shown) also demonstrated similar results. The 2SLS method (assuming linear effect of the IV on the phenotype) gives

the same estimates as the ratio method, but the intervals are symmetric and so deviate from the ratio and Bayesian methods for the weaker instruments. In particular here, the confidence intervals for the 2SLS method with the weak instrument include zero; the ratio and Bayesian intervals both exclude zero.

A difference between the ratio and Bayesian method (5.3) is that the ratio method assumes a linear association of the genetic variant and phenotype with a constant increase in mean phenotype for each copy of the variant allele (here called a "per allele" model), whereas the Bayesian method (5.3) models the mean phenotype separately for each number of variant alleles (here called a two-degree of freedom or "2df" model). We shall see that the Bayesian and 2SLS methods can incorporate either per allele or 2df models for the G-X association.

Weak instrument - ($\mathbb{E}(F) = 7$)	Estimate	95% CI/CrI
Observational estimate	-0.358	-0.506, -0.210
Ratio method	1.637	0.563, 6.582
2SLS method	1.637	-0.126, 3.400
Bayesian method	1.496	0.536, 7.190
Moderate instrument - ($\mathbb{E}(F) = 20$)	Estimate	95% CI/CrI
Observational estimate	-0.251	-0.393, -0.109
Ratio method	2.555	1.481, 6.007
2SLS method	2.555	0.801, 4.309
Bayesian method	2.417	1.473, 4.592
Strong instrument - ($\mathbb{E}(F) = 75$)	Estimate	95% CI/CrI
Observational estimate	0.108	-0.061, 0.276
Ratio method	2.136	1.632, 2.906
2SLS method	2.136	1.469, 2.804
Bayesian method	2.107	1.633, 2.817

Table 5.1: Causal parameter estimates and confidence/credible intervals using ratio, 2SLS and Bayesian methods compared with observational estimate for the weak, moderate and strong instrument simulated examples of Section 5.2.1

This Bayesian method assumes that the variances σ^2_{xj} and σ^2_{yj} are known, whereas in fact they need to be estimated from the data, an issue which is addressed in the next section.

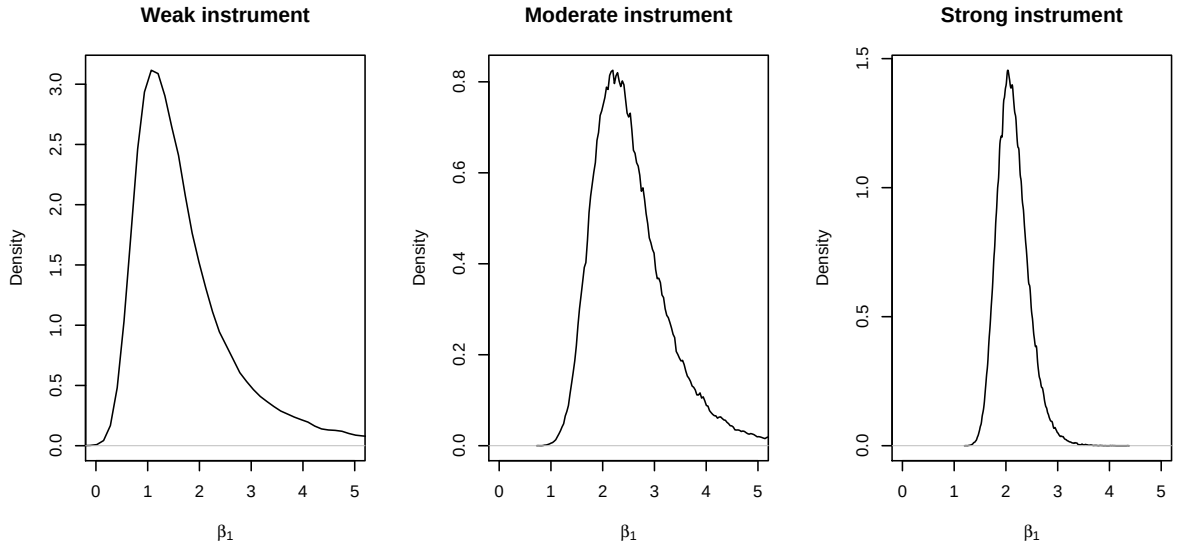


Figure 5.3: Kernel-smoothed density of posterior distribution of the causal parameter for the weak, moderate and strong instrument simulated examples of Section 5.2.1 using the group-based Bayesian method of Section 5.2.2

5.3 Continuous outcome — Multiple genetic markers in one study

5.3.1 Methods

If we have data in the study from more than one SNP then, provided they satisfy the IV assumptions above, all SNPs can be used simultaneously to divide the population into many subgroups. For each diallelic SNP, there are three genotypic subgroups, corresponding to 0, 1 or 2 variant alleles. For a dataset with K diallelic SNPs, we have a maximum 3^K subgroups, for each of which we can measure the mean phenotype and outcome, and examine the regression as in (5.3) above to estimate β_1, the causal association. In practice, fewer than the maximum number of genotypic groups will be observed, due to linkage disequilibrium (LD) between SNPs.

If the number of groups is large, and so their sizes N_j are small, then the assumption of exact knowledge of σ^2_{xj} and σ^2_{yj} for each group is not appropriate. Indeed if $N_j = 1$, a group-specific estimate of variance cannot even be expressed. It is then preferable to base the analysis on the standard deviation in the whole population for the phenotype (σ_x) and the outcome (σ_y), using an individual-based model for phenotype and outcome. For each

individual i in subgroup j, we have

$$\begin{aligned} X_{ij} &\sim \mathcal{N}(\xi_j, \sigma_x^2) \\ Y_{ij} &\sim \mathcal{N}(\eta_j, \sigma_y^2) \\ \eta_j &= \beta_0 + \beta_1 \xi_j \end{aligned} \tag{5.4}$$

The observed phenotype and outcome for each individual are here modelled using normal distributions, although other distributions might be more appropriate for some applications. The information about ξ_j now depends on the population standard deviation for the phenotype as well as the size of the group. In the application below, vague Uniform[0,20] priors are used for σ_x and σ_y, while the other priors remain as before.

An alternative analysis is to assume a linear relationship between the phenotype and the number of variant alleles for each SNP which is also additive across SNPs. If this structure is appropriate, the analysis should be more efficient as the correlation between similar genotypes is accounted for and fewer parameters are estimated. Then we use these modelled values in the second-stage regression.

$$\begin{aligned} \xi_i &= \alpha_0 + \sum_k \alpha_k \, g_{ik} \\ X_i &\sim \mathcal{N}(\xi_i, \sigma_x^2) \\ Y_i &\sim \mathcal{N}(\eta_i, \sigma_y^2) \\ \eta_i &= \beta_0 + \beta_1 \xi_i \end{aligned} \tag{5.5}$$

where g_{ik} is the number of variant alleles in SNP k for individual i, and α_k are the first-stage genetic regression coefficients. Independent vague $\mathcal{N}(0, 100^2)$ priors are now placed on the α_k rather than the ξ_i. The values of the α and β regression parameters depend, through feedback, on all the data including the outcome Y.

Models (5.4) and (5.5) are the equivalent of 2SLS in a Bayesian setting, except that there is feedback on the first-stage coefficients from the second-stage regression; the posterior distribution of the causal association parameter β_1 naturally incorporates the uncertainty in the first-stage regression, but with no assumption of asymptotic normality on its distribution. The models are also analogous to the likelihood-based FIML/LIML, except that here the correlation between X and Y is set to be zero; we discuss the role of this correlation further in Chapter 6.

5.3.2 Application to C-reactive protein and fibrinogen

C-reactive protein (CRP) is an acute-phase protein produced by the liver as part of the inflammation response pathway. Fibrinogen is a soluble blood plasma glycoprotein, which

enables blood-clotting and is also associated with inflammation. The pathway of inflammation is not well understood, but is important as both CRP and fibrinogen are proposed as risk markers of coronary heart disease (CHD) (82). Furthermore, although CRP is associated with CHD risk, this association reduces on adjustment for various risk factors, and attenuates to near null on adjustment for fibrinogen (84). It is important therefore to assess whether CRP causally affects levels of fibrinogen, since if so adjusting for fibrinogen would represent an overadjustment. The CRP gene has several common variations which are associated with different blood concentrations of CRP. We use IV techniques to estimate the causal effect of CRP on fibrinogen. As CRP has a positively skewed distribution, we take its natural logarithm, and assume a linear relationship between fibrinogen and log(CRP). All SNPs used here as IVs are in the CRP regulatory gene on chromosome 1.

The Cardiovascular Health Study (203) is an observational study of risk factors for cardiovascular disease in adults 65 years or older. We use cross-sectional baseline data for 4469 white subjects from this study, in which four diallelic SNPs relevant to CRP were measured: rs1205, rs1800947, rs1417938 and rs2808630. Each of these SNPs was found to be associated with CRP levels. We checked their associations with seven known CHD risk factors (age, body mass index, triglycerides, systolic blood pressure, total cholesterol, low and high density lipoproteins) for each SNP, and found no significant associations ($P < 0.05$) out of the 28 examined. This suggests that the SNPs are valid instruments.

We used the ratio, 2SLS, and Bayesian methods using models (5.3), (5.4) and (5.5) for estimating causal associations. The ratio method for each SNP separately is based on per allele regressions. For the 2SLS method, we use first a per allele model additive across SNPs and secondly a fully factorial version of the 2df model where each observed genotype is placed in a separate subgroup. The 2SLS per allele model is equivalent to the structural-based Bayesian model (5.5) and the 2SLS factorial model is equivalent to the individual-based Bayesian model (5.4). When using the group-based regression (5.3), we excluded all genotypic groups with less than 5 subjects (14 subjects excluded, Figure 5.4). The individual-based (5.4), structural-based (5.5), ratio and 2SLS analyses include all subjects. A sensitivity analysis was performed excluding from the 2SLS factorial and Bayesian individual-based analyses all individuals from genotypic groups with less than 5 subjects. The observational increase in fibrinogen (μmol/l) per unit increase in log(CRP) is 0.937 (s.e. 0.024) and correlation between fibrinogen and log(CRP) is 0.501. The $F_{4,4464}$ statistic in the regression of log(CRP) on the SNPs additively per allele is 27.2, indicating that the instruments together are moderately strong (92; 161). As we have used more IVs than we have phenotypes, we can perform an overidentification test. The Sargan test (158) is a test of the validity of the IV and linearity assumptions in the model. The test

statistic is 7.15, which compared to a χ^2_3 distribution gives a p-value of 0.067, meaning that the validity of the instruments is not rejected at the 5% level.

The ratio method gives a different point estimate for each SNP, all of which are compatible with zero association (Table 5.2). Using the 2SLS methods on all of the SNPs together, we obtain answers which synthesize all of the relevant data for each of the SNPs. The Bayesian methods give causal estimates consistent with the 2SLS estimates (Table 5.2). The Bayesian structural-based and 2SLS per allele models give lower estimates of causal association than the other models, with 95% CIs that include zero. The Bayesian credibility intervals are (appropriately) asymmetric, as no normal assumption has been made. The Bayesian individual-based and the 2SLS factorial methods both give different results when individuals from small genotypic groups are excluded. The direction of the differences in the estimates is consistent with weak instrument bias.

Method	Estimate	95% CI
Ratio using rs1205	0.234	-0.169 to 0.660
Ratio using rs1417938	-0.608	-1.581 to 0.137
Ratio using rs1800947	0.203	-0.478 to 0.940
Ratio using rs2808630	2.722	$-\infty$ to ∞
2SLS factorial using all SNPs	0.376	0.088 to 0.665
2SLS factorial (excluding small groups)	0.280	-0.041 to 0.601
2SLS per allele using all SNPs	0.200	-0.138 to 0.538
Bayesian methods	Estimate	95% CrI
Group-based (excluding small groups)	0.342	0.004 to 0.698
Individual-based	0.389	0.049 to 0.728
Individual (excluding small groups)	0.300	-0.045 to 0.666
Structural-based	0.212	-0.157 to 0.586

Table 5.2: Comparison of the causal estimates of increase in fibrinogen (μmol/l) per unit increase in $\log_e$(CRP) in the Cardiovascular Health Study. 95% confidence/credible intervals (CI/CrI) are shown. Small groups are genotypic groups with less than 5 subjects

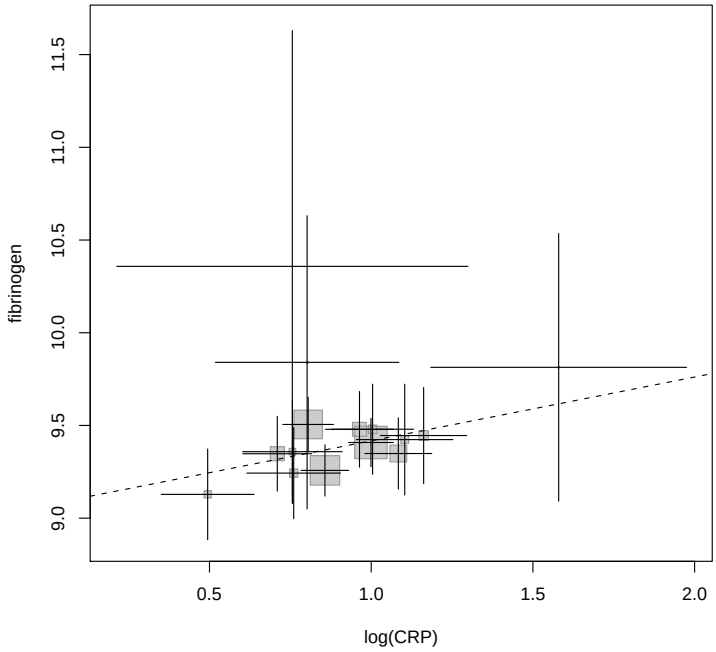


Figure 5.4: Plot of mean fibrinogen against mean log(CRP) in the Cardiovascular Health Study stratified by genotypic group. Error bars are 95% CIs. Groups with less than 5 subjects omitted. The size of the shaded squares is proportional to the number of subjects in each group. The dashed line is the estimate of causal association from the group-based method

5.4 Continuous outcome — Multiple genetic markers in multiple studies

5.4.1 Methods

The above framework leads naturally to a model for meta-analysis across multiple studies. IV assumption iii. in Sections 1.2.2 and 2.2 states that the IV is conditionally independent of the outcome given the phenotype and confounders. This ensures that, in principle, the same parameter β_1 is being estimated regardless of how many and which SNPs are available in each study. This is because the outcome is independent of the IV given the phenotype (which is measured) and the confounders (which are averaged over). We thus propose a hierarchical model for β_1 estimated across multiple studies as follows. For a fixed-effect meta-analysis, we assume the same value of β_1 for each study. For a random-effects meta-analysis, we allow β_{1m} from study m to come from a distribution with mean β_1 and variance

ψ^2. This acknowledges the possibility that the causal parameters are somewhat different across studies, as is plausible due to the influences of different population characteristics, but that they are expected to have generally similar values.

For the group-based regression (5.3), for group j in study m, a fixed-effect meta-analysis is:

$$\begin{aligned} \bar{X}_{jm} &\sim \mathcal{N}(\xi_{jm}, \sigma^2_{xjm}) \\ \bar{Y}_{jm} &\sim \mathcal{N}(\eta_{jm}, \sigma^2_{yjm}) \\ \eta_{jm} &= \beta_{0m} + \beta_1 \, \xi_{jm} \end{aligned} \tag{5.6}$$

Values for β_{0m}, the constant terms in the regression, will vary depending on the average level of outcome in the population in each study, and are thus given independent vague $\mathcal{N}(0, 100^2)$ priors for each study.

For a random-effects meta-analysis, the last line of (5.6) is replaced by:

$$\begin{aligned} \eta_{jm} &= \beta_{0m} + \beta_{1m} \, \xi_{jm} \\ \beta_{1m} &\sim \mathcal{N}(\beta_1, \psi^2) \end{aligned} \tag{5.7}$$

We use a Uniform[0,20] prior for ψ in the example below.

These modifications to the simple group-based analysis (5.3) for a meta-analysis context can also be similarly made to the individual-based model (5.4), and to the structured model (5.5). For example, the full model using a structured model (5.5), assuming heterogeneity between studies, for individual i and SNP $k = 1 \ldots K_m$ in study $m = 1, \ldots, M$ is:

$$\begin{aligned} \xi_{im} &= \alpha_{0m} + \sum_{k=1}^{K_m} \alpha_{km} \, g_{ikm} \\ X_{im} &\sim \mathcal{N}(\xi_{im}, \sigma^2_{xm}) \\ Y_{im} &\sim \mathcal{N}(\eta_{im}, \sigma^2_{ym}) \\ \eta_{im} &= \beta_{0m} + \beta_{1m} \, \xi_{im} \\ \beta_{1m} &\sim \mathcal{N}(\beta_1, \psi^2) \end{aligned} \tag{5.8}$$

The standard deviation parameters $(\sigma_{xm}, \sigma_{ym})$ are given independent priors. In this model, we assume that the first-stage regression coefficients α_{km} are unrelated in the different studies. An extra sophistication would be to assume that these coefficients are common or related when different studies involve the same set of SNPs (see Section 5.6.2). Example WinBUGS code is given in the appendix to this chapter.

5.4.2 Application to C-reactive protein and fibrinogen

We give an example of meta-analysis of eleven studies (82) using the methods described. In addition to the Cardiovascular Health Study (CHS) used in Section 5.3.2, we incorporate data from a further eight general population cohort studies: British Women's Heart and Health Study (BWHHS), Copenhagen City Heart Study (CCHS), Copenhagen General Population Study (CGPS), English Longitudinal Study of Ageing (ELSA), Framingham Health Study (FRAM), Northwick Park Heart Study II (NPHS2), Rotterdam Study (ROTT), and Whitehall II Study (W2). In each of these the analyses presented here are cross-sectional, based on baseline measurements of CRP and fibrinogen. We also use data from two case-control studies, the Nurses' Health Study (NHS) and Stockholm Heart Epidemiology Program (SHEEP), again with CRP and fibrinogen measured at baseline. We use the data from controls alone since these better represent cross-sectional population studies. Details of these studies are summarized in Table 5.3.

To avoid problems with weak instruments, we want to choose genetic instruments which together are strongly related to log(CRP). For this, the instrument was chosen to maintain the F statistic above 10 and to include sequentially, where available, each of SNPs rs1205, one of rs1130864 and rs1417938 (these SNPs are in complete LD), rs3093077, rs1800947 and rs2808630. In the meta-analysis we use between 2 and 4 SNPs as instruments in each study; the Sargan overidentification tests were satisfied (Table 5.3). The choice of instruments here is not made a priori, as should ideally be the case, but pragmatically to exemplify the method. For comparison with the Bayesian methods, we use the study-specific 2SLS causal estimates and corresponding asymptotic standard errors in a standard two-step inverse variance weighted meta-analysis (using a moment estimator of the between-study variance in the case of random-effects meta-analysis). Mean log(CRP) and fibrinogen levels for the genotypic groups for six of the studies are shown in Figure 5.5. We note that the treatment of the two-stage method is not the same as that of the Bayesian method, as the two-stage results are combined in a two-step summary effects meta-analysis rather than an one-step IPD meta-analysis. A two-step approach is used as it is difficult to specify an error structure for the phenotype in a possible hierarchical two-stage analysis, and because a two-step analysis is usually used in practice, and so provides a more relevant comparison than a one-step method.

Table 5.4 shows a causal association of log(CRP) on fibrinogen which does not significantly differ from the null, except for the structural-based fixed-effect meta-analysis, which suggests a weak negative causal association. Groups of size less than 5 have been omitted in the 2SLS factorial, group-based and individual-based analyses. There is no clear preference for the random-effects models from the Deviance Information Criterion (DIC)

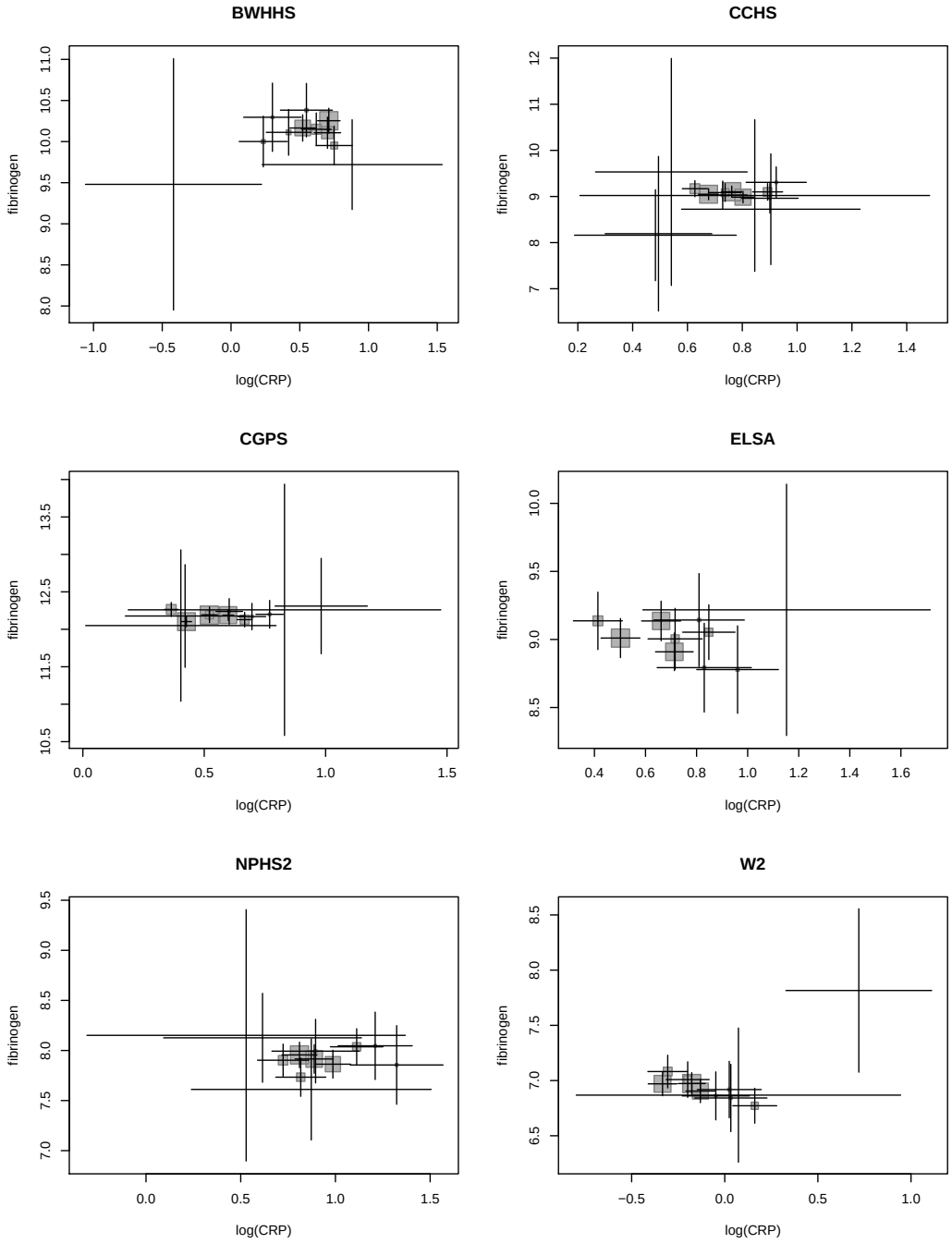


Figure 5.5: Plot of mean fibrinogen against mean log(CRP) for six studies from Section 5.4.2 stratified by genetic group. Error bars are 95% CIs. Groups with less than 5 subjects omitted. The size of the shaded squares is proportional to the number of subjects in each group.

Study	SNPs used [1] as IV	Participants	Excluded	F statistic	df	Overidentification p-value
BWHHS	g1, g3, g5	3188	7	16.7	(3, 3184)	0.638
CCHS	g1, g2, g4	7998	5	29.6	(3, 7994)	0.358
CGPS	g1, g2, g4	35679	5	152.0	(3, 35675)	0.439
CHS	g1, g3, g5, g6	4469	15	27.2	(4, 4464)	0.067
ELSA	g1, g2, g4	4409	8	24.7	(3, 4405)	0.367
FRAM	g1, g2, g4	1575	4	10.0	(3, 1571)	0.447
NHS	g1, g6	414	0	13.2	(2, 411)	0.984
NPHS2	g1, g2, g4	2153	3	11.6	(3, 2149)	0.344
ROTT	g1, g2	2077	2	11.9	(2, 2074)	0.983
SHEEP	g1, g2, g4	1044	4	10.5	(3, 1040)	0.680
W2	g1, g2, g4	4354	5	21.5	(3, 4350)	0.469
Total		67361	58			

Table 5.3: Summary of studies in meta-analysis of Section 5.4.2: SNPs used as instrumental variable (IV), number of participants with complete genetic data, number of participants in genotypic groups of size less than 5 excluded from some analyses, F value with degrees of freedom (df), p-value from Sargan test of overidentification from additive per allele regression of phenotype on SNPs used as IVs

[1] g1 = rs1205, g2 = rs1130864, g3 = rs1417938, g4 = rs3093077, g5 = rs1800947, g6 = rs2808630

(204). The DIC should only be used to compare between a fixed- or random-effect model, and not between models based on different data structures. Again, the structural-based models give lower estimates of causal association than the other methods.

5.5 Binary outcome — Genetic markers in one study

We now consider methods for use with a binary outcome, assuming a logistic model of association and targeting an odds ratio parameter. A log-linear model could also be considered; in this case a relative risk parameter would be estimated.

5.5.1 Conventional methods

We again consider the case of a single SNP in one study, where confounding causes the observational estimate of the association of phenotype and outcome to be different from the causal relationship. Let individual i have phenotype level x_i, outcome y_i, genotype

Fixed-effect meta-analysis	Estimate	95% CI/CrI	DIC [1]	
2SLS factorial	-0.005	-0.139 to 0.130		
2SLS per allele	-0.086	-0.255 to 0.082		
Group-based	-0.008	-0.142 to 0.125	-242.1	
Individual-based	-0.036	-0.164 to 0.090	500692	
Structural-based	-0.136	-0.276 to -0.002	501037	
Random-effects meta-analysis	Estimate	95% CI/CrI	DIC	ψ
2SLS factorial	-0.007	-0.151 to 0.137		0.072
2SLS per allele	-0.086	-0.255 to 0.082		0.000
Group-based	-0.017	-0.234 to 0.177	-244.5	0.188
Individual-based	-0.039	-0.228 to 0.153	500692	0.155
Structural-based	-0.150	-0.365 to 0.048	501037	0.169

Table 5.4: Estimates of increase in fibrinogen (μmol/l) per unit increase in log(CRP), 95% confidence/credible interval (CI/CrI), deviance information criterion (DIC) and heterogeneity parameter (ψ) in meta-analysis of eleven studies using 2SLS and Bayesian methods. Genotypic groups with less than 5 individuals excluded from the 2SLS factorial, group-based and individual-based analyses

[1]We note that DIC should be used to compare between a fixed- or random-effect model and not between models.

g_i taking values 0,1,2, and unmeasured confounder u_i. We consider the model of logistic association:

$$\begin{aligned} x_i &= \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_{xi} \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 x_i + \beta_2 u_i \\ y_i &\sim \text{Binomial}(1, \pi_i) \\ u_i &\sim \mathcal{N}(0, \sigma_u^2), \epsilon_{xi} \sim \mathcal{N}(0, \sigma_1^2) \text{ independently} \end{aligned} \tag{5.9}$$

As an example, we simulate data for a sample of size 1200, containing 48 individuals with $g_i = 2$, 384 with $g_i = 1$ and 768 with $g_i = 0$, corresponding to Hardy-Weinberg equilibrium for a minor allele frequency of 20%. We consider the same parameter values as in Section 5.2.1 above except for $\beta_0 = -2$: $(\alpha_0, \alpha_2, \beta_0, \beta_1, \beta_2, \sigma_u^2, \sigma_1^2) = (0, 1, -2, 2, -3, 1, 0.25)$. Setting $\beta_0 = -2$ ensures a large but realistic number of cases, as the probability of an event for an individual with $x_i = 0, u_i = 0$ is expit$(-2) = 0.12$. We consider the cases of a weak instrument ($\alpha_1 = 0.15$, giving an expected F statistic for the regression of X on G of 7), a moderate instrument ($\alpha_1 = 0.25$, F statistic 20) and a strong instrument

($\alpha_1 = 0.5$, F statistic 75):

$$\begin{aligned} x_i &= \alpha_1 g_i + 1u_i + \epsilon_{xi} \\ \text{logit}(\pi_i) &= -2 + 2x_i - 3u_i \\ u_i &\sim \mathcal{N}(0,1), \epsilon_{xi} \sim \mathcal{N}(0, 0.25) \text{ independently} \end{aligned} \tag{5.10}$$

Figure 5.6 shows the simulated data grouped by genotype graphically. The standard error for the log odds of an event in each group has been estimated using a normal approximation.

The observational estimates obtained by regressing Y on X (Table 5.5) are far from the true causal association ($\beta_1 = 2$) as expected because of the strong negative confounding (U is positively related to X but negatively to Y). The ratio method (assuming zero correlation between coefficients) gives estimates compatible with $\beta_1 = 2$, but with a wide confidence interval in the case of the weak or moderate instrument. Sensitivity analyses taking values for correlation of $\pm 0.1, \pm 0.2$ gave similar wide asymmetric confidence intervals.

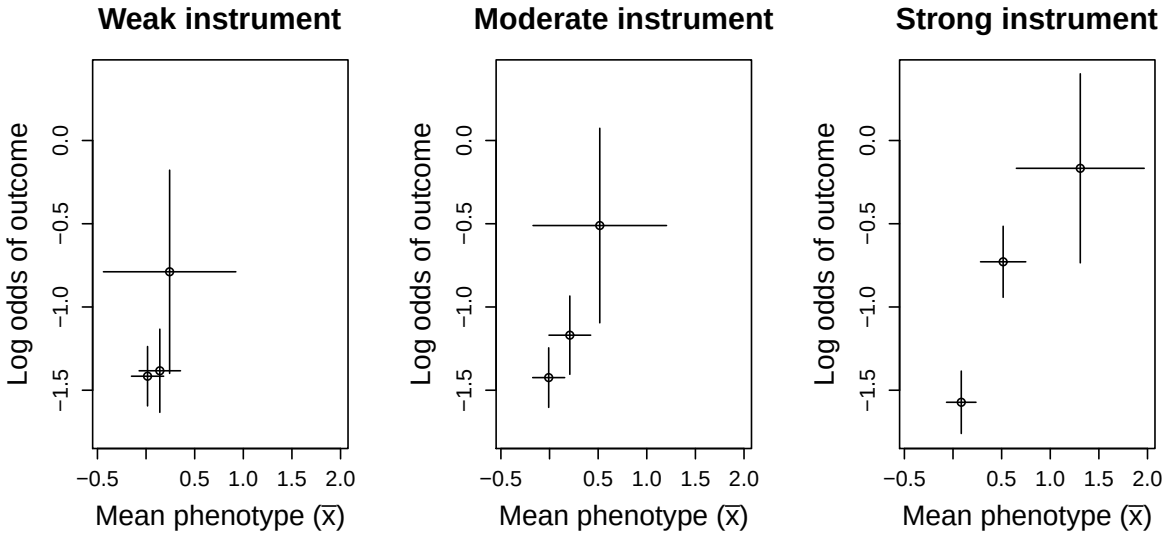


Figure 5.6: Graphs of log odds of event against mean phenotype ($\bar{x}$) in three genetic groups for the weak, moderate and strong instrument simulated examples of Section 5.2.1. Error bars are 95% CIs for the mean and log odds

5.5.2 A Bayesian method

As in the continuous setting, we can reformulate the problem as one of linear regression with heterogeneous error in X. For each genotype value $j = 0, 1, 2$ we calculate the mean

level of the phenotype $\bar{x}_j$ with its variance σ^2_{xj} and log odds of event $\bar{y}_j$ with its asymptotic variance σ^2_{yj}. The model is

$$\begin{aligned} \bar{X}_j &\sim \mathcal{N}(\xi_j, \sigma^2_{xj}) \\ \bar{Y}_j &\sim \mathcal{N}(\eta_j, \sigma^2_{yj}) \\ \eta_j &= \beta_0 + \beta_1 \xi_j \end{aligned} \tag{5.11}$$

where β_1 represents the increase in log odds of event for unit increase in true phenotype and is the parameter of interest. This corresponds to the group-based regression from above.

Alternatively, we can model on an individual level. An individual-based model for phenotype and outcome can be constructed, using a normal distribution for the phenotype and a binomial distribution for the outcome with a logistic link function. Let the number of individuals in genotypic subgroup j be N_j and n_j be the number of them who have events. Then for each individual i in subgroup j, we have

$$\begin{aligned} X_{ij} &\sim \mathcal{N}(\xi_j, \sigma^2_x) \\ n_j &\sim \text{Binomial}(N_j, \pi_j) \\ \eta_j = \text{logit}(\pi_j) &= \beta_0 + \beta_1 \xi_j \end{aligned} \tag{5.12}$$

Equivalently, we would obtain the same model by taking the likelihood contributions to the binomial density for each individual separately:

$$\begin{aligned} Y_{ij} &\sim \text{Binomial}(1, \pi_{ij}) \\ \eta_{ij} = \text{logit}(\pi_{ij}) &= \beta_0 + \beta_1 \xi_j \end{aligned} \tag{5.13}$$

Models (5.12) and (5.13) correspond to the individual-based regression model (5.4) with a continuous outcome.

Finally, we can consider a structural-based model:

$$\begin{aligned} \xi_i &= \alpha_0 + \sum_k \alpha_k g_{ik} \\ X_i &\sim \mathcal{N}(\xi_i, \sigma^2_x) \\ Y_i &\sim \text{Binomial}(1, \pi_i) \\ \eta_i = \text{logit}(\pi_i) &= \beta_0 + \beta_1 \xi_i \end{aligned} \tag{5.14}$$

where g_{ik} is the number of variant alleles in SNP k for individual i. Equivalent models could be constructed to estimate a relative risk in a log-linear model, by replacing the logistic function with a logarithm function in each of the above models.

Results for the group-based, individual-based and structural-based Bayesian methods, as well as the confounded observational estimate and estimates from the ratio and two-stage IV methods are shown in Table 5.5. Posterior distributions for the structural-based Bayesian method are displayed as Figure 5.7. We see that the group-based method gives wider confidence intervals, but similar point estimates to the ratio/two-stage estimate. This is partially due to the lack of a linearity assumption in the gene-phenotype association. The estimate from the structural-based method did not converge for the weak instrument, but similar estimates to the ratio method are given especially with the strong instrument.

Weak instrument - ($\mathbb{E}(F) = 7$)	Estimate	95% CI/CrI
Observational estimate	-0.25	-0.38, -0.12
Ratio method	1.33	-0.83, 21.16
Two-stage method	1.33	-0.68, 3.33
Group-based Bayesian method	0.82	-12.13, 12.76
Individual-based Bayesian method	Did not converge	
Structural-based Bayesian method	Did not converge	
Moderate instrument - ($\mathbb{E}(F) = 20$)	Estimate	95% CI/CrI
Observational estimate	-0.15	-0.28, -0.03
Ratio method	1.48	0.47, 3.36
Two-stage method	1.48	0.49, 2.47
Group-based Bayesian method	1.44	-0.84, 8.61
Individual-based Bayesian method	1.62	0.48, 3.38
Structural-based Bayesian method	1.58	0.48, 3.80
Strong instrument - ($\mathbb{E}(F) = 75$)	Estimate	95% CI/CrI
Observational estimate	-0.10	-0.22, 0.02
Ratio method	1.55	1.04, 2.20
Two-stage method	1.55	1.10, 1.99
Group-based Bayesian method	1.56	0.93, 3.05
Individual-based Bayesian method	1.51	0.99, 2.14
Structural-based Bayesian method	1.57	1.06, 2.23

Table 5.5: Causal parameter estimates of $\beta_1 = 2$ and confidence/credible intervals using ratio, two-stage and Bayesian group-based, individual-based and structural-based methods compared with observational estimate for the weak, moderate and strong instrument simulated examples of Section 5.5.1

These methods can be naturally extended for meta-analysis of multiple studies by use

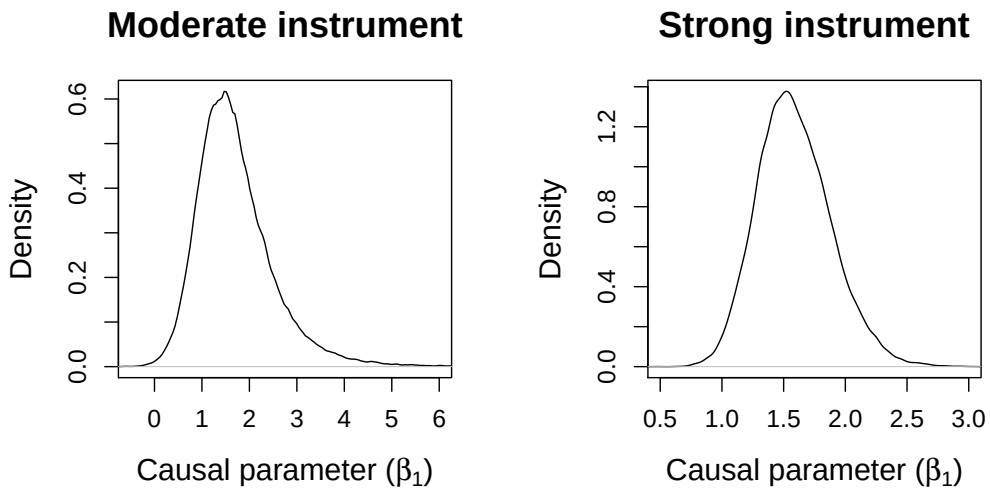


Figure 5.7: Kernel-smoothed density of posterior distribution of the causal parameter for the moderate and strong instrument simulated examples of Section 5.5.1 using the structural-based Bayesian method of Section 5.5.2

of a hierarchical model as in Model 5.8 for the continuous outcome setting:

$$\begin{aligned} \xi_{im} &= \alpha_{0m} + \sum_{k=1}^{K_m} \alpha_{km}\, g_{ikm} \\ X_{im} &\sim \mathcal{N}(\xi_{im}, \sigma^2_{xm}) \\ Y_{im} &\sim \text{Binomial}(1, \pi_{im}) \\ \eta_{im} = \text{logit}(\pi_{im}) &= \beta_{0m} + \beta_{1m}\, \xi_{im} \\ \beta_{1m} &\sim \mathcal{N}(\beta_1, \psi^2) \end{aligned} \tag{5.15}$$

Having introduced the Bayesian models for continuous and binary outcomes in this chapter, we discuss extensions to this model, before returning to consider the properties of the models under simulation in Chapter 6, where we discuss estimation and interpretation of the causal parameter β_1 in the light of the work in Chapter 3 (weak instrument bias) and Chapter 4 (non-collapsibility).

5.6 Dealing with issues of evidence synthesis in meta-analysis

In this section, we detail how the problems of combining evidence of heterogenous sources can be efficiently accomplished in the Bayesian models detailed above, with a focus on

specific features exhibited in the CCGC dataset. Aside from the first subsection on cohort studies, these extensions are relevant in both continuous and binary outcome cases.

5.6.1 Cohort studies

In a cohort study, if individuals are not excluded from study entry at baseline due to history of disease, each participant has two windows of opportunity to become a case: one before study entry and one after. We want to include participants in cohort studies up to twice in the analysis, once in the study viewed retrospectively and once prospectively. A cross-sectional or retrospective analysis is performed by viewing the cohort at baseline as a cross-sectional study with cases taken as individuals with previous history of disease (prevalent cases) and controls as all non-diseased individuals. A prospective analysis excludes all prevalent cases and considers events within the reporting period. An individual who is censored at the end of the follow-up period is taken as a control in both the retrospective and prospective analyses as he has two separate opportunities to become a case. However, we do not want to include the individual's phenotype twice, and we want to ensure that the same parameter is estimated in both analyses.

In the corresponding model (5.16), we consider genotypic subgroup j. This subgroup contains N_{1j} individuals, n_{1j} of whom are prevalent cases, and $N_{2j}(= N_{1j} - n_{1j})$ non-prevalent individuals, n_{2j} of whom have incident events.

$$\begin{aligned} X_{ij} &\sim \mathcal{N}(\xi_j, \sigma^2) \text{ for } i = 1, \ldots N_{2j} \text{ non-prevalent individuals} \\ n_{1j} &\sim \text{Binomial}(N_{1j}, \pi_{1j}) \\ n_{2j} &\sim \text{Binomial}(N_{2j}, \pi_{2j}) \\ \text{logit}(\pi_{1j}) &= \eta_{1j} = \beta_{01} + \beta_1 \xi_j \\ \text{logit}(\pi_{2j}) &= \eta_{2j} = \beta_{02} + \beta_1 \xi_j \end{aligned} \tag{5.16}$$

This model ensures that the same fitted values of phenotype are used in both logistic regressions without including individuals twice in the regression of phenotype on genotype.

5.6.2 Common SNPs

Where the same subset of SNPs has been used in several studies, we can combine the estimates of genetic association α_{km} across studies. This should give a more precise model of association in smaller studies and should reduce weak instrument bias, as instrument strength will be combined across the studies. Due to possible heterogeneity between populations, we use a random-effects model, where we impose a multivariate normal distribution

on the study level parameters α_{km} with mean vector μ_α and variance-covariance matrix Ψ. Note that the intercept parameters α_{0m} are not pooled.

$$\begin{aligned} X_{ijm} &\sim \mathcal{N}(\xi_{jm}, \sigma_m^2) \\ \xi_{jm} &= \alpha_{0m} + \sum_{k=1}^{K} \alpha_{km}\, g_{jkm} \\ \alpha_{km} &\sim \mathcal{N}_K(\mu_\alpha, \Psi) \end{aligned} \tag{5.17}$$

5.6.3 Common haplotypes

Alternatively, we can model the phenotype additively across haplotypes as in model (5.18). Each individual has two haplotypes h_{1i} and h_{2i} and phenotype is modelled additively in a meta-analysis as the sum of three components, a study specific intercept γ_{0m} in study m and a component from each haplotype γ_{km} for haplotype k. The haplotype parameters are modelled as random-effects to allow for heterogeneity between genetic effects in each study. The study-specific estimates $\boldsymbol{\gamma}_m = (\gamma_{2m}, \ldots, \gamma_{Km})^T$ are modelled as being drawn from a multivariate normal distribution with mean μ_γ and variance-covariance matrix Ψ.

$$\begin{aligned} X_{im} &\sim \mathcal{N}(\xi_{jm}, \sigma_m^2) \\ \xi_{im} &= \gamma_{0m} + \gamma_{h_{1i}m} + \gamma_{h_{2i}m} \\ \boldsymbol{\gamma}_m &\sim \mathcal{N}_K(\mu_\gamma, \Psi) \end{aligned} \tag{5.18}$$

A multivariate normal distribution is assumed for each $\boldsymbol{\gamma}_m$. A multivariate prior is assumed for the mean vector μ_γ with mean $\mathbf{0}$ and diagonal variance-covariance matrix with 10 as each diagonal element, and a non-informative inverse-Wishart prior is assumed for Ψ, where the scale matrix in the Wishart distribution is diagonal with 10 as each diagonal element.

Due to collinearity from each individual having exactly two haplotypes, one of the haplotype effects (γ_{1m}) is arbitrary fixed to zero throughout. The parameter γ_{km} ($k = 2, \ldots, K$) is then interpreted as the increase in log(CRP) for an individual in study m having a copy of haplotype k relative to haplotype 1. As each of the γ_{km} is estimated relative to the effect of haplotype 1, it seems prudent to label the most common haplotype category as haplotype 1, to reduce the uncertainty in estimation of the γ_{km}, although this should not affect the overall causal estimate of β_1.

5.6.4 Lack of phenotype data

Where a study has not measured the phenotype (X) but has genetic data in common with other studies, we use the random-effects distributions for the genetic association parameters defined in Sections 5.6.2 or 5.6.3 as a predictive distribution or implicit prior for the unknown parameters. This requires an assumption of exchangeability that the change in phenotype per additional allele is similar (i.e. can be drawn from the same random-effects distribution) as the other studies. For identifiability, we set $\alpha_{0m} = 0$ in (5.17) or $\gamma_{0m} = 0$ in (5.18) as with no data on the G-X association, this parameter cannot be identified.

Incorporation of studies with information on only some of the gene–phenotype–outcome triangle needed for Mendelian randomization analysis is known in econometrics circles as the "two sample problem" (205).

5.6.5 Tabular data

For studies providing tabular data only, we were provided for each genetic subgroup j with binary outcome data on the total number of individuals (N_j) and the number with an event (n_j). We are able to incorporate such studies into an analysis using the random-effects distributions for the parameters of genetic association as above.

5.7 Discussion

In this chapter, we have described a Bayesian approach to analysis of Mendelian randomization studies. We introduced the approach in a simple example of a confounded association with one IV. We extended the method to use multiple IVs, to use individual participant data and to incorporate an explicit, here additive, genetic model. We then show how this leads naturally to a meta-analysis, which can be performed even with heterogeneous genetic data. These methods have been applied in the estimation of the causal association of CRP levels on fibrinogen.

5.7.1 Bayesian methods in IV analysis

The Bayesian approach has similarities to the 2SLS method. In both, fitted values of phenotype are estimated for each genotypic group, which are then used in a regression of outcome on phenotype. In 2SLS, these fitted values are assumed to be precisely known in the second-stage regression. In the Bayesian framework, the fitted values of phenotype

and outcome are estimated simultaneously, and the standard error in the causal parameter is directly estimated from the MCMC sampling process. This means that no assumption is made on the distribution of the causal parameter, giving appropriately sized standard errors and skew CIs. The Bayesian approach allows us to be explicit about the assumptions made. This gives us flexibility to determine the model according to what we believe is plausible without being limited to linear or normal assumptions.

Additionally, the Bayesian approach provides a framework to perform analyses that are difficult or not possible using 2SLS. These include meta-analysis in a single hierarchical model, imputation of missing data (see Chapter 7), inclusion of studies with partial information on the gene-phenotype-outcome associations, and hierarchical random-effects modelling of the first-stage genetic association parameters.

Bayesian methods have not been widely proposed for IV analyses or applied in Mendelian randomization studies. Although Bayesian methods for IV analysis have been suggested in the econometrics literature (134; 135), their use is not common and differences between the fields mean that methods cannot easily be translated into an epidemiological setting (31). McKeigue et al. (141) have performed a Bayesian analysis in the single SNP and single study situation, but regarding the parameter of interest as the "ratio of the causal effect to crude [i.e. observational] effect". We prefer to regard β_1, the causal association, as the parameter of interest.

5.7.2 Bayesian analysis as a likelihood-based method

Although this chapter focuses the advantages of a Bayesian IV framework, several of these advantages are inherited from the fact that the Bayesian methods examine the full likelihood of the model, and would be shared by other likelihood-based methods such as a full information maximum likelihood (FIML) approach. Such advantages include the propagation of uncertainty through the model. Indeed, it could be argued that the Bayesian method is not a truly Bayesian method, but simply a MCMC method. As the prior distributions are not informative, the Bayesian approach simply gives a sample from the posterior distribution, which approximates the likelihood. In a FIML approach, the mode of the likelihood is considered, rather than the median or mean usually considered in a Bayesian analysis, but otherwise the estimates will be similar. Similarly, bootstrapping could be used in non-Bayesian approaches like FIML to remove the dependence of inference for the causal effect on asymptotic assumptions.

A specific advantage of the Bayesian framework over a FIML approach is the possibility of the use of informative prior information in estimation. This is often important

for hyperparameters, such as the between-study variance ψ^2, where the information in the dataset on the parameter may be limited. Another advantage is the computational problems associated with FIML. Maximizing the likelihood of such a complex function is computationally expensive, especially in a meta-analysis context. It is not clear that such a likelihood would be unimodal. Bootstrapping to give robust confidence intervals may be theoretically possible, but impractical in large datasets. If a particular genetic variant had a low minor allele frequency, all individuals in a particular genotypic subgroup may be omitted from a given bootstrap sample, leading to possibly unidentifiable parameters. By contrast, the Bayesian approach is robust to these difficulties. Although the MCMC algorithm is computationally expensive, it is not prohibitively so. The Bayesian model can be fitted using standard software, meaning that diagnostics for convergence and fit are really available, whereas a FIML approach would need to be fitted 'by hand'.

Generally, we do not consider the FIML method in this chapter as it is not widely used in practice. One reason for this is that, even though asymptotic assumptions are not required for inference, the parametric and distributional assumptions made by fully likelihood methods, such as Bayesian and FIML methods, are strong and may be violated in practice. We discuss this trade-off further in Section 6.5.3.

5.7.3 Meta-analysis

Methods for meta-analysis of Mendelian randomization studies have not been extensively explored, and have been restricted to studies measuring one identical SNP (71; 89; 138). In applications, meta-analyses of studies have concentrated on testing for a causal effect, without accounting for the uncertainty in the estimated mean difference in phenotype values between genotypic groups (76; 172). Where this uncertainty has been accounted for, confidence intervals for the causal association have been too wide to exclude a moderate causal association (100; 174). Our proposed analysis thus extends this previous work in a number of ways: first by using a flexible Bayesian framework that eliminates the problems caused by non-normal causal estimates, second by presenting a coherent framework for estimation of the causal association using data from multiple studies, and third by allowing the use of different genetic markers in different studies.

An advantage of the Bayesian setting for meta-analysis is that the whole meta-analysis can be performed in one step. This keeps each study distinct within the hierarchical model, only combining studies at the top level. This is more effective at dealing with heterogeneity, both statistical and in study design, than performing separate meta-analyses on each of the genotype-phenotype and genotype-outcome associations (71). An alternative approach

where the causal association estimate and its precision are estimated in each study, and these estimates combined in a meta-analysis in a second stage, is not recommended for two reasons. First, the distribution of each causal estimate is not normal (especially if the instrument is not strong), and so the uncertainty is not well represented by its standard error, and secondly, some causal estimates from individual studies may have infinite variance. Examples of these problems are apparent in Figure 5.3 and Table 5.2.

5.7.4 Conclusion

The validity of IV analyses relies on assumptions specified in previous chapters. These assumptions can only be partially verified from data, and there are a number of ways in which they may be violated for Mendelian randomization studies (2). Nevertheless, this proposed Bayesian method for meta-analysis of Mendelian randomization studies is a useful methodological advance. It should also find application in the context of the increasing number of consortia that are now collating the relevant individual genetic, phenotype and outcome data from multiple studies (82).

5.7.5 Key points from chapter

- A Bayesian approach for IV analysis gives similar results to other established methods, while allowing extensions to analyses not possible in other frameworks.
- Estimation is possible with both continuous and binary outcomes and extension to a hierarchical meta-analysis model is natural.

Appendix: WinBUGS code

WinBUGS code for random-effects meta-analysis of group-based model

```
model {
# prior for hierarchical causal estimate (parameter of interest)
 betatrue ~ dnorm(0, 0.000001)
# prior for standard deviation of individual study estimates
 betasd ~ dunif(0, 20)
 betatau <- pow(betasd, -2)
for(m in 1:M) {         # M = number of studies
# prior for regression intercept parameter
```

```
  beta0[m] ~ dnorm(0, 0.000001)
# distribution of study-specific causal estimates
  beta[m] ~ dnorm(betatrue, betatau)
 for (j in 1:G[m]) {    # G[m] = number of genetic subgroups in study m
# distribution of phenotype in subgroup j, study m
   x[j, m] ~ dnorm(xi[j, m], xtau[j, m])
# distribution of outcome in subgroup j, study m
   y[j, m] ~ dnorm(eta[j, m], ytau[j, m])
# prior for true value of phenotype in subgroup j, study m
   xi[j, m] ~ dnorm(0, 0.000001)
# linear model of true outcome on true phenotype
   eta[j, m] <- beta0[m] + beta[m] * xi[j, m]
   } } }
```

WinBUGS code for fixed-effect meta-analysis of structural-based model

```
model {
# prior for fixed causal estimate (parameter of interest)
 beta ~ dnorm(0, 0.000001)
for(m in 1:M) {
# prior for regression intercept parameter
 beta0[m] ~ dnorm(0, 0.000001)
 alpha0[m] ~ dnorm(0, 0.000001)
# prior for study phenotype standard deviation
 xsd[m] ~ dunif(0, 20)
 xtau[m] <- pow(xsd[m], -2)
# prior for study outcome standard deviation
 ysd[m] ~ dunif(0, 100)
 ytau[m] <- pow(ysd[m], -2)
 for(k in 1:G[m]) {  # G[m] = number of genes in study m
# prior for gene-phenotype regression parameters
  alpha[k, m] ~ dnorm(0, 0.000001)
  }
 for (i in 1:N[m]) {  # N[m] = number of individuals in study m
# linear model of true phenotype on genes
  xi[i, m] <- inprod(alpha[1:G[m], m], gene[i, 1:G[m], m]) + alpha0[m]
```

```
# distribution of phenotype in individual i, study m
  x[i, m] ~ dnorm(xi[i, m], xtau[m])
# distribution of outcome in individual i, study m
  y[i, m] ~ dnorm(eta[i, m], ytau[m])
  eta[i, m] <- beta0[m] + beta * xi[i, m]
  } } }
```

Chapter 6

Improvement of bias and coverage in instrumental variable analysis

6.1 Introduction

In this chapter, we explore the bias and coverage properties of some commonly used methods for calculating instrumental variable (IV) estimates of causal association, and specifically the Bayesian methods introduced in Chapter 5.

We investigate two specific issues related to bias and coverage of estimates. The first is weak instrument bias (see Chapter 3) (101; 160). A weak instrument is an IV which does not explain a large proportion of the variation in the risk factor (102). Weak instruments are known to produce biased estimates with incorrectly sized confidence intervals (161).

The second issue is that of non-collapsibility in analyses involving binary outcomes and logistic modelling (see Chapter 4) (110). When a log odds ratio is marginalized over the distribution of a confounder, its value changes (33). So the interpretation of a regression parameter in a logistic association model depends on the distribution and choice of covariates in the model. With binary outcomes, several different parameters of interest and estimation methods have been proposed (97; 125). We seek to estimate an individual odds ratio conditional on all covariates, as this is the parameter targeted in a standard logistic regression analysis with adjustment for confounders (94; 153), and a population odds ratio marginal across all covariates (153), as this is the parameter typically estimated in a randomized controlled trial (109).

Although the results in this chapter can be applied generally to IV problems, the models and parameters of the simulations will correspond to those typical in a Mendelian randomization analysis. Specifically, we consider G as discrete and thus dividing the data into 'genetic subgroups'.

We first present data from the British Women's Heart and Health Study, one of the studies in the CRP CHD Genetics Collaboration (CCGC), to give a background to the estimation problem (Section 6.2). We give methods and a simulation study with continuous outcomes (Section 6.3), introducing a novel development in the continuous outcome Bayesian model introduced in Chapter 5 to model the observational correlation between risk factor and outcome, which reduces bias to near zero with even moderately weak instruments. We show methods for binary outcomes corresponding to those with continuous outcomes (Section 6.4), and demonstrate that adjusting for the first-stage residuals in a logistic model, which is similar to modelling the correlation in a continuous setting, changes the target parameter from a population odds ratio to an odds ratio conditional on variation in the phenotype. In the discussion, we relate these results to the analysis of Mendelian randomization studies (Section 6.5).

6.2 Example — British Women's Heart and Health Study

We motivate our methodological discussion using data from the British Women's Heart and Health Study (BWHHS) on C-reactive protein (CRP) and fibrinogen with complete data on three SNPs in the CRP coding region as IVs: rs1205, rs1130864, rs1800947. Although CRP and fibrinogen are positively correlated ($\hat{\beta} = 0.807$, SE $0.029, r = 0.45$), it is not thought that long-term variation in CRP is causally associated with increased levels of fibrinogen (140). As CRP has a skewed distribution, a linear association is assumed between log-transformed CRP and fibrinogen.

Figure 6.1 gives several graphical representations of the BWHHS data which will help us understand the requirements of methods for data analysis later in the chapter. The top-left graph shows the levels of log(CRP) and fibrinogen for all 3188 individuals in the study. The line plotted represents the observational association with 95% confidence interval obtained by linear regression. 122 individuals have CRP reported as 0.16 or 0.17 as this is the minimum level detectable by the assay used. A sensitivity analysis omitting these individuals made little difference to the overall results. The top-right graph shows the distribution of the mean of log(CRP) and fibrinogen for all individuals, estimated by a 1000 iterations of a non-parametric bootstrap. In each iteration, a sample of the population (with replacement) of the same size as the dataset is taken, and the mean of log(CRP) and fibrinogen are evaluated. Both graphs show a positive correlation between log(CRP) and fibrinogen.

The middle row of graphs shows the bootstrapped distributions of the mean of log(CRP) and fibrinogen for the group of individuals with each number of variant alleles of a SNP. We see that the within-group correlation (due to confounding) is represented by the direction of the major axis of the oval-shaped distribution, and the between-group causal effect is estimated by the regression line through the centres of these distributions. With one instrument, there are only three genetic groups, and the causal effect is not estimated precisely.

The bottom row of graphs shows the bootstrapped distributions of the mean of log(CRP) and fibrinogen for each of the genotypic groups based on each of the SNPs. The bottom-left graph illustrates the four groups containing more than 400 individuals and the bottom-right graph the nine groups containing more than 10 individuals (minimum groups size is 108). Each of these groups consists of all individuals with the same genotype across the three SNPs. We see that the correlation between the means of log(CRP) and fibrinogen is similar for each of the groups. The lines plotted in the bottom row represent the causal association with 95% confidence interval obtained from the 2SLS method. These lines through the means of the groups do not seem to have a clear positive or negative gradient. This visual inspection of the distribution of the means indicates that modelling the correlation between the means may be important and that the within-group correlation appears to be similar in each group.

These graphs provide an illustration of Mendelian randomization data. The observational correlation between phenotype and outcome, and the correlation between the mean phenotype and the mean outcome, are both positive. However, when the participants are divided into genotypic subgroups, as in the lower graphs, the causal effect is seen to be the gradient of the line through the means of phenotype and outcome for each group. Although the correlation between the mean phenotype and mean outcome for each group is strongly positive, this could be due to confounding. The gradient between the groups, representing the change in outcome for a unit change in phenotype where the confounder levels are the same in each group, is null. Assuming that the SNPs used as IVs are valid instruments, it is this between-group gradient which is the causal association.

6.3 Continuous outcomes and linear models

We describe both established and novel IV methods to estimate causal associations with continuous outcomes and linear models, and then examine how they perform in simulations. We are specifically interested in the bias and coverage properties of different

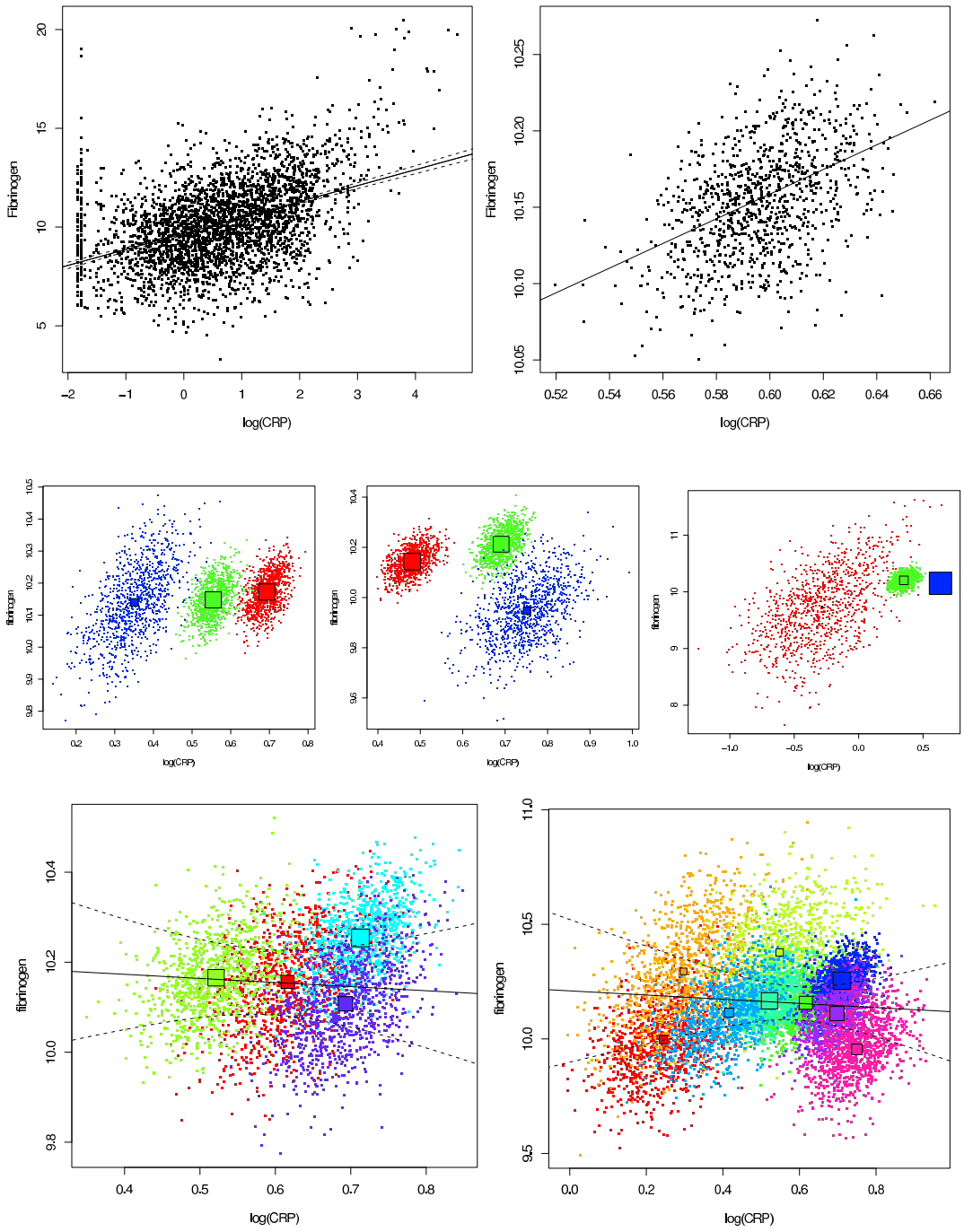


Figure 6.1: British Women's Heart and Health Study data on log-transformed CRP (mg/l) against fibrinogen (μmol/l): *top row* - left: raw data with observational association; - right: bootstrapped distribution of means of log(CRP) and fibrinogen for population with observational association; *middle row* - distribution of bootstrapped means for each number of variant alleles for - left: rs1205; - middle: rs1130864; right: rs1800947; *bottom row* - distribution of bootstrapped means for each genetic subgroup with estimate of causal association - left: four largest subgroups; right - nine largest subgroups (dashed lines are 95% confidence intervals throughout, area of squares is proportional to number of individuals in the group)

estimators. Apart from the Bayesian method introduced in Chapter 5, these methods were introduced in Chapter 2, and the salient features of the methods are recalled here.

6.3.1 Methods

a) Two-stage: The two-stage least squares (2SLS) estimator is so called because it can be calculated using two regression stages (93). The first stage (G-X regression) regresses X on G to give fitted values $\hat{X}|G$. The second stage (X-Y regression) regresses Y on the fitted values $\hat{X}|G$ from the first stage regression. The causal estimate is this second-stage regression coefficient for the change in outcome caused by unit change in the risk factor. The variance for the two-stage estimator with continuous outcomes is here calculated using a sandwich variance estimator to account for possible misspecification of the first-stage regression (121). The estimated causal parameter is generally assumed to be normally distributed (119). The 2SLS estimator has a finite kth moment with $(k+1)$ instruments when all the associations are linear and the error terms heteroscedastic and normally distributed (124).

Estimates using the 2SLS method are known to be biased in the direction of the confounded association between the risk factor and outcome (160; 206). The magnitude of the bias depends on the statistical strength of association between the instrument and risk factor (101). The relative bias, defined as the bias of the IV estimator divided by the bias of the observational estimator (ie. from ordinary least squares regression of Y on X) (102), is asymptotically approximately $1/F$, where F is the expected F statistic for the IVs in the first-stage regression (102; 161). Hence an expected F-value of 10 leads to a relative bias of about 10%. Weak instruments also lead to overly narrow confidence intervals and poor coverage properties (161), and methods which do not allow for the possibility of an infinite confidence set will not be robust to weak instruments (119). When the IV is weak, the IV estimator has a long-tailed distribution, which is not well approximated by a normal distribution (122).

b) Ratio: The ratio of coefficients (or Wald (111)) estimator is calculated as the ratio of two regression coefficients: from the regression of Y on G (G-Y regression) and the regression of X on G (G-X regression) (2). The ratio method can only be used when there is a single instrument, in which case the causal estimate coincides with that from the 2SLS method. Confidence intervals for the ratio estimator can be calculated using Fieller's theorem (100; 114), assuming a bivariate normal distribution of the regression estimates with zero correlation.

c) LIML: The limited information maximum likelihood (LIML) method is the "maximum likelihood counterpart of 2SLS" (132). It is calculated by a maximum likelihood procedure on the unrestricted reduced form (where each endogenous variable is expressed in terms of the exogenous variables) on the assumption of homoscedastic errors (27). With a single instrument, the estimate coincides with that from 2SLS. LIML is close to median unbiased for all but the weakest instruments (118), although it does not have any finite moments for any number of instruments (133). We also perform an analysis using full information maximum likelihood (FIML), which is similar to LIML, except that in LIML each equation is estimated separately, whereas in FIML all equations are estimated simultaneously.

d) Bayesian: We use a Bayesian method (Section 5.3) which is analogous to the 2SLS model for a normally distributed risk factor and exposure (140). For each individual i, we model the measured risk factor x_i as coming from a normal distribution for X_i with mean ξ_i and variance σ_x^2; similarly, measured outcome y_i comes from an independent normal distribution for Y_i with mean η_i and variance σ_y^2. The mean risk factor ξ_i is assumed to be a linear function of the instruments $g_{ik}, k = 1, \ldots, K$. The model is estimated in one stage, allowing propagation of uncertainty and feedback between the two regression stages. There is no assumption on the distribution of the causal parameter β_1 (141).

$$\begin{aligned} X_i &\sim \mathcal{N}(\xi_i, \sigma_x^2) \\ Y_i &\sim \mathcal{N}(\eta_i, \sigma_y^2) \\ \xi_i &= \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} \\ \eta_i &= \beta_0 + \beta_1 \xi_i \end{aligned} \tag{6.1}$$

e) Adjusted Bayesian: The above model (6.1) assumes that an individual's risk factor and outcome are uncorrelated within genetic subgroups. This is not true, since there will be a correlation between X and Y due to the true causal association and to confounding, as seen in Section 6.2. The correlation due to confounding is the cause of weak instrument bias in the 2SLS method (101; 206). In the Bayesian formulation, we introduce a new model which explicitly includes the correlation between risk factor and outcome by using a bivariate normal distribution for (X_i, Y_i) with correlation ρ. We replace the first two lines of (6.1) by

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right) \tag{6.2}$$

Equivalently, to avoid bivariate distributions, we use the properties of the bivariate normal distribution to model X_i by its univariate marginal distribution and Y_i by its conditional distribution given $X_i = x_i$.

$$\begin{aligned} X_i &\sim \mathcal{N}(\xi_i, \sigma_x^2) \\ Y_i|X_i = x_i &\sim \mathcal{N}(\eta_i + \frac{\sigma_y}{\sigma_x}\rho(x_i - \xi_i), (1-\rho^2)\sigma_y^2) \end{aligned} \tag{6.3}$$

6.3.2 Simulations for continuous outcomes

We take a simple model of confounded association. Risk factor x_i for individual i is a linear combination of three instruments g_{ik} for $k = 1, 2, 3$ which take values 0 or 1, normally distributed confounder u_i, and error ϵ_{xi} terms. Outcome y_i is a linear combination of x_i and u_i with normally distributed error ϵ_{yi}. The true causal effect of X on Y is represented by β_1. To simplify, the constant terms in the equations are set to be zero:

$$\begin{aligned} x_i &= \alpha_{11}\, g_{i1} + \alpha_{12}\, g_{i2} + \alpha_{13}\, g_{i3} + \alpha_2\, u_i + \epsilon_{xi} \\ y_i &= \beta_1\, x_i + \beta_2\, u_i + \epsilon_{yi} \\ u_i &\sim \mathcal{N}(0, \sigma_u^2); \epsilon_{xi} \sim \mathcal{N}(0, \sigma_x^2); \epsilon_{yi} \sim \mathcal{N}(0, \sigma_y^2) \text{ independently} \end{aligned} \tag{6.4}$$

Since each instrument is dichotomous, there are 8 possible IV combinations. We simulated 100 000 datasets from this model for each set of parameters with 200 individuals divided equally between the combinations. The instruments can be thought of as uncorrelated SNPs with dominant minor allele frequency 0.293. We considered four sets of parameter values covering a range of typical situations, with $\sigma_x^2 = \sigma_y^2 = \sigma_u^2 = 1$ throughout:

a) null causal effect, moderate positive confounding ($\beta_1 = 0, \alpha_2 = 1, \beta_2 = 2$);

b) null causal effect, strong positive confounding ($\beta_1 = 0, \alpha_2 = 1, \beta_2 = 4$);

c) positive causal effect, moderate positive confounding ($\beta_1 = 1, \alpha_2 = 1, \beta_2 = 2$);

d) positive causal effect, strong negative confounding ($\beta_1 = 1, \alpha_2 = 1, \beta_2 = -4$).

The instruments are taken to be of equal strength, $\alpha_{11} = \alpha_{12} = \alpha_{13} = \alpha_1$ with α_1 taking five values from 0.2 to 0.6, corresponding to mean $F_{3,196}$ statistic values between 2.0 and 10.1.

6.3.3 Implementation

For computational reasons, we only perform 10000 simulations in a Bayesian framework for each scenario. Results using classical methods for 10000 and 100 000 simulations are given to assess the validity of results based on only 10000 simulations. We firstly consider median bias rather than mean bias, as mean bias is not defined for the LIML estimator.

Results for 2SLS were obtained using the *sem* package in R (167) and for LIML using the *ivreg2* command in Stata (117). In the Bayesian analyses, we use vague prior distributions on all parameters: normal priors with mean zero, variance 10^2 for all regression parameters, uniform priors on $[0, 20]$ for standard deviations and a uniform prior on $[-1, 1]$ for the correlation ρ. We use Markov chain Monte Carlo (MCMC) methods in WinBUGS (207) with at least 5000 iterations, of which the first 500 are discarded as 'burn-in'. We assess convergence by examining the Monte Carlo error, re-running simulations which have failed to converge. We regard the mean of the posterior distribution as the 'estimate' of the parameter of interest and the standard deviation of the posterior distribution as the 'standard error (SE)'; the posterior mean gave better properties than the posterior median for the median bias. We used the 2.5th to the 97.5th percentile range as the '95% confidence interval' to estimate coverage.

Although "credible interval" is the more appropriate term with Bayesian estimates, the term "confidence interval" is here used to encompass both Bayesian and non-Bayesian interval estimates.

6.3.4 Results

Table 6.1 shows the median bias and coverage of a 95% confidence interval for the first 10000 simulations using the 2SLS, LIML and Bayesian methods and for all 100 000 simulations using 2SLS and LIML. . Results are not presented for the FIML method as for each dataset the FIML estimate typically differed from the LIML estimates only in the fourth decimal place, and the standard error in the third decimal place (the FIML standard error was consistently less than the LIML standard error). The Monte Carlo standard error (MCSE), representing the uncertainty due to the limited number of simulations, for 10 000 simulations is 0.003–0.008 for the median estimate (depending on the strength of the instrument) and 0.002 for the coverage.

We can see that all methods exhibit some bias. When the instruments are very weak ($\mathbb{E}(F) = 2.0, 3.3$), the 2SLS and Bayesian methods are severely biased in the direction of the observational association. When the instrument has a mean strength of $\mathbb{E}(F) = 10.1$, the 2SLS method has a substantial median bias of around 0.07 with moderate confounding

and 0.14 with strong confounding. In contrast, the LIML method shows minimal bias throughout for all but the weakest instruments ($\mathbb{E}(F) = 2.0$). The unadjusted Bayesian method has results similar to that for the 2SLS method with a null causal effect, and is biased in the same direction when there is a true causal effect.

The adjusted Bayesian method has an absolute median bias less than 0.03 in the eight simulations when $\mathbb{E}(F) = 7.3$ or 10.1. As the standard error of the median estimate due to the number of simulations is of the order of 0.005 to 0.015, these results are compatible with the adjusted Bayesian method being median unbiased for $\mathbb{E}(F) > 7$. When the instrument is very weak, the posterior distributions for β_1 have a long-tailed distribution which is often skew. With a single instrument, the confidence interval in the ratio method using Fieller's theorem (114) may include infinity (89). This corresponds in the Bayesian analysis with multiple instruments to a bimodal posterior distribution. In both the skewed and bimodal cases, neither the posterior median nor mean is a good summary of the distribution, and the corresponding median bias across simulations is not close to zero despite analyzing the data under the correct model.

The 2SLS method underestimates CIs throughout, with coverage consistently less than 95% and as small as 80% with the weakest instruments under strong confounding. LIML again underestimates coverage throughout, especially with weak instruments, though not as severely as 2SLS. The adjusted Bayesian method has correct coverage throughout, with coverage within 2 standard deviations (0.44%) of 95% for 18 of the 20 sets of parameter values. Both the 2SLS and LIML methods rely on asymptotic normality to perform inference on the causal effect. As the true distribution of the causal effect with a finite population is not normal, but in fact heavy-tailed, the asymptotic standard error is an underestimate of the true uncertainty in the causal effect, and so coverage is underestimated.

The unadjusted Bayesian method usually has good coverage with a null causal effect, but incorrectly estimated confidence intervals throughout when there is a true effect. This is because the error structure between X and Y is incorrectly specified. The true contour lines of the joint probability density function of X and Y within genetic subgroups (for people with the same mean risk factor and outcome) should be elliptical with major axis in the direction of the confounded association. However, by ignoring the correlation, circular contour lines are assumed. Figure 6.2 shows simulations for the mean risk factor and outcome of three genetic subgroups assuming positive (left), zero (centre) and negative (right panel) correlation between X and Y with a positive true causal effect to illustrate the within-subgroup density function. We see that when the correlation is positive, the variation in the gradient between the groups (the causal effect, estimates shown as grey

		First 10000 simulations				100 000 simulations	
α_1	Mean F	2SLS	Bayesian	Adjusted Bayesian	LIML	2SLS	LIML
a) Null causal effect ($\beta_1 = 0$), moderate positive confounding							
0.2	2.0	0.460 (0.838)	0.433 (0.944)	0.452 (0.951)	0.170 (0.898)	0.4585 (0.840)	0.1650 (0.900)
0.3	3.3	0.256 (0.881)	0.275 (0.950)	0.132 (0.957)	0.019 (0.926)	0.2536 (0.875)	0.0189 (0.922)
0.4	5.1	0.159 (0.900)	0.176 (0.949)	0.032 (0.955)	0.006 (0.936)	0.1521 (0.899)	0.0050 (0.935)
0.5	7.3	0.103 (0.917)	0.115 (0.951)	0.007 (0.954)	0.003 (0.944)	0.1006 (0.914)	0.0004 (0.942)
0.6	10.1	0.069 (0.924)	0.074 (0.948)	0.001 (0.951)	-0.001 (0.945)	0.0705 (0.926)	-0.0001 (0.948)
b) Null causal effect ($\beta_1 = 0$), strong positive confounding							
0.2	2.0	0.910 (0.798)	0.846 (0.938)	0.860 (0.946)	0.288 (0.886)	0.9204 (0.801)	0.3072 (0.885)
0.3	3.3	0.522 (0.848)	0.564 (0.940)	0.269 (0.948)	0.048 (0.910)	0.5061 (0.853)	0.0361 (0.915)
0.4	5.1	0.313 (0.888)	0.350 (0.948)	0.046 (0.953)	0.014 (0.932)	0.3049 (0.887)	0.0046 (0.930)
0.5	7.3	0.215 (0.905)	0.239 (0.945)	0.030 (0.950)	0.024 (0.936)	0.2011 (0.908)	0.0014 (0.940)
0.6	10.1	0.139 (0.920)	0.153 (0.950)	-0.009 (0.950)	-0.012 (0.946)	0.1419 (0.920)	0.0005 (0.946)
c) Positive causal effect ($\beta_1 = 1$), moderate positive confounding							
0.2	2.0	1.464 (0.832)	1.322 (0.997)	1.450 (0.950)	1.155 (0.895)	1.4593 (0.836)	1.1646 (0.897)
0.3	3.3	1.251 (0.869)	1.266 (0.997)	1.142 (0.952)	1.029 (0.917)	1.2546 (0.874)	1.0248 (0.921)
0.4	5.1	1.151 (0.897)	1.197 (0.999)	1.027 (0.953)	1.003 (0.933)	1.1532 (0.897)	1.0003 (0.932)
0.5	7.3	1.109 (0.913)	1.150 (0.998)	1.010 (0.953)	1.007 (0.940)	1.1015 (0.915)	1.0010 (0.943)
0.6	10.1	1.075 (0.925)	1.105 (0.998)	1.004 (0.950)	1.002 (0.946)	1.0721 (0.925)	1.0013 (0.947)
d) Positive causal effect ($\beta_1 = 1$), strong negative confounding							
0.2	2.0	0.072 (0.798)	0.079 (0.915)	0.131 (0.947)	0.672 (0.882)	0.0884 (0.802)	0.6998 (0.885)
0.3	3.3	0.496 (0.856)	0.579 (0.918)	0.756 (0.952)	0.982 (0.913)	0.4856 (0.852)	0.9449 (0.915)
0.4	5.1	0.696 (0.886)	0.822 (0.914)	0.956 (0.950)	0.995 (0.932)	0.6956 (0.887)	1.0020 (0.930)
0.5	7.3	0.814 (0.906)	0.936 (0.915)	1.019 (0.953)	1.026 (0.940)	0.8000 (0.907)	1.0001 (0.940)
0.6	10.1	0.855 (0.923)	0.942 (0.913)	0.995 (0.952)	0.996 (0.948)	0.8594 (0.919)	1.0011 (0.945)

Table 6.1: Simulations for continuous outcome – Median estimate of $\beta_1 = 0$ or 1 (coverage probability of 95% confidence interval) from 2SLS, LIML and Bayesian methods across 10000 and 100 000 (2SLS and LIML only) simulations for various scenarios and strengths of instrument

lines passing through the true mean of the middle subgroup) is less than when zero correlation is assumed, which in turn is less than when there is a true negative correlation. Hence when the confounded correlation within groups is in the same direction as the causal effect, ignoring this correlation will result in overly wide confidence intervals, and when the correlation is in the opposite direction, confidence intervals will be underestimated.

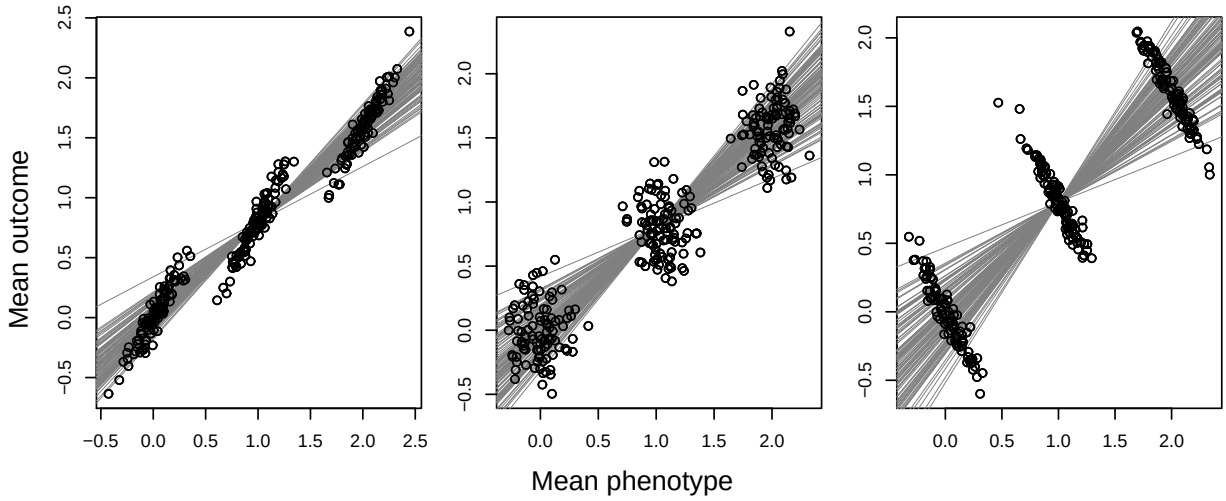


Figure 6.2: Simulated data illustrating joint distribution of mean phenotype and outcome in three genetic subgroups and causal estimate of association (grey lines) with positive between-group association and positive (left), null (centre) and negative (right panel) within-group correlation

6.3.5 Comparing mean and median bias

Table 6.2 explores the mean and median estimates of bias for the 2SLS and adjusted Bayesian approaches. Although LIML has no finite moments, the estimators for the 2SLS and Bayesian methods with three IVs have finite first moments, and so the mean estimate (and equally the mean bias) is a sensible quantity to consider for these estimators. The Bayesian method would even give an estimate with a finite mean even if there were no data, due to the prior distribution. The true distribution of the IV estimator is biased in the direction of the observational correlation between X and Y, and skewed in the opposite direction. This can be observed in the difference between the mean and median 2SLS estimates across simulations. The posterior distribution from the adjusted Bayesian method is also skewed in the same direction, as it reflects the true uncertainty of the sampling distribution of the IV estimate. As the instrument becomes stronger, there is

no clear pattern since this skew has become less pronounced. For the adjusted Bayesian method with $\mathbb{E}(F) = 10$, as already noted, the median bias is close to zero when the posterior mean is considered as a point estimate. We see here that the mean bias is close to zero when the posterior median is considered as a point estimate. The effect of considering the posterior median or the sample median biases the point estimate in the opposite way to considering the posterior mean or the sample mean. When the sample mean of the posterior medians or the sample median of the posterior means is considered, with moderately strong instruments the two effects seem to cancel each other out, leading to bias being close to zero.

For a Bayesian, bias is an odd concept as it requires reducing the posterior distribution to a single point value. As is seen in this example, depending on how bias is defined, different summaries of the posterior distribution will be more or less biased. A Bayesian would rather report the entire posterior distribution, as this represents their true belief about the parameter of interest. Coverage is a much more important property to a Bayesian, as this depends on the entire posterior distribution.

6.3.6 Different strength instruments

In response to concerns in Kleibergen and Zivot's paper (134), that the adjusted Bayesian method may perform badly when instruments of different strength were used, we perform simulations similar to those in Section 6.3.2, except with the ratio of the genetic association parameters ($\alpha_{1.}$) set at 1:3:5.

$$\begin{aligned} x_i &= 1\alpha_1\, g_{i1} + 3\alpha_1\, g_{i2} + 5\alpha_1\, g_{i3} + \alpha_2\, u_i + \epsilon_{xi} \\ y_i &= \beta_1\, x_i + \beta_2\, u_i + \epsilon_{yi} \\ u_i &\sim \mathcal{N}(0, \sigma_u^2); \epsilon_{xi} \sim \mathcal{N}(0, \sigma_x^2); \epsilon_{yi} \sim \mathcal{N}(0, \sigma_y^2) \text{ independently} \end{aligned} \tag{6.5}$$

Five values of α_1 were considered (0.1, 0.15, 0.2, 0.25 and 0.3) corresponding to mean F statistics between 4.0 and 27.5. All other parameters were taken to be the same as in the original simulation, and the same four scenarios are considered.

Results are presented in Table 6.3 for the 2SLS, Bayesian, adjusted Bayesian and LIML methods across 1000 simulations. Although the reduced number of simulations means that the MCSE for the median estimates are around 0.005–0.020 (depending on the strength of the instrument) and for the coverage are 0.007, the pattern of results is very similar to that in the equal strength instrument case considered previously. The adjusted Bayesian estimates are consistent with zero median bias for $\mathbb{E}(F) \geq 7.6$, and the coverage of the adjusted Bayesian method is close to the nominal level throughout. This is compared to

	Method:	2SLS		Posterior mean from adjusted Bayesian model (6.2)		Posterior median from adjusted Bayesian model (6.2)	
α_1	Mean F	Mean	Median	Mean	Median	Mean	Median
a) Null causal effect ($\beta_1 = 0$), moderate positive confounding							
0.2	2.0	0.4138	0.4598	0.4246	0.4520	0.4753	0.4826
0.3	3.2	0.1824	0.2555	0.0755	0.1324	0.1707	0.2067
0.4	5.1	0.1023	0.1589	-0.0410	0.0323	0.0573	0.1049
0.5	7.2	0.0596	0.1031	-0.0539	0.0072	0.0165	0.0609
0.6	10.2	0.0375	0.0695	-0.0409	0.0007	0.0056	0.0371
b) Null causal effect ($\beta_1 = 0$), strong positive confounding							
0.2	2.7	0.8336	0.9098	0.8189	0.8601	0.9223	0.9212
0.3	4.6	0.3726	0.5223	0.1581	0.2688	0.3484	0.4140
0.4	7.5	0.1901	0.3132	-0.1023	0.0463	0.0970	0.1965
0.5	11.2	0.1253	0.2154	-0.1046	0.0297	0.0371	0.1331
0.6	15.7	0.0722	0.1390	-0.0752	-0.0093	0.0146	0.0687
c) Positive causal effect ($\beta_1 = 1$), moderate positive confounding							
0.2	2.0	1.4111	1.4637	1.4245	1.4500	1.4732	1.4810
0.3	3.3	1.1920	1.2513	1.0864	1.1420	1.1814	1.2130
0.4	5.1	1.0914	1.1511	0.9458	1.0265	1.0455	1.1010
0.5	7.3	1.0607	1.1089	0.9456	1.0100	1.0169	1.0630
0.6	10.2	1.0425	1.0752	0.9666	1.0040	1.0122	1.0415
d) Positive causal effect ($\beta_1 = 1$), strong negative confounding							
0.2	2.0	0.1489	0.0718	0.1626	0.1311	0.0572	0.0650
0.3	3.3	0.6397	0.4957	0.8781	0.7558	0.6790	0.6059
0.4	4.9	0.8074	0.6962	1.1115	0.9555	0.9111	0.8091
0.5	7.4	0.8948	0.8138	1.1192	1.0190	0.9771	0.9068
0.6	10.2	0.9226	0.8549	1.0749	0.9945	0.9846	0.9177

Table 6.2: Simulations for continuous outcome – Mean and median estimates of $\beta_1 = 0$ or 1 for 2SLS, posterior mean and posterior median of adjusted Bayesian method across 10000 simulations for various scenarios and strengths of instrument

the 2SLS method, which still shows signs of non-zero median bias even with $\mathbb{E}(F) = 27.5$, and to each of the other methods, which display incorrect coverage for weak instruments with $\mathbb{E}(F) = 4.0$.

α_1	Mean F	2SLS	Bayesian	Adjusted Bayesian	LIML
a) Null causal effect ($\beta_1 = 0$), moderate positive confounding					
0.1	4.0	0.192 (0.878)	0.212 (0.944)	0.060 (0.951)	-0.015 (0.927)
0.15	7.6	0.098 (0.904)	0.108 (0.937)	-0.002 (0.941)	-0.012 (0.931)
0.2	12.8	0.053 (0.926)	0.057 (0.938)	-0.007 (0.942)	-0.009 (0.939)
0.25	19.4	0.042 (0.949)	0.042 (0.964)	0.003 (0.960)	0.004 (0.958)
0.3	27.5	0.017 (0.951)	0.018 (0.959)	-0.013 (0.960)	-0.010 (0.960)
b) Null causal effect ($\beta_1 = 0$), strong positive confounding					
0.1	4.0	0.474 (0.845)	0.520 (0.926)	0.198 (0.939)	0.099 (0.900)
0.15	7.6	0.205 (0.912)	0.214 (0.951)	-0.014 (0.949)	-0.022 (0.948)
0.2	12.8	0.114 (0.905)	0.123 (0.944)	0.020 (0.948)	0.014 (0.935)
0.25	19.4	0.049 (0.939)	0.051 (0.950)	-0.018 (0.953)	-0.016 (0.949)
0.3	27.5	0.036 (0.940)	0.037 (0.953)	-0.010 (0.955)	-0.009 (0.949)
c) Positive causal effect ($\beta_1 = 1$), moderate positive confounding					
0.1	4.0	1.188 (0.885)	1.206 (1.000)	1.034 (0.959)	0.991 (0.932)
0.15	7.6	1.068 (0.922)	1.086 (0.997)	0.970 (0.956)	0.966 (0.952)
0.2	12.8	1.059 (0.918)	1.072 (1.000)	1.007 (0.948)	1.007 (0.939)
0.25	19.4	1.040 (0.938)	1.050 (0.999)	1.004 (0.950)	1.005 (0.956)
0.3	27.5	1.022 (0.939)	1.034 (1.000)	0.993 (0.949)	0.994 (0.947)
d) Positive causal effect ($\beta_1 = 1$), strong negative confounding					
0.1	4.0	0.638 (0.876)	0.712 (0.923)	0.853 (0.949)	0.998 (0.923)
0.15	7.6	0.805 (0.922)	0.912 (0.928)	0.983 (0.951)	1.003 (0.946)
0.2	12.8	0.854 (0.917)	0.923 (0.907)	0.964 (0.937)	0.973 (0.936)
0.25	19.4	0.925 (0.932)	0.972 (0.916)	0.997 (0.947)	0.998 (0.945)
0.3	27.5	0.952 (0.946)	0.981 (0.921)	0.998 (0.957)	0.996 (0.953)

Table 6.3: Simulations for continuous outcome with unequal strength instruments – Median estimate of $\beta_1 = 0$ or 1 (coverage probability of 95% confidence interval) for 2SLS, LIML and Bayesian methods across 1000 simulations for various scenarios and strengths of instrument

We conclude from this limited simulation exercise that the results and conclusions of this section are likely to apply equally in situations where the instruments used have different strength, and where they have the same strength.

6.3.7 Summary

We conclude that modelling the correlation between the risk factor and outcome is necessary in a Bayesian model in cases where there is a true causal effect or in any model where the instrument is weak. Compared with a 2SLS approach, the adjusted Bayesian method gives an improvement in coverage properties, and a marked reduction in bias for all but the weakest instruments.

6.4 Binary outcomes and logistic models

We firstly recall the individual and population log odds ratios from Chapter 4. These are typically different quantities. We list IV methods to estimate causal effects with binary outcomes, showing how they are analogous to methods for continuous outcomes, and then present simulations with binary outcomes to investigate bias and coverage properties in these methods.

6.4.1 Collapsibility

A measure of association is collapsible over a variable if it is constant across the strata of the variable, and if this constant value equals the value obtained from the marginal analyses (110). In a logistic model, the odds ratio is non-collapsible, as it differs depending on the distribution of confounders (33). As defined in Chapter 4, the individual log odds ratio (ILOR) represents the difference in log-odds when the risk factor $X = x$ is increased by one to $x + 1$ conditional on all other covariates ($V = v$). This is a constant function of x and v in a logistic-linear model, and so the dependence on these variables is dropped. The population log odds ratio (PLOR) represents the difference in log-odds for an unit increase across the distribution of X marginal in all other variables (V):

$$\text{ILOR} = \log(\text{odds}(Y(x+1,v))) - \log(\text{odds}(Y(x,v))) \tag{6.6}$$

$$\text{PLOR} = \log(\text{odds}(Y(X+1,V))) - \log(\text{odds}(Y(X,V))) \tag{6.7}$$

where $\text{odds}(Y) = \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}$ and $Y(x,v)$ is $Y(x,v) = Y|(X = x, V = v)$ is the outcome random variable with phenotype level x and covariate level v. The probabilities in the definition of the PLOR are taken across the joint distribution of X and V.

In a logistic risk model linear in X and V, the ILOR can be estimated by logistic regression of Y on X and V. The PLOR cannot be estimated without knowledge of the distribution of X or V, even if V is not a confounder in the X-Y association. We calculate the PLOR here by numerical integration using the *adapt* package in R (193) as per equation 4.10. Additionally, we refer to the confounded 'observational' association, calculated by logistic regression of Y on X ignoring V. This will be biased compared to the ILOR due to confounding, with direction of bias depending on the Y-V and X-V associations.

6.4.2 Methods

a) Two-stage: The analogue of 2SLS with binary outcomes is a two-stage estimator where the second stage (X-Y regression) uses logistic regression. The standard error is taken from the logistic regression with no correction. This will be underestimated, as the uncertainty in the first stage regression is not acknowledged.

b) Ratio: Similarly, with a single instrument, a ratio estimator with binary outcomes can be calculated using logistic regression in the G-Y regression (54; 173). Again, this coincides with the two-stage estimator. However, such regression methods do not yield consistent estimators of the ILOR and have been called "forbidden regressions" (118; 129). This is because the non-linear model does not guarantee that the residuals from the second-stage regression are uncorrelated with the instruments. As we have seen in Chapter 4, this leads to population-based causal estimates which are marginal with respect to the covariates for Y.

c) Adjusted two-stage: The adjusted two-stage approach uses the estimated residuals from the first stage (G-X) regression in the second stage (X-Y) regression, as they are unbiased estimates of the covariates for X, some of which will be related to Y (94). Including these residuals in the second stage regression is an attempt to adjust for unmeasured covariates in estimating the ILOR. This is known as a control function approach (131). We note that this adjustment is not relevant in the linear case, as the first-stage fitted values and residuals are uncorrelated, meaning that the second-stage regression coefficient for X would not change if the orthogonal first-stage residuals were added to the regression model.

d) Maximum likelihood: By jointly modelling the risk factor and outcome distributions, a maximum likelihood estimate of the causal effect can be calculated. We model the risk factor as normally distributed in a linear regression on the number of genetic variants, and the outcome as a Bernoulli random variable in a logistic model on the mean risk factor.

$$\begin{aligned} x_i &\sim \mathcal{N}(\bar{X}_i, \sigma_x^2) \\ y_i &\sim \text{Bernoulli}(\pi_i) \\ \bar{x}_i &= \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 \bar{x}_i \end{aligned} \tag{6.8}$$

The joint likelihood ℓ is given by:

$$\ell = \prod_{i=1,\ldots,N} \left(\pi_i^{y_i} (1-\pi_i)^{1-y_i} \frac{1}{\sqrt{2\pi}\sigma_x} \{\exp(-\frac{1}{2\sigma}(x_i - \bar{x}_i)^2)\} \right) \tag{6.9}$$

As maximization is performed over the joint model, this is a full information maximum likelihood (FIML) approach. The *optim* command in R can be used to perform maximization of the log-likelihood. Although results using the FIML method were not considered in this dissertation, they were added to the paper produced from work in this chapter, which can be found in Appendix D.

e) Bayesian: In the Bayesian approach with binary outcomes (Section 5.5), we assume the probability of an event (π_i) for each individual i is associated with the mean risk factor (ξ_i) in a logistic model (140; 141). The outcome Y_i is modelled as a Bernoulli random variable:

$$\begin{aligned} X_i &\sim \mathcal{N}(\xi_i, \sigma_x^2) \\ Y_i &\sim \text{Bernoulli}(\pi_i) \\ \xi_i &= \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} \\ \eta_i = \text{logit}(\pi_i) &= \beta_0 + \beta_1 \xi_i \end{aligned} \tag{6.10}$$

f) Adjusted Bayesian: Similarly to the adjusted two-stage model, we can adjust for the underlying first stage residuals in a Bayesian model.

$$\begin{aligned} Y_i | X_i = x_i &\sim \text{Bernoulli}(\pi_i) \\ \eta_i = \text{logit}(\pi_i) &= \beta_0 + \beta_1 \xi_i + \delta(x_i - \xi_i) \end{aligned} \tag{6.11}$$

We note that the coefficient δ for the residual association is analogous to the correlation parameter ρ in the continuous model (6.3), both algebraically as a coefficient for the first-stage residuals, and conceptually as a way of adjusting for unmeasured covariates. As in the continuous case, uncertainty in both the G-X and G-Y associations feeds back into the model through the joint distribution of the variables.

6.4.3 Simulations for binary outcomes

In order to investigate the bias associated with different levels of confounding for different strengths of instrument, we consider a model of confounded association with three instruments. Data were simulated from model (6.12), a binary outcome analogue of model

(6.4):

$$\begin{aligned} x_i &= \alpha_{11}g_{1i} + \alpha_{12}g_{2i} + \alpha_{13}g_{3i} + \alpha_2 u_i + \epsilon_i \\ \eta_i = \text{logit}(\pi_i) &= \beta_0 + \beta_1 x_i + \beta_2 u_i \\ y_i &\sim \text{Bernoulli}(\pi_i) \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_x^2); u_i \sim \mathcal{N}(0,1) \text{ independently} \end{aligned} \tag{6.12}$$

The continuous outcome η_i is converted to a binary outcome by drawing Bernoulli random variables with probability of event $\pi_i = \text{expit}(\eta_i)$, where expit is the inverse of the logit function. The three instruments were thought of as uncorrelated SNPs in Hardy-Weinberg equilibrium with minor allele frequencies of $\frac{1}{3}$, $\frac{1}{3}$ and $\frac{1}{5}$, and the sample size was set to 2025 ($= 3^4 \times 5^2$). We set $\beta_0 = 0$, giving close to 50% prevalence of disease and took two sets of values of $(\alpha_{11}, \alpha_{12}, \alpha_{13}) = (0.5, 0.4, 0.6), (0.10, 0.08, 0.12)$ with $\sigma_x^2 = 1$ corresponding to strong and weak instrument scenarios, with mean F statistics of 100 and 5 respectively in the regression of phenotype on the instruments. Parameter values were chosen to correspond to a case-control study with approximately equal number of cases and controls, with large enough sample size to give reasonable precision of the causal effect in the simulation. The genetic association parameters corresponding to the strong instruments were chosen to examine the effect of non-collapsibility in the absence of weak instrument bias; the weak instrument parameters were chosen to correspond roughly to the parameters in a recent study of the causal effect of C-reactive protein on coronary heart disease (64).

We consider two values for β_1 of 0.4 and -0.8 (corresponding to odds ratio 1.49 and 0.45) and four values for β_2 of 0.0, 0.2, -0.6 and 1.0 corresponding to different directions and levels of confounding, with $\alpha_2 = 1$. We perform 100 000 simulations for each set of parameter values (results using Bayesian methods from 1000 simulations).

6.4.4 Results

Table 6.4 gives the median estimate and coverage for $\beta_1 = 0.4$ or -0.8 for the two-stage and Bayesian methods both adjusted and unadjusted for the first-stage residuals, together with the median observational estimate and population log odds ratio (PLOR). Monte Carlo standard error across 1000 simulations is approximately 0.002 with the strong instrument and 0.02 with the weak instrument. The individual log odds ratio (ILOR) is equal to β_1 throughout. Coverage for the PLOR is also given for the unadjusted two-stage and Bayesian methods. The MCSE based on 1000 simulations for the coverage is 0.007, and

for the median estimate is about 0.012 with the weak instrument and 0.002 with the strong instrument.

With the stronger instrument, the two-stage and Bayesian estimators are attenuated compared to the ILOR, even when there is no confounding ($\beta_2 = 0$). The median estimates from both methods approximate the PLOR throughout. The estimates from the adjusted approaches are much closer to the ILOR, but there is still some attenuation especially when the confounding is strong.

By comparing the results with the weak and strong instruments, we see that there is an effect of weak instrument bias in the two-stage methods depending on the direction of confounding. The weak instrument bias generally appears to be less in the adjusted Bayesian method than in the adjusted two-stage method, although it is difficult to make a firm conclusion because of Monte Carlo standard error due to the small number of simulations.

Although with the weak instrument, coverage is fairly close to the nominal level, this may be due to the lack of precision in the causal estimates rather than because the inference is good. With the strong instrument, the unadjusted estimators have poor coverage of β_1 but reasonable coverage of the PLOR. The adjusted estimators have coverage of the ILOR close to 95% throughout, except with strong confounding. Except for 1 of the 16 sets of parameter values, the adjusted Bayesian method has coverage within two standard deviations (1.4%) of 95%.

6.4.5 Simulations for semi-parametric estimators

Two other approaches to IV estimation are the Generalized Method of Moments (GMM) and Generalized Structural Mean Models (GSMM). GMM is designed as a more flexible form of 2SLS to deal with problems of heteroscedasticity of error distributions and non-linearity in the two-stage structural equations (126; 142). GSMM were designed in the context of randomized trials with incomplete compliance (145; 146). In the IV setting, the potential outcome $Y(x)$ is defined as the outcome which would have been observed if the risk factor X were set to x. A structural form is assumed for $Y(X) - Y(0)|X = x$ and the causal parameter is found using "G-estimation" (149; 150), using the independence of the 'exposure-free outcome' $Y(0)|X = x$ and the IV. Both of these methods are described as semi-parametric, with a parametric form assumed for the structural equations but no assumption on the error distribution. With a normal error distribution, the logistic GSMM is equivalent to an adjusted two-stage approach (153). With retrospective data, the risk factor for the cases can be omitted or down-weighted in a GSMM approach (208).

In order to learn more about the behaviour of these semi-parametric estimators, we repeat the simulation of Section 4.4.3 which uses Model (6.13) below. This is equivalent to Model (6.12), expect that there is only one IV.

$$\begin{aligned} x_i &= \alpha_1 g_i + \alpha_2 u_i + \epsilon_i \\ \text{logit}(\pi_{1i}) &= \beta_0 + \beta_1 x_i + \beta_2 u_i \\ y_{1i} &\sim \text{Binomial}(1, \pi_{1i}) \\ u_i &\sim \mathcal{N}(0,1), \epsilon_i \sim \mathcal{N}(0, \sigma_x^2) \text{ independently} \end{aligned} \tag{6.13}$$

For computational reasons, it was not practical to run the GMM and GSMM algorithms for 2 500 000 simulations as in Chapter 4. Hence, we changed some of the parameters from the simulation in Section 4.4.3 to give more precise estimation of the causal effect at the cost of making the generating model slightly less realistic. We set $\alpha_1 = 0.5, \alpha_2 = 1, \sigma_x^2 = 0.1^2$, corresponding to a slightly stronger instrument with a mean F statistic of 60 on a reduced sample size of 1000. We set $\beta_1 = -1$, corresponding to a greater number of events. We retained the same range of three values for β_1 of 0.4, -0.8 and 1.2 and seven values for β_2 of -1.0, -0.6, -0.2, 0, 0.2, 0.6, 1.0. The 100 000 simulations below of the GMM and GSMM methods took 17 CPU-days on a 2.2GHz processor.

We calculated the GMM and GSMM estimators "by hand" using the *optim* command in R for computational speed. Similar results were obtained using a number of user-written packages including the *ivpois* and *gmm* commands in Stata, and the *gmm* package in R. We calculated the GSMM estimator following the work of Vansteelandt and Goetghebeur (152) in two ways: firstly, with a logistic associational model of outcome on phenotype, instrument and the interaction term (as recommended by the original authors), and secondly omitting the interaction term in the associational model. We refer to these variations as GSMM-1 and GSMM-2.

The results in Table 6.5 show that the two-stage method again gives attenuated results compared to β_1, but similar to the PLOR. The adjusted two-stage and GSMM (especially GSMM-2) methods give similar results throughout, as has been theoretically shown when the distribution of X is normal (208). In this example, neither the adjusted two-stage nor the GSMM estimators are far from the ILOR. This is because the majority of the variation in X is due to U and not the error term, and so the residual in the adjusted two-stage method $R = X - \hat{X}$ is close to the true residual term in the second-stage regression. The GMM estimate is generally biased in the opposite direction to the direction of the bias of the observational estimate due to confounding. The bias of the GMM estimate is large, especially with large values of the true causal effect ($\beta_1 = 1.2$).

Additionally, we calculated the GMM and GSMM results for 1000 simulations of Model (6.12) from Section 6.4.3 above. We here use the *gmm* package in R to calculate GMM and GSMM estimates (170). Results, together with the PLOR and those from the adjusted two-stage method are given in Table 6.6. As before, Monte Carlo standard error across 1000 simulations is approximately 0.002 with the strong instrument and 0.02 with the weak instrument.

In Table 6.6, much of the variation in X is due to the error term ϵ_i, and the adjusted two-stage and GSMM estimators are attenuated compared to the ILOR. In the $\mathbb{E}(F) = 5$ case, the GSMM often failed to converge, leading to discrepancies between the adjusted two-stage and GSMM estimates.

We conclude that the median values of the GSMM estimates are close to the ILOR with no confounding, but attenuate especially when confounding is large. The median values of the GMM estimate were close to the ILOR with no confounding, but biased for the ILOR when confounding is present. Both estimators display problems of convergence, especially when the instrument is weak.

6.4.6 Summary

In conclusion, these simulations for binary outcomes show the effect of both non-collapsibility and weak instrument bias. Although all of the estimates give some bias in estimation of a causal effect, the results suggest that the two-stage and Bayesian methods estimate the PLOR, and that adjusting for the first-stage residuals in either approach gives better estimation of the ILOR. There is nothing to choose between the classical and Bayesian adjusted estimators with a strong instrument, but the adjusted Bayesian approach generally shows less bias due to weak instruments.

Having discussed estimation of the ILOR, we recall from Chapter 4 that adjusting for the first-stage residual in either a two-stage or Bayesian analysis only adjusts for the variation in the phenotype not explained by the IV. While estimation of the ILOR is a noble goal, when the covariates for the outcome are unknown, this is not possible. In general, it is not clear what effect is being estimated in an adjusted analysis. Hence, while the binary outcome adjusted methods are of theoretical interest, the conclusion from Chapter 4 that their use in practice should not be recommended still holds.

		Obs.	PLOR	1000 simulations				100 000 simulations	
				Two-stage	Bayesian	Adj. two-stage	Adj. Bayesian	Two-stage	Adj. two-stage
Strong instrument ($\mathbb{E}(F) = 100$)									
$\beta_1 = 0.4$	$\beta_2 = 0$	0.400	0.370	0.373 (0.939) [0.948]	0.375 (0.946) [0.949]	0.400 (0.944)	0.402 (0.939)	0.374 (0.945) [0.957]	0.401 (0.950)
	$\beta_2 = 0.2$	0.485	0.357	0.359 (0.939) [0.969]	0.364 (0.953) [0.969]	0.400 (0.954)	0.400 (0.955)	0.360 (0.935) [0.961]	0.400 (0.951)
	$\beta_2 = -0.6$	0.133	0.379	0.383 (0.932) [0.943]	0.386 (0.943) [0.948]	0.386 (0.934)	0.387 (0.939)	0.381 (0.941) [0.948]	0.383 (0.941)
	$\beta_2 = 1.0$	0.754	0.287	0.285 (0.747) [0.965]	0.287 (0.792) [0.968]	0.361 (0.935)	0.359 (0.942)	0.291 (0.759) [0.965]	0.369 (0.929)
$\beta_1 = -0.8$	$\beta_2 = 0$	-0.801	-0.635	-0.645 (0.595) [0.956]	-0.650 (0.704) [0.974]	-0.805 (0.943)	-0.808 (0.946)	-0.644 (0.591) [0.968]	-0.802 (0.951)
	$\beta_2 = 0.2$	-0.710	-0.659	-0.663 (0.696) [0.957]	-0.669 (0.790) [0.973]	-0.792 (0.940)	-0.794 (0.944)	-0.670 (0.703) [0.964]	-0.797 (0.949)
	$\beta_2 = -0.6$	-1.024	-0.548	-0.558 (0.158) [0.977]	-0.561 (0.269) [0.984]	-0.781 (0.943)	-0.781 (0.945)	-0.555 (0.169) [0.975]	-0.776 (0.941)
	$\beta_2 = 1.0$	-0.326	-0.690	-0.700 (0.761) [0.949]	-0.704 (0.826) [0.964]	-0.724 (0.818)	-0.730 (0.846)	-0.702 (0.794) [0.945]	-0.725 (0.854)
Weak instrument ($\mathbb{E}(F) = 5$)									
$\beta_1 = 0.4$	$\beta_2 = 0$	0.400	0.373	0.390 (0.958) [0.960]	0.415 (0.966) [0.966]	0.423 (0.955)	0.418 (0.962)	0.374 (0.957) [0.956]	0.402 (0.949)
	$\beta_2 = 0.2$	0.498	0.359	0.382 (0.952) [0.953]	0.415 (0.973) [0.968]	0.421 (0.937)	0.405 (0.947)	0.373 (0.961) [0.961]	0.414 (0.950)
	$\beta_2 = -0.6$	0.098	0.382	0.329 (0.944) [0.944]	0.373 (0.964) [0.965]	0.330 (0.944)	0.370 (0.959)	0.340 (0.943) [0.945]	0.342 (0.942)
	$\beta_2 = 1.0$	0.818	0.287	0.348 (0.970) [0.966]	0.380 (0.985) [0.974]	0.433 (0.946)	0.374 (0.961)	0.342 (0.967) [0.964]	0.434 (0.949)
$\beta_1 = -0.8$	$\beta_2 = 0$	-0.801	-0.639	-0.624 (0.947) [0.974]	-0.673 (0.979) [0.989]	-0.783 (0.949)	-0.782 (0.955)	-0.638 (0.951) [0.972]	-0.804 (0.951)
	$\beta_2 = 0.2$	-0.698	-0.665	-0.668 (0.944) [0.960]	-0.719 (0.979) [0.988]	-0.801 (0.941)	-0.805 (0.945)	-0.651 (0.949) [0.968]	-0.784 (0.949)
	$\beta_2 = -0.6$	-1.062	-0.549	-0.563 (0.932) [0.977]	-0.613 (0.974) [0.984]	-0.808 (0.955)	-0.776 (0.959)	-0.579 (0.946) [0.977]	-0.819 (0.918)
	$\beta_2 = 1.0$	-0.273	-0.690	-0.607 (0.918) [0.942]	-0.672 (0.964) [0.962]	-0.628 (0.922)	-0.676 (0.947)	-0.631 (0.914) [0.940]	-0.653 (0.949)

Table 6.4: Simulations with binary outcomes – Median estimate (coverage probability of 95% confidence interval for β_1) [coverage for population log odds ratio (PLOR) given in square brackets for unadjusted two-stage and Bayesian analyses] from IV analyses using two-stage, Bayesian and adjusted methods across 1000 and 100 000 simulations in strong and weak instrument scenarios

		$\beta_2 = -1.0$	$\beta_2 = -0.6$	$\beta_2 = -0.2$	$\beta_2 = 0$	$\beta_2 = 0.2$	$\beta_2 = 0.6$	$\beta_2 = 1.0$
$\beta_1 = 0.4$	PLOR	0.3708	0.3955	0.3955	0.3857	0.3708	0.3333	0.2945
	Two-stage method	0.3744	0.3981	0.3977	0.3886	0.3726	0.3358	0.2973
	Adjusted two-stage	0.4013	0.4016	0.4015	0.4018	0.3996	0.4003	0.4008
	GSMM-1	0.4018	0.4023	0.4021	0.4023	0.4005	0.3998	0.4017
	GSMM-2	0.4017	0.4019	0.4014	0.4018	0.3998	0.4003	0.4011
	GMM	0.3858	0.4117	0.4113	0.4013	0.3837	0.3444	0.3034
$\beta_1 = -0.8$	PLOR	-0.5366	-0.6098	-0.6873	-0.7233	-0.7538	-0.7903	-0.7903
	Two-stage method	-0.5285	-0.6008	-0.6793	-0.7164	-0.7509	-0.7938	-0.7930
	Adjusted two-stage	-0.8002	-0.8025	-0.8015	-0.8008	-0.8020	-0.8006	-0.7990
	GSMM-1	-0.8019	-0.8042	-0.8015	-0.8030	-0.8031	-0.8019	-0.7999
	GSMM-2	-0.8008	-0.8026	-0.8013	-0.8008	-0.8020	-0.8005	-0.7989
	GMM	-0.5633	-0.6510	-0.7506	-0.7996	-0.8458	-0.9042	-0.9035
$\beta_1 = 1.2$	PLOR	1.1600	1.0873	0.9791	0.9226	0.8678	0.7676	0.6821
	Two-stage method	1.1863	1.1127	1.0002	0.9412	0.8846	0.7811	0.6899
	Adjusted two-stage	1.1991	1.2020	1.2024	1.2020	1.2035	1.2045	1.1990
	GSMM-1	1.2026	1.2027	1.2049	1.2035	1.2044	1.2058	1.2001
	GSMM-2	1.2001	1.2019	1.2023	1.2019	1.2036	1.2046	1.1989
	GMM	1.8202	1.5983	1.3235	1.1997	1.0931	0.9185	0.7806

Table 6.5: Simulation for semi-parametric estimators — Population log odds ratio (PLOR) compared to two-stage, adjusted two-stage, generalized method of moments (GMM) and two generalized structural mean model (GSMM) methods from Model (6.13) of confounded association. Median estimates across 100 000 simulations

		PLOR	Adjusted two-stage	GMM	GSMM-1	GSMM-2
Strong instrument ($\mathbb{E}(F) = 100$)						
$\beta_1 = 0.4$	$\beta_2 = 0$	0.370	0.400	0.400 (0.952)	0.417 (0.926)	0.415 (0.937)
	$\beta_2 = 0.2$	0.357	0.400	0.385 (0.954)	0.413 (0.934)	0.407 (0.934)
	$\beta_2 = -0.6$	0.379	0.386	0.414 (0.960)	0.398 (0.944)	0.391 (0.947)
	$\beta_2 = 1.0$	0.287	0.361	0.297 (0.788)	0.369 (0.949)	0.365 (0.951)
$\beta_1 = -0.8$	$\beta_2 = 0$	-0.635	-0.805	-0.815 (0.963)	-0.792 (0.931)	-0.794 (0.930)
	$\beta_2 = 0.2$	-0.659	-0.792	-0.842 (0.981)	-0.786 (0.921)	-0.786 (0.908)
	$\beta_2 = -0.6$	-0.548	-0.781	-0.657 (0.745)	-0.763 (0.935)	-0.771 (0.945)
	$\beta_2 = 1.0$	-0.690	-0.724	-0.911 (0.988)	-0.721 (0.831)	-0.721 (0.824)
Weak instrument ($\mathbb{E}(F) = 5$)						
$\beta_1 = 0.4$	$\beta_2 = 0$	0.370	0.423	0.422 (0.997)	0.424 (0.979)	0.422 (0.985)
	$\beta_2 = 0.2$	0.357	0.421	0.416 (0.998)	0.368 (0.955)	0.383 (0.970)
	$\beta_2 = -0.6$	0.379	0.330	0.352 (0.989)	0.413 (0.950)	0.378 (0.956)
	$\beta_2 = 1.0$	0.287	0.433	0.367 (0.999)	0.322 (0.948)	0.347 (0.958)
$\beta_1 = -0.8$	$\beta_2 = 0$	-0.635	-0.783	-0.784 (0.960)	-0.640 (0.914)	-0.651 (0.916)
	$\beta_2 = 0.2$	-0.659	-0.801	-0.855 (0.960)	-0.679 (0.973)	-0.687 (0.961)
	$\beta_2 = -0.6$	-0.548	-0.808	-0.692 (0.945)	-0.599 (0.959)	-0.682 (0.977)
	$\beta_2 = 1.0$	-0.690	-0.628	-0.751 (0.945)	-0.620 (0.979)	-0.601 (0.963)

Table 6.6: Simulations with binary outcomes – Population log odds ratio (PLOR) and median estimate from adjusted two-stage method compared to median estimate (coverage probability of 95% confidence interval for β_1) from IV analyses of Model (6.12) using generalized method of moments (GMM) and two generalized structural mean model (GSMM) methods across 1000 simulations in strong and weak instrument scenarios

6.5 Discussion

In this chapter, we observed how weak instrument bias gives rise to biased causal estimates in instrumental variable analyses. We introduced a Bayesian method to explicitly model the correlation between risk factor and outcome, which reduced median bias from weak instruments to close to zero when the mean F statistic is around 10 in a simulation exercise. We saw how this adjustment is analogous to a control variable approach with binary outcomes, which targets a conditional odds ratio in a logistic model as opposed to unadjusted methods, which target the population odds ratio.

6.5.1 Comparison with previous work

This chapter builds on previous work, which has shown that the likelihood-based LIML method is median unbiased in the continuous outcome case. The adjusted Bayesian approach, also likelihood-based, is also median unbiased for moderately strong instruments, and does not suffer from the problems of underestimated confidence intervals given by LIML with weak instruments. Although the adjusted Bayesian estimate is not median unbiased for very weak instruments, this problem is at least partially due to the shape of the posterior distribution, which cannot always be summarized by a single value. We have seen how failure to take into account the correlation in a Bayesian approach leads to weak instrument bias and incorrect coverage.

In the binary case, the adjusted two-stage estimator has been shown to estimate a conditional odds ratio which more closely estimates the individual odds ratio than the unadjusted two-stage estimator, as seen in Chapter 4 (94). We build on this and the Bayesian estimator of Chapter 5 by introducing an adjusted Bayesian approach, which adjusts for the first-stage residuals in a Bayesian framework. With a strong instrument, the adjusted two-stage and adjusted Bayesian approach give similar answers. With a weak instrument, the adjusted Bayesian approach seems to be less biased than the adjusted two-stage approach.

Previous work on improving coverage for weak instruments have proposed methods based on the inversion of tests which are robust to weak instruments (92; 209; 210) or permutation tests (122). We provide an alternative method which simultaneously estimates the causal parameter of interest and provides a confidence interval, uses available statistical software, and appears to generalize more easily, for example to the case of multiple phenotypes.

6.5.2 Retrospective data

In Mendelian randomization, when retrospective data have been measured, it is usual to make inference on the G-X association using non-diseased individuals, such as a control population in a case-control setting (89). This makes the assumption that the distribution of X in the controls is similar to that of the general population, which is true for a rare disease (208) and is necessary to prevent bias of the causal estimate due to reverse causation and ascertainment of case-control status (33). In the two-stage method, fitted values for the diseased individuals can be estimated from the G-X model on the non-diseased individuals only. However, residual values cannot be used as there is no pre-event exposure measurement, so the adjusted two-stage approach is not possible. In a Bayesian MCMC setting, the adjusted approach is possible even with retrospective data. An exposure value can be imputed for diseased individuals from the distribution of X in the model fitted on the healthy individuals only. At each iteration in the MCMC procedure, a value of x_i is drawn from this distribution, which is used to form the residual $(x_i - \xi_i)$. Feedback from these imputed values to the parameters of genetic association (α_k) should be cut (211), as otherwise the imputation process will affect the parameters in the G-X association.

6.5.3 Comparison with semi-parametric methods

As all of the simulations considered in this chapter have used a correctly specified model, the advantages of the semi-parametric approach of the GMM and GSMM estimators are not apparent. What is clear however is that, in the simple setting considered, the GMM estimator suffers from bias. With a normally distributed phenotype, the GSMM estimate can be approximated by the adjusted two-stage estimate, and so suffers from the same problems of attenuation from the ILOR when the variation in X is not correlated with variation in Y.

Although the Bayesian method introduced is parametric, the posterior distribution enables hypothesis testing without the need for asymptotic approximation, giving accurate coverage even with weak instruments. This contrasts with asymptotic assumptions of normality for the causal estimate in each of the other methods, which can give incorrect coverage especially with weak instruments. Although the assumptions necessary to estimate the model are stronger, the assumptions for accurate inference are less strong. It is not generally the case that either semi-parametric or parametric models should be preferred in practice. In this chapter, and in this dissertation as a whole, we only consider correctly specified models. This means that the robustness advantages of semi-parametric

models will not be evident from the results shown. The main reason for excluding sensitivity analyses with misspecified models from this dissertation is the multitude of different simulation results which could be presented by adjusting different aspects of the data-generating model. While such sensitivity analyses would be interesting to consider, conclusions about the methods based on these simulations would be difficult to generalize and limited in relevance to the specific departures from the model considered.

6.5.4 Key points from chapter

- Explicitly modelling the correlation between phenotype and a continuous outcome within genetic subgroups is preferable in a Bayesian model, and results in an estimator which is less affected by weak instrument bias than a two-stage method.
- Adjusting for the first-stage residuals in a binary outcome analysis is analogous to adjusting for this correlation, and results in estimates closer to the individual log odds ratio, especially when the majority of variation in the phenotype is correlated with variation in the outcome.
- However, this adjustment gives an estimate which is marginal in some covariates but conditional in others, and so does not have an obvious interpretation. A logistic generalized structural mean model targets this same estimand.
- Uncertainty in the causal parameter is accurately represented in the Bayesian method by the shape of the posterior distribution, resulting in better inference and coverage properties compared with classical two-stage methods.

Acknowledgement

We thank John Thompson (Leicester) for helpful discussions that led to the proposal of Bayesian model (6.2).

Appendix: WinBUGS code

WinBUGS code for adjusted Bayesian model with continuous outcome

```
model {
 beta ~ dnorm(0, 0.000001)
```

```
 beta0 ~ dnorm(0, 0.000001)
 xtau <- pow(xsd, -2)
 xsd ~ dunif(0, 20)
 ytau <- pow(ysd, -2)
 ysd ~ dunif(0, 20)
 rho ~ dunif(-1, 1)
 tauy <- ytau/(1-pow(rho,2))
 # tauy is the precision of y conditional on x
 alpha0 ~ dnorm(0, 0.000001)
for(k in 1:K) {
  alpha[k] ~ dnorm(0, 0.000001)
 }
for (i in 1:N) {
  xi[i] <- alpha0 + inprod(alpha[1:K], g[i,1:K])
  x[i] ~ dnorm(xi[i], xtau)
  eta[i] <- beta0 + beta * xi[i]
  muy[i] <- eta[i] + sqrt(xtau/ytau)*rho*(x[i]-xi[i])
 # muy[i] is the mean of y[i] conditional on x[i]
  y[i] ~ dnorm(muy[i], tauy)
  } } }
```

WinBUGS code for adjusted Bayesian model with binary outcome and case-control data

```
model {
 beta ~ dnorm(0, 0.000001)
 beta0 ~ dnorm(0, 0.000001)
 xtau <- pow(xsd, -2)
 xsd ~ dunif(0, 20)
 gamma ~ dnorm(0, 0.000001)
 alpha0 ~ dnorm(0, 0.000001)
for(k in 1:K) {
  alpha[k] ~ dnorm(0, 0.000001)
  }
for (i in 1:N) {
  xi[i] <- alpha0 + inprod(alpha[1:K], g[i,1:K])
  x[i] ~ dnorm(xi[i], xtau)
```

```
  logit(pi[i]) <- beta0 + beta * xi[i] + gamma * xres[i]
  y[i] ~ dbern(pi[i])
  }
for (i in 1:C) {
  xres[i] <- cut(x[i]-xi[i])
  } # C = number of cases, which are placed first in the data file
for (i in (C+1):N) {
  xres[i] <- x[i]-xi[i]
  } } }
```

Chapter 7

Missing data methods with multiple instruments

7.1 Introduction

One difficulty with applied Mendelian randomization studies is that, although the IV estimate is consistent (and so asymptotically unbiased) for the causal association, its variance is typically much larger than the variance from a standard analysis (ie. regression of Y on X adjusted for known confounders) (40). This is because the variation in the phenotype explained by the instrumental variable is usually small (100; 104). To test some causal associations, sample sizes of several thousands are needed (47).

A possible solution is to use multiple IVs. Where there are several genetic variants which can be used as IVs and each explains independent variation in the phenotype, the IV estimate using all of the instruments will have lower variance than the IV estimate using a subset of the IVs (44; 212). However, a problem arising from including multiple IVs in an analysis is missing data (45). Sporadically missing genetic data typically arise due to difficulty in interpreting the output of genotyping platforms. If the output is not clear, a "missing" result is recorded. Hence, although efficiency will be gained from using multiple instruments, this may be offset in a complete-case analysis due to more participants with missing data being omitted.

Rather than omitting participants, we seek to use the structure of the genetic data, in particular the correlation between genetic markers known as linkage disequilibrium (LD), to impute missing data and include all participants in an analysis, acknowledging uncertainty in the imputation. In this chapter, we introduce four methods for imputing missing data under the missing at random (MAR) assumption (i.e. the pattern of missingness in the genotype data does not depend on the values of the missing genetic data but only on

data that are observed) (180). We use the Bayesian method introduced in Chapter 5, and discuss possible modifications if data are missing not at random (MNAR, i.e. missingness depends also on the unobserved missing values). We apply these methods in a simulation study and to real data from the British Women's Heart and Health Study on the association between C-reactive protein (CRP) and each of fibrinogen and coronary heart disease (CHD). The observational associations between CRP and fibrinogen, and CRP and CHD are both positive, but attenuate on adjustment for known confounders. It is thought that the true causal associations are null (64; 175).

While missing data methods have been proposed for longitudinal analysis of non-compliance in a randomized trial (213), these are limited to a single IV and a continuous outcome. Neither a general purpose method for imputing missing data in an IV analysis, nor specific methods for Mendelian randomization data, are known to exist.

7.2 Methods for incorporating missing data

We conduct our analyses in a Bayesian framework as this lends itself naturally to data imputation. We use the complete-case Bayesian methods introduced in Chapters 5 and 6, and introduce four methods for imputing genetic data under the MAR assumption which can be incorporated into the Bayesian model to include subjects with missing genetic data.

We assume throughout that all IVs are single nucleotide polymorphisms (SNPs) with two possible alleles. We code each SNP as 0, 1 or 2, representing the number of variant alleles. Individuals with 1 variant allele on a SNP are heterozygotes; otherwise they are homozygotes. A per-allele genetic model is presumed for each SNP; another model could be used if considered more appropriate.

Genetic data may be missing for several reasons: an individual may fail to provide a sample for analysis, consent may not be given for genetic testing, DNA extracted may be of insufficient quality or quantity for analysis, or the reading from a genetic platform may be difficult to interpret and hence a missing result may be recorded. In the first three cases, no genetic data would be available for the individual, but they could be included in the analysis. Although they would be informative about the distribution of the phenotype and outcome, they would not generally contribute greatly to the estimation of a causal effect. The focus of this work is on individuals who have missing data for only some SNPs, as these would contribute most to the estimation of the causal effect.

7.2.1 Bayesian model

With continuous outcomes, we use the adjusted Bayesian model (6.2) and with binary outcomes, the unadjusted Bayesian model (6.10). In all binary outcome analyses, we only make inference on the gene-phenotype association in individuals without prior history of disease (33).

In each case, the causal parameter of interest is β_1, the increase in mean outcome (or log-odds of outcome) per unit increase in the phenotype. We use vague prior distributions on all parameters: in our example these are normal priors with mean zero and variance 1000^2 for all regression parameters, uniform priors on $[0, 20]$ for standard deviations, and a uniform prior on $[-1, 1]$ for the correlation ρ. We employ Markov chain Monte Carlo (MCMC) sampling using WinBUGS (207) for all analyses, with at least 50000 iterations, of which the first 1000 are discarded as 'burn-in'. We assess convergence by running three parallel chains with different starting values, examining the Gelman-Rubin plots (214).

Missingness in either phenotype or outcome is easily dealt with by the model, as information on ξ and η is gained from all other individuals with data on phenotype and outcome. However, missingness in the IVs is less simple, as it is not clear what the underlying distribution of the genetic parameters is. We present four methods for addressing missing genetic data below.

7.2.2 Multiple imputations method

We first impute the genetic data multiple times using a genetic software package (we used Beagle (215; 216) in this chapter), and incorporate the imputations into the Bayesian model using the WinBUGS *dpick* function (207) to choose one of the imputed datasets at random in each MCMC iteration. Beagle imputes genetic data using a hidden Markov model and empirical Bayes methods under a MAR assumption. The *dpick* function gives a discrete uniform categorical random variable taking integer values such that feedback from the rest of the model to this random variable is not permitted (211), so that the imputed datasets are used equally often on average. We add to the Bayesian model:

$$\begin{aligned} m &\sim \text{Discrete Uniform}(1, M) \\ \xi_i &= \alpha_0 + \sum_{k=1}^{K} \alpha_k \, g_{ikm} \end{aligned} \tag{7.1}$$

where g_{ikm} is the number of variant alleles of SNP k for individual i in imputed dataset $m, m = 1, \ldots, M$. When $M = \infty$, this is equivalent to imputing from the posterior distribution of the genotypes given by the genetic software package without feedback.

This is a similar idea to classical multiple imputation, but implemented in a Bayesian setting. In the examples below, we use $M = 10$ imputations.

7.2.3 SNP imputation method

Instead of using the multiple imputations approach, we can use the posterior probabilities of genotypes given by the same software package for each SNP directly in the Bayesian model. The output from Beagle gives us posterior probabilities p_{ijk} that SNP k for individual i takes value j. We model the number of variant alleles of SNP k for individual i as a categorical random variable taking values in {0, 1, 2}. We add to the Bayesian model:

$$g_{ik} \sim \text{Categorical}(p_{i0k}, p_{i1k}, p_{i2k}) \tag{7.2}$$

A disadvantage of this method is that it does not account for known correlation between SNPs when imputing multiple SNPs in the same individual. Additionally, in both the multiple and SNP imputation methods, only the genetic data are used to impute missing values. As the phenotype and outcome data contain information about the missing genetic data values, they should also be used in the imputation model (217). However, if the genetic markers are highly correlated and the genetic data do not explain much variation in the phenotype, then we would not expect the bias caused by this omission to be large.

7.2.4 Multivariate latent variable method

In this method, we extended our Bayesian model to include the Bayesian model for imputation of correlated SNPs proposed by Lunn et al. (218). Genetic material in humans is arranged in two haplotypes, each consisting of combinations of alleles which are inherited together. We use latent vectors $\psi_{1i} = (\psi_{1i1}, \ldots, \psi_{1iK})$ and $\psi_{2i} = (\psi_{2i1}, \ldots, \psi_{2iK})$ to model each of the haplotypes for an individual i by a multivariate normal random variable with one component corresponding to each SNP. If ψ_{1ik} is positive, SNP k on the first haplotype (numbered arbitrarily) has a variant allele; otherwise not. Hence the number of variant alleles for SNP k is $I(\psi_{1ik} > 0) + I(\psi_{2ik} > 0)$, where $I(.)$ is an indicator function. We use the WinBUGS function *dgene.aux* to model the number of variant alleles (218). This function describes a discrete distribution on {0,1,2} taking two arguments. When both arguments are negative, *dgene.aux* is 0 with probability 1; when the arguments have opposite sign, *dgene.aux* is 1 with probability 1; when both are positive, *dgene.aux* is 2 with probability 1. The function is coded as a probability distribution rather than a deterministic function for technical reasons: missing genetic data values are required to be stochastic, rather than deterministic nodes. The latent variables are a convenient way

of modeling correlations in discrete distributions with analogy to the underlying biological structure of the problem.

$$\begin{aligned} \psi_{1i} &\sim \mathcal{N}_K(\mu, \Sigma) \\ \psi_{2i} &\sim \mathcal{N}_K(\mu, \Sigma) \\ g_{ik} &\sim dgene.aux(\psi_{1ik}, \psi_{2ik}) \end{aligned} \tag{7.3}$$

The parameters of the multivariate normal distribution are given vague priors. The prior for the mean (μ) is multivariate independent normal with mean $\mathbf{0}$ and diagonal variance-covariance matrix with 10 as each diagonal element. The prior for the variance-covariance matrix (Σ) is inverse Wishart, where the scale matrix in the Wishart distribution is diagonal with 10 as each diagonal element.

7.2.5 Haplotype imputation method

If the variation in the genetic data can be summarized by a small number of haplotypes, then instead of using an additive SNP-based model of genetic association, we can use an additive haplotype-based model. If individual i has haplotypes h_{1i} and h_{2i}, we have:

$$\xi_i = \gamma_{h_{1i}} + \gamma_{h_{2i}} \tag{7.4}$$

There is no need of a constant term γ_0, as each individual has exactly two haplotypes.

Often, when there is limited genetic variation, SNPs are chosen to tag haplotypes and there is a one-to-one correspondence between SNPs and haplotypes. In this case, a per-allele additive SNP-based model is equivalent to this additive haplotype model. When there is uncertainty in haplotype assignment due to missing data, we use the available SNPs to reduce the genetic variation in the data to a set of candidate haplotypes, and model each unknown haplotype value by a categorical random variable with probabilities for each haplotype estimated from the relative proportions of each of the possible haplotypes in the dataset. This method is illustrated for a specific dataset below.

A disadvantage of this method is that it is difficult to write a general model which could be used for arbitrary genetic data. A separate imputation model is needed for each genotypic pattern of observed and missing data in the study population. This method is not recommended when there is an uncertainty in haplotype assignment for individuals with complete data, as the model may lose identifiability.

7.2.6 Use of Beagle for genetic imputation

Out of the four methods given for incorporation of missing data, two of them are self-contained methods for imputation (the latent variable and haplotype imputation methods), whereas the other two use an external program to impute the genetic variables (the multiple imputations and SNP imputation methods). In the second case, the output from the genetic imputation method is incorporated into a Mendelian randomization model, while allowing for uncertainty in the output from the model. From a Bayesian point of view, we use the posterior output from the genetic imputation model as a prior for the genetic variables in the instrumental variable analysis. Where the genetic variables are imputed with no uncertainty, either due to complete genetic information or linkage between genetic variants, this prior is equivalent to inclusion of the variables in the model with no allowance for uncertainty, as in the methods of Chapters 5 and 6.

Several reviews on models and algorithms for imputation of genetic variables are available (for example, (219)). Although the focus of this dissertation is not the comparison of different imputation software, we provide some details on how the algorithm used in the Beagle program works, as this is the program used in this chapter.

The Beagle input comprises genetic markers and their respective positions. These positions are used to form a "localized haplotype-cluster model" based on the biological principle that markers which are physically closer are more likely to be correlated than those which are physically distant. In this chapter, we do not specify the distances between markers, and so the markers are considered to be equidistant. The localized haplotype-cluster model is a hidden Markov model (HMM) where the states are diplotype pairings. Phased haplotypes for each individual are drawn from this HMM conditional on the observed genotype data. The haplotypes drawn are used to construct a new localized haplotype-cluster model, and the procedure is repeated for 10 iterations. The most-likely diplotype for each individual is outputted, and the probabilities of missing genotypes are calculated from the model that is fitted at the final iteration (216). No other information than the measured and missing genotype values are used in the imputation.

7.3 Simulation study

We perform a simulation study to assess the performance of the four imputation methods.

7.3.1 Set-up

Three genetic variants (G_1, G_2, G_3) are used as IVs. The data for each individual $i = 1, \ldots, N$ are generated from the model:

$$\begin{aligned} x_i &= \alpha_1 g_{1i} + \alpha_2 g_{2i} + \alpha_3 g_{3i} + u_i + \epsilon_{xi} \quad (7.5) \\ y_i &= \beta_1 x_i - 2u_i + \epsilon_{yi} \\ u_i, \epsilon_{xi}, \epsilon_{yi} &\sim \mathcal{N}(0,1) \text{ independently} \quad (7.6) \end{aligned}$$

where U is a confounder, and ϵ_x and ϵ_y are independent error terms. Missing data are introduced by random draws R_k for SNP k for each individual, where G_k is observed if $R_k = 1$, and missing if $R_k = 0$. The true causal effect was $\beta_1 = 1$. Datasets of 1000 individuals were generated for a range of five realistic scenarios:

- Scenario 1 has $\mathbb{P}(R_1 = 1) = \mathbb{P}(R_2 = 1) = 1, \mathbb{P}(R_3 = 1) = 0.8$, so that only SNP 3 contains any missingness. SNPs 2 and 3 are taken to be in complete LD. Minor allele frequencies (MAF) are all 0.4.
- Scenario 2 has correlated SNPs tagging four haplotypes with frequencies 0.4, 0.3, 0.2 and 0.1. R_1, R_2 and R_3 are independent with $\mathbb{P}(R_j = 1|G_1, G_2, G_3) = 0.93$.
- Scenario 3 has the same missingness mechanism as Scenario 2 but SNPs are uncorrelated. MAFs are 0.4, 0.4 and 0.2.
- Scenario 4 has the same haplotypes as Scenario 2 but R_1, R_2 and R_3 are independent with $\mathbb{P}(R_j = 1|G_1, G_2, G_3) = 0.98$ if $G_j = 0$ or 2 (i.e. homozygous at SNP j), and $\mathbb{P}(R_j = 1|G_1, G_2, G_3) = 0.88$ if $G_j = 1$ (i.e. heterozygous).
- Scenario 5 has the same missingness mechanism as Scenario 4 but same uncorrelated SNPs as Scenario 3.

Parameters of genetic association $(\alpha_1, \alpha_2, \alpha_3)$ were chosen as in Table 7.1 to give an average F statistic of around 16–20 in Scenarios 2–5. The relation between the four haplotypes and three SNPs in Scenarios 2 and 4 is given in Table 7.2. In each scenario, the complete-case analysis contains on average around 20% fewer individuals than the complete-data analysis due to missingness. Scenarios 1–3 follow the MCAR assumption, while Scenarios 4 and 5 do not.

The simulation study was very computer-intensive. For each method, we performed 11 000 iterations of the MCMC algorithm to estimate the posterior distribution of the causal effect. The first 1000 iterations were discarded as "burn-in". Calculations were

performed on a multi-core computer with 2.20GHz central processing units (CPUs). For the complete-data, complete-case and haplotype imputation methods, analysis of each simulated dataset took 8-12 minutes. The multiple imputations method took 20-30 minutes, the latent variable method took 40-50 minutes, and the SNP imputation method took 100-120 minutes. Analyses of 1000 simulated datasets were performed for each scenario (100 for the SNP imputation method for computational reasons). Convergence was assessed by examination of the posterior variance of the causal effect parameter. Results for simulations with a high estimated posterior variance were discarded, the MCMC algorithm was re-run with different initial parameter values, and convergence was checked by examination of the trace plot and empirical posterior distribution. In total, the simulation study took over 1 CPU-year of processing time.

We regard the mean of the posterior distribution as the 'estimate' of the parameter of interest and the standard deviation of the posterior distribution as the 'standard error (SE)'. We used the 2.5th to the 97.5th percentile range as the '95% confidence interval' to estimate coverage.

	α_1	α_2	α_3	Expected F statistic
Scenario 1	0.5	0.6	0.8	55 [1]
Scenario 2	0.5	0.6	0.8	16
Scenario 3	0.2	0.3	0.4	20
Scenario 4	0.5	0.6	0.8	16
Scenario 5	0.2	0.3	0.4	20

Table 7.1: Parameters of genetic association used in simulation study and expected F statistic from the regression of X on G (on 3 and 996 degrees of freedom) with complete data

[1] G_2 and G_3 are collinear, so the relevant F statistic here is on 2 and 997 degrees of freedom

	G_1	G_2	G_3	Frequency
Haplotype 1	1	0	0	0.4
Haplotype 2	0	1	0	0.3
Haplotype 3	0	0	1	0.2
Haplotype 4	0	0	0	0.1

Table 7.2: Relation between haplotypes and SNPs in Scenarios 2 and 4 and frequency of haplotypes

	Analysis method	Estimate of β_1 (MCSE)	Sample relative efficiency (MCSE)	Coverage of 95% CI (MCSE)	Mean width of 95% CI (MCSE)	Per dataset relative efficiency (MCSE)
Scenario 1	Complete-data	1.012 (0.005)	1	0.961 (0.007)	0.628 (0.003)	1
	Complete-case	1.019 (0.006)	0.788 (0.024)	0.960 (0.007)	0.713 (0.004)	0.791 (0.004)
	Multiple imputations	1.013 (0.005)	1.000 (0.002)	0.961 (0.007)	0.627 (0.003)	1.003 (0.001)
	SNP imputation	0.998 (0.016)	0.989 (0.008)	0.96 (0.02)	0.622 (0.009)	1.016 (0.005)
	Latent variable	1.005 (0.005)	1.053 (0.004)	0.961 (0.007)	0.613 (0.003)	1.046 (0.001)
Scenario 2	Complete-data	1.028 (0.008)	1	0.953 (0.007)	1.039 (0.011)	1
	Complete-case	1.031 (0.009)	0.768 (0.021)	0.947 (0.007)	1.186 (0.013)	0.801 (0.007)
	Multiple imputations	1.024 (0.009)	0.818 (0.022)	0.945 (0.007)	1.150 (0.011)	0.848 (0.009)
	SNP imputation	1.032 (0.030)	0.941 (0.078)	0.94 (0.02)	1.134 (0.036)	0.904 (0.024)
	Latent variable	1.012 (0.008)	0.960 (0.027)	0.946 (0.007)	1.086 (0.009)	0.943 (0.014)
	Haplotype imputation	1.010 (0.009)	0.910 (0.021)	0.944 (0.007)	1.097 (0.011)	0.916 (0.007)
Scenario 3	Complete-data	1.012 (0.007)	1	0.949 (0.007)	0.891 (0.007)	1
	Complete-case	1.017 (0.008)	0.797 (0.023)	0.948 (0.007)	1.014 (0.009)	0.804 (0.006)
	Multiple imputations	1.012 (0.008)	0.904 (0.019)	0.949 (0.007)	0.939 (0.008)	0.919 (0.005)
	SNP imputation	0.998 (0.024)	0.904 (0.049)	0.92 (0.02)	0.913 (0.022)	0.976 (0.012)
	Latent variable	1.001 (0.007)	1.000 (0.033)	0.944 (0.007)	0.901 (0.006)	1.013 (0.010)
Scenario 4	Complete-data	1.006 (0.008)	1	0.948 (0.007)	1.006 (0.009)	1
	Complete-case	1.009 (0.009)	0.841 (0.022)	0.948 (0.007)	1.107 (0.010)	0.857 (0.006)
	Multiple imputations	1.008 (0.009)	0.848 (0.022)	0.947 (0.007)	1.107 (0.010)	0.856 (0.006)
	SNP imputation	1.000 (0.029)	0.963 (0.054)	0.98 (0.02)	1.057 (0.033)	0.942 (0.019)
	Latent variable	1.002 (0.008)	0.955 (0.022)	0.946 (0.007)	1.037 (0.008)	0.957 (0.006)
	Haplotype imputation	0.996 (0.008)	0.925 (0.018)	0.945 (0.007)	1.044 (0.010)	0.949 (0.005)
Scenario 5	Complete-data	1.013 (0.007)	1	0.939 (0.007)	0.888 (0.006)	1
	Complete-case	1.018 (0.008)	0.814 (0.021)	0.933 (0.007)	0.986 (0.008)	0.836 (0.005)
	Multiple imputations	1.015 (0.007)	0.972 (0.015)	0.943 (0.007)	0.916 (0.007)	0.949 (0.004)
	SNP imputation	0.971 (0.021)	0.972 (0.041)	0.94 (0.02)	0.864 (0.017)	0.977 (0.010)
	Latent variable	1.008 (0.007)	1.021 (0.013)	0.941 (0.007)	0.891 (0.006)	1.000 (0.003)

Table 7.3: Mean estimate of causal effect (β_1), sample relative efficiency, coverage of the 95% confidence interval (CI) for $\beta_1 = 1$, mean width of the 95% CI, and per dataset relative efficiency (Monte Carlo standard error (MCSE) in brackets) from simulation study for up to six analyses in five scenarios

7.3.2 Results

Table 7.3 gives results for each scenario and method. We note that the haplotype imputation analysis is possible in Scenario 1, but results would be as the complete-data analysis, because data would be imputed without uncertainty. The haplotype imputation is not attempted in Scenarios 3 and 5 as the SNPs are uncorrelated. All other models, including the latent variable model (which estimates a variance-covariance matrix with near zero correlation in Scenarios 3 and 5), have been applied in each scenario.

In addition to the mean causal effect estimate, we give the coverage and mean width of the 95% confidence interval (CI) and two estimates of relative efficiency. The sample relative efficiency is calculated as the ratio of variance of the 1000 estimates of β_1 from the method in question to the variance of the estimates of β_1 from the complete-data analysis. We also give the mean estimate of the relative efficiency from each dataset, referred to here as the per dataset relative efficiency. The relative efficiency from each dataset is calculated as the ratio of the variance of the estimate of β_1 for that dataset from the method in question to the variance of the estimate of β_1 for that dataset from the complete-data analysis. For each estimate, we give the Monte Carlo standard error (MCSE), which represents the uncertainty in the result due to the limited number of simulations performed (220). Consistent estimation of the relative efficiency in each dataset relies on consistent estimation of the standard error, and so the per dataset relative efficiency does not necessarily tend to the true relative efficiency for large numbers of simulations. However the per dataset relative efficiency has a lower MCSE than the estimate of the relative efficiency, and so is informative for efficiency in this simulation study.

The additional results in Table 7.3 help to inform us about the analysis methods. The MCSEs of the sample and per dataset relative efficiency indicate that sufficient simulations have been performed to estimate the relative efficiency with a reasonable level of precision. We firstly note that the estimate of β_1 in the complete-data analysis is slightly larger than 1 in each of the scenarios. This is due to weak instrument bias and is a result of the skew posterior distribution for β_1. Although the sample and per simulation relative efficiencies of the complete-case analysis are close to 0.80 (as expected) in the MAR Scenarios 1–3, they are greater than 0.80 in the MNAR Scenarios 4 and 5, with the difference for the per simulation relative efficiency greater than would have been expected by chance. In these scenarios, the majority of the data lost is in the heterozygote group. The minor homozygotes, who constitute the smallest group and exhibit the greatest mean difference in phenotype level from the overall mean phenotype, contribute disproportionately to the precision of the causal effect. Less missingness in the minor homozygotes means that the precision of the causal effect is not so greatly reduced compared to the complete-data

analysis as in the MAR scenarios. However, the coverage of the complete-case method in Scenario 5 suggests that CIs of the complete-case analysis are maybe narrow in this MNAR scenario, although further simulations would be required to reach a firm conclusion.

Taking each of the missing data methods in turn, we see that the multiple imputations method performs better, especially in terms of efficiency, when used with uncorrelated SNPs (Scenarios 3 and 5) than with correlated SNPs (Scenarios 2 and 4). In Scenario 4 with correlated SNPs and a MNAR model, its efficiency is no better than that of the complete-case analysis; it does however outperform the complete-case analysis in the other four scenarios. It may be that using more than 10 imputations in the multiple imputations method would give better performance. A limited sensitivity analysis found that increasing the number of iterations in the multiple imputations method did not appreciably change the results; a further large-scale simulation study would be required to verify this. In practice, it would seem prudent to use a larger number of imputations than 10 when computing resources allow. The SNP imputation method seems to be an improvement on the multiple imputations method, although the reduced number of simulations does limit this conclusion.

Although the latent variable method gives the most precise estimation of the causal effect, there are signs in the simulation study that the method may give underestimated standard errors. In Scenario 1, the mean width of the 95% CI is narrower for the latent variable method than for the complete-data method, and the relative efficiency is greater than 1. However, there does not seem to be a problem with the coverage, which is close to the nominal 95% level throughout. Reasons for this phenomenon are given in Section 7.3.3.

The haplotype imputation method is only implemented in Scenarios 2 and 4, but in these scenarios, it gives the good performance in terms of mean width of the 95% CI and per dataset relative efficiency. We recommend the haplotype imputation method where it can be used. Otherwise, we favour the multiple imputations method over the SNP imputation method because of its better mathematical properties (such as imputing multiple missing SNPs in an individual taking account of the correlation between SNPs), lack of inferiority in Scenarios 3 and 5 where the SNPs are not correlated, and additional evidence (1000 versus 100 simulated datasets) from the simulation study.

7.3.3 Apparent precision of the latent variable method

It is not clear why the latent variable method gives more precise estimates than the complete data analysis, but we here give three observations about the methods and two plausible reasons for this phenomenon.

Firstly, the latent variable method gives estimates which are smaller (closer to zero) than the complete-data method. In Table 7.3, the mean estimates of β_1 across simulations are lower for the latent variable method than for the complete case method in each of the five scenarios. In Scenario 1, in 807 of the 1000 datasets the latent variable estimate is smaller than that of the complete-case analysis. In the other scenarios, the respective totals are 545, 549, 524, and 520. If the probability of one method giving a smaller estimate were 0.5, we would expect 95% of the totals to lie between 469 and 531.

Secondly, the mean estimate of β_1 from the complete-data analysis is greater than 1 in all scenarios. While in any individual scenario (with the possible exception of Scenario 2) this could be explained as a chance finding, the consistency of the finding combined with the known non-normality of the posterior distribution suggest that this is not just the result of random variation. Therefore, if the latent variable method estimates are slightly attenuated, there may be no deviation in the mean bias from $\beta_1 = 1$. Although the coverage is not calculated using the posterior mean, it is plausible that a slight attenuation in the distribution of β_1 would not severely affect the coverage. Hence we search for reasons why the latent variable method gives attenuated results compared to the complete-data analysis.

Thirdly, in Scenario 1, about 3-5% of the missing data values are incorrectly imputed at any one iteration for any dataset by the latent variable method. This compares to a posterior probability of incorrect imputation from Beagle of $< 0.01\%$.

The term 'non-differential misclassification' refers to the incorrect classification of an observation into a certain category due to measurement error which is independent of the true value. It is well-known that such misclassification generally biases results towards the null (221). Although that the misclassification from the latent variable imputation may not be non-differential, this is a possible reason for the attenuation.

A second possible reason for the attenuation is the feedback in the Bayesian model. When a two-stage Bayesian model is fitted, uncertainty propagates throughout the model. It may be that uncertainty in the model of causal association is in some way traded off with uncertainty in the imputation model, and so the uncertainty in β_1 is underestimated at a cost of some incorrect imputation in the latent variable model.

Whatever the reason, as compared to the multiple imputations method which gives mean estimates of β_1 within 0.004 of the mean estimate from the complete-data analysis throughout, the latent variable method gives results close to $\beta_1 = 1$ in all scenarios. Paradoxically, the aim of the missing data analysis is not to "give the correct answer" (that is to estimate $\beta_1 = 1$), but to give the same inference as would have been obtained if no data were missing. In this case, the attenuation from the latent variable method

"cancels out" the skewness of the posterior distribution, giving additional precision at no apparent cost of bias or coverage. However, that does not mean that the latent variable method is superior; it may be possible to construct a scenario where the latent variable method is either severely biased or gives incorrect coverage for the same reasons that the method works suspiciously well in the scenarios presented.

7.4 British Women's Heart and Health Study

We illustrate our methods using data from the British Women's Heart and Health Study (BWHHS) to assess the impact of using multiple instruments and missing data on Mendelian randomization analyses. BWHHS is one of the constituent studies of the CRP CHD Genetics Collaboration (CCGC). We examine the causal effect of CRP on fibrinogen (continuous outcome) and on coronary heart disease (binary outcome) using three SNPs in the CRP gene region as instrumental variables: rs1205, rs1130864, rs1800947. These three SNPs tag four haplotypes (Table 7.4) which comprise over 99% of the variation in the CRP gene in European descent populations (82).

Haplotype	rs1205	rs1130864	rs1800947
1	C	T	G
2	C	C	G
3	T	C	G
4	T	C	C

Table 7.4: Haplotypes in the CRP gene region tagged by three SNPs used as instruments

BWHHS is a prospective cohort study of heart disease in British women between the ages of 60 and 79. We use cross-sectional baseline data on 3693 participants who have complete or partial data for CRP, fibrinogen and the three SNPs. There is missingness in 2.1% of participants for CRP, 2.4% for fibrinogen, 10.8% for rs1205, 1.9% for rs1130864, and 2.6% for rs1800947. Genotyping was undertaken by Kbioscience on two separate occasions for SNP rs1205, and then for SNPs rs1130864 and rs1800947. Table 7.5 shows the pattern of missingness of SNPs. Although it is unusual to see so much more missing data in one SNP than in another, this may be due to the individual characteristics of that SNP or region of the DNA. 3188 individuals (86% of the total) had data on all three SNPs; of these 12 (0.4%) individuals had a genotype which did not conform to the haplotype patterns of Table 7.4. CRP measurements were assessed using an immunonephelometric high-sensitivity assay supplied by Behring. Only CRP measurements

from non-diseased individuals were considered to rule out reverse causation. CHD was defined as non-fatal myocardial infarction (using World Health Organization criteria). We assessed CHD at baseline, comparing individuals with a definite previous myocardial infarction (6.9%) against all other individuals. CRP was log-transformed throughout. We found that a per-allele model of genetic association was appropriate for each of the SNPs. Each of the SNPs was in Hardy-Weinberg equilibrium. Only participants of European descent were included to ensure homogeneity of the population in question.

The Sargan overidentification test (158) gives $p = 0.72$ with fibrinogen and $p = 0.08$ with CHD. This indicates that there is no more heterogeneity between the causal estimates using different IVs than might be expected by chance. Failure of an overidentification test is taken as evidence that there is a violation of the IV assumptions (30). We also tested a range of six continuous and three binary coronary risk factors: body mass index, total cholesterol, systolic blood pressure, diastolic blood pressure, low density lipoprotein, triglycerides, history of diabetes (definite vs other), history of hormone replacement therapy (never vs current/ex) and use of hypertensive medicine (current vs never/ex). Out of 27 tests of association between the 3 SNPs and 9 risk factors, none gave $p < 0.05$. We conclude that the IVs appear to be valid instruments for the data in question.

rs1205	rs1130864	rs1800947	Participants
✓	✓	✓	3201
✓	✓	✗	32
✓	✗	✓	20
✗	✓	✓	373
✓	✗	✗	43
✗	✓	✗	17
✗	✗	✓	4
✗	✗	✗	3

Table 7.5: Patterns of missingness in three SNPs used as instruments

7.4.1 Complete-case analyses

We analyze the BWHHS data using each of the three SNPs measured as the sole IV, and with all of the SNPs included as IVs. We perform two sets of analyses: firstly including all participants with complete data on the IV in question, and secondly using the common set of 3188 participants with measured values for all three SNPs. The F statistic in mul-

tivariate regression of phenotype on all the instruments is 16.7, indicating little potential bias from weak instruments (161).

Continuous outcome: mean difference in fibrinogen			
IV	N	Participants with complete data on IV (sample size = N)	Participants with complete data on all IVs (sample size = 3188)
rs1205	3283	0.029 (0.399)	0.021 (0.488)
rs1130864	3609	-0.146 (0.340)	-0.266 (0.432)
rs1800947	3584	-0.217 (0.428)	-0.166 (0.409)
All three			-0.102 (0.274)
Binary outcome: log odds ratio of CHD			
rs1205	3283	1.04 (0.77)	1.06 (0.80)
rs1130864	3609	-0.50 (0.61)	-0.55 (0.75)
rs1800947	3584	1.24 (0.90)	1.16 (0.84)
All three			0.44 (0.55)

Table 7.6: Estimate (SE) from IV analysis of causal effect of unit increase in log(CRP) on fibrinogen (μmol/l) and coronary heart disease (CHD) (β_1) for various instrumental variables (IVs): complete-case analysis for participants (N) with complete data on SNP used as IV in analysis and for participants with complete data on all SNPs

Table 7.6 shows that, considering the data on participants with complete data for each of the SNPs, using all the SNPs as the IV gives the most precise estimator, with at least 20% reduction in SE compared to using any of the SNPs individually. However, a substantial proportion of the data has been discarded in the complete-case analyses. If we only use SNP rs1130864 as the IV, an additional 421 participants can be included in the analysis, resulting in about a 20% reduction in SE. Although this gain in precision is not uniform across all SNPs, with a slight loss of precision in the causal estimates using SNP rs1800947 as the IV despite a sample size increase of 396, this analysis motivates us to use methods for incorporating individuals with missing data.

7.4.2 Haplotype-based analysis

For the haplotype imputation method, we note that each of the SNPs available here tags one haplotype. This means that the haplotype assignment of an individual with complete genetic data that are consistent with the haplotypes 1-4 of Table 7.4 can be determined without uncertainty. Where there is missing data, we consider the possible haplotype assignments consistent with the four haplotypes of Table 7.4. For example, an individual measured as heterozygous in SNPs rs1205 and rs1800947 (CT and CG) with a missing

data value for SNP rs1130864 must have one copy of haplotype 4 and one copy of either haplotype 1 or 2. An individual measured as homozygous CC in SNP rs1205 and GG in rs1800947 with a missing data value for SNP rs1130864 has two haplotypes which must each be either 1 or 2. For each individual, we model the unknown haplotypes using categorical random variables. For example, the variables in these examples would each have a binomial distribution taking value 1 or 2 with probabilities corresponding to the relative proportions of the haplotypes in the population. To estimate the proportions of each haplotype, we assume independence of haplotypes within and between individuals and maximize the likelihood of a multinomial distribution with the correct likelihood contributions from individuals with complete and missing data. These probabilities are used to form the priors for the categorical variables in the Bayesian analysis. The 12 individuals with genotypes not conforming to the haplotype patterns of Table 7.4 (hereafter labeled as 'rogue') were omitted from the analysis.

7.4.3 Results under the MAR assumption

We applied each of the four methods described above. Each of the imputation methods gives similar answers, which differ somewhat from the complete-case analysis results in terms of point estimate (Table 7.7), especially in the binary case. The exception is the latent variable method, which reported poor convergence for the parameters in the multivariate latent variable distribution, even when the number of iterations was substantially increased. However, the distribution of the causal parameter seemed to have converged. The reduction in the standard error for all missing data methods compared to the complete case analysis is 8-12%, corresponding to a 17-29% increase in sample size (assuming that the precision of the causal estimate increases proportional to the sample size), slightly more than the true increase in sample size of 16%. The Monte Carlo standard error, which describes the uncertainty about the value of the causal estimate due to using MCMC, is approximately 0.002 for the continuous outcome and 0.01 for the binary outcome.

It is perhaps surprising to find a gain in precision greater than the gain in sample size. However, the increase in sample size within each of the genotypic subgroups, each containing all individuals with a particular genotype, is not uniform. In this case, the individuals with imputed data fall disproportionately into the smaller subgroups. This means that most of the smaller subgroups increase in size by more than 16%, giving rise to a greater than expected increase in precision. These results assume that the data is missing at random (MAR), meaning that the fact that a data value is missing gives no

information about the true value of the data point beyond that provided by the observed data.

	Continuous outcome: mean difference in fibrinogen		Binary outcome: log odds ratio of CHD	
Imputation method	Effect (SE)	95% CI	Log odds ratio (SE)	95% CI
Complete case analysis	-0.102 (0.274)	-0.699, 0.382	0.44 (0.55)	-0.57, 1.59
Multiple imputations	-0.088 (0.249)	-0.619, 0.358	0.22 (0.50)	-0.75, 1.25
SNP imputation	-0.075 (0.250)	-0.613, 0.369	0.22 (0.49)	-0.73, 1.22
Latent variable method [1]	-0.040 (0.241)	-0.552, 0.401	0.20 (0.48)	-0.72, 1.15
Haplotype imputation	-0.061 (0.250)	-0.590, 0.391	0.23 (0.51)	-0.75, 1.25

Table 7.7: Estimates of causal effect of unit increase in log(CRP) on fibrinogen (μmol/l) and coronary heart disease (CHD) (β_1) in complete-case analysis ($N = 3188$) and in entire study population ($N = 3693$) using different imputation methods for missing genetic data

[1]The latent variable results are presented with the caveat that the parameters in the multivariate normal distribution of the latent variables did not converge, although the causal parameter did seem to have converged.

7.4.4 Assessing the missingness assumption

One plausible way that the data may be MNAR is that heterozygotes are thought to be harder to determine than homozygotes on many high-throughput genotypic platforms, and so a missing value is more likely to be assigned to a heterozygote than to a homozygote (222). We therefore describe a test of the null hypothesis that a missing value is equally likely for heterozygotes and homozygotes. In the absence of knowledge of the true genetic data for all individuals, we use correlations between the SNPs in the observed haplotype to infer missing SNP values. If the whole cohort is assumed to have genotypes conforming to the four haplotypes of Table 7.4, then the true missing SNP values can sometimes be determined. Although there may truly be individuals with rogue genotypes, the appearance of such individuals in the data may be due to genotyping error, which occurs typically in about 1% of instances.

Assuming that all of the individuals with missing data conform to the four haplotypes of Table 7.4, we see that if an individual is homozygous TT in SNP rs1205, then the individual must be homozygous CC in SNP rs1130864. If an individual is homozygous CC or heterozygous in SNP rs1205, then the individual's genotype for SNP rs1130864 cannot be determined. Of the 331 individuals homozygous TT in rs1205, 326 are homozygous CC

in rs1130864, 2 are rogue and 3 (0.9%) are missing. Out of the 3033 individuals with CC or CT in rs1205, 60 (2.0%) are missing in rs1130864.

We see that 0.9% of individuals who are by assumption homozygous in SNP rs1130864 are missing, compared to 2.0% of individuals who may be heterozygous in this SNP. We apply similar logic to SNPs rs1205 and rs1800947 to construct Table 7.8 in such a way that an individual cannot be included as having missing data more than once. We fit a logistic selection model for missingness, assuming that the probability that the SNP k is missing (π_{jk}) depends on the SNP, and on whether an individual is definitely a homozygote ($j = 1$) or possibly a heterozygote ($j = 0$).

$$\text{logit}(\pi_{jk}) = \delta_j + \gamma_k \qquad (j = 0, 1;\ k = 1, 2, 3;\ \delta_0 = 0)$$

Such selection models are rightly criticized as being very sensitive to the specification of the model (223). The analysis is presented not because the given model is assumed to be correct or of interest, but purely to provide an informal assessment of whether the differences in missingness rates between different groups can be explained by chance alone. If the selection model does fit the data well with a negative value of δ, this may mean that the data are MNAR. However, a poorly fitting model does not necessarily mean that the data are MAR; it may simply mean that the selection model considered is the wrong model.

Missing SNP:	rs1205	rs1130864	rs1800947
Must be homozygous	30/331 (9.1%)	3/331 (0.9%)	11/1486 (0.7%)
Could be heterozygous	331/3043 (10.9%)	60/3033 (2.0%)	22/1815 (1.2%)

Table 7.8: Proportions of missingness for each SNP for individuals who are definitely homozygous in that SNP versus those whose true genetic data cannot be determined by reference to haplotypes in Table 7.4

In this model, $\hat{\delta}_1 = -0.325$ (SE 0.169, $p = 0.06$), providing weak evidence against the null hypothesis that an individual with a missing result for a SNP is as likely to be a heterozygote as a homozygote. This suggests that MAR may be violated. We note that as we are using observed data to test for patterns in the missing data where the true values of the missing data can be determined, any pattern observed in the missingness would not violate the MAR assumption. However, it strongly suggests that the pattern would also be present in individuals where the true values of the missing data cannot be determined, which would violate MAR. As there is both a biologically plausible reason for potential violation of the MAR assumption and weak evidence from a selection model, we proceed to

perform a sensitivity analysis. Even in the absence of evidence from the selection model, a sensitivity analysis would seem prudent.

7.4.5 Sensitivity to the MAR assumption

In the following analyses, we assess sensitivity of the results to departure from the MAR assumption.

The SNP imputation method (Section 7.2.3) is the simplest of the four missing data techniques to modify under the MNAR assumption. For example, if we believe that heterozygotes are more likely to have missing data than homozygotes, we can increase the probability of being a heterozygote p_{i1} for each individual i.

To assess sensitivity to the MAR assumption, we increased the probability of an individual with missing data being a heterozygote in the SNP imputation method. We logit-transformed the probability of being a heterozygote (p_{i1}) for each individual i, added a constant d (here referred to as the heterozygote-missingness parameter), and back-transformed to the probability scale. This ensured that when the genotype of an individual is known with high probability, there would be little change in the posterior probabilities, whereas when the genotype was uncertain, the probability of the individual being a heterozygote would increase; the probabilities of major and minor homozygotes would remain in the same ratio. We varied d from 0 to 2 in steps of 0.5, where $d = 0$ corresponds to the MAR assumption. For example, a probability of being a heterozygote of 0.2 increases to 0.65 when $d = 2$.

		Continuous outcome (mean difference in fibrinogen)		Binary outcome (log odds ratio of CHD)	
		Effect (SE)	95% CI	Log odds ratio (SE)	95% CI
MAR	$d = 0.0$	-0.075 (0.250)	-0.613, 0.369	0.22 (0.49)	-0.73, 1.22
MNAR	$d = 0.5$	-0.073 (0.249)	-0.608, 0.370	0.24 (0.48)	-0.68, 1.23
	$d = 1.0$	-0.075 (0.246)	-0.603, 0.366	0.29 (0.48)	-0.63, 1.26
	$d = 1.5$	-0.080 (0.245)	-0.605, 0.362	0.29 (0.49)	-0.62, 1.29
	$d = 2.0$	-0.078 (0.244)	-0.595, 0.360	0.28 (0.48)	-0.65, 1.26

Table 7.9: Sensitivity analysis on the heterozygote-missingness parameter in a MNAR model for estimates of causal effect of unit increase in log(CRP) on fibrinogen (μmol/l) and coronary heart disease (CHD) (β_1) using SNP imputation method

We see that the estimates are not particularly sensitive to departures from the MAR assumption (Table 7.9). Part of the reason for this may be that, for many individuals, geno-

type can be imputed with little uncertainty due to the LD between the genetic markers. There is a slight increase in precision for the continuous outcome as the heterozygote-missingness parameter increases, possibly due to decreased uncertainty in genotype assignment, and a slight increase in the association for the binary outcome.

7.5 Discussion

In this chapter, we have considered using multiple instruments in IV analyses. Using multiple instruments has the potential to reduce the variance of the causal estimates, but if there are sporadic missing data, this increase is offset by a decrease in sample size in a complete-case analysis. The missing data methods we have described can be used to include all participants and gain precision in the analysis under the assumption of missingness at random (MAR). Even though this assumption may not be fully valid, the results in our example were not sensitive to departures from this assumption. A further assumption of the imputation methods is that of Hardy–Weinberg equilibrium (HWE). However, violation of HWE is often an indication of a population substructure; if a SNP is not in HWE, then this may call into question its use as an IV for Mendelian randomization in the dataset.

Although the haplotype imputation model is the most natural of the methods, relying on only the independent inheritance of haplotypes in the study population, it is not necessarily applicable to all Mendelian randomization studies. A characteristic of the BWHHS dataset is that the SNPs can be summarized as a small number of haplotypes with certainty; haplotype imputation in this dataset is the preferred analysis.

Out of the three general purpose methods for missing data imputation, the latent variable method is the most interpretable in terms of the underlying biology. One concern may be that the impact of the distributional assumptions of the latent variables on the analysis is not clear. There is a danger of lack of convergence or poor mixing in complicated Bayesian models such as this, which resulted in a somewhat different estimate from the other methods in the BWHHS example with the continuous outcome, although less difference was observed with the binary outcome.

The SNP imputation and the multiple imputations methods are both easy to implement and based on the same idea. In the multiple imputations method, we rely on sampling from a discrete number of imputations rather than from the entire probability distribution, although the number of imputations could be increased if this were thought to be a problem. A drawback of the SNP imputation model is the assumption of prior independence of the SNPs in the imputation. One problem with these two methods is that the genetic

data are imputed without using the phenotype and outcome. Although we would expect some attenuation in the causal estimate due to the omission of the phenotype, the results seem fairly similar to those of the haplotype and latent variable models, both of which allow feedback from the phenotype and outcome in the imputation process. In this chapter we have used Beagle for genetic imputation; results were similar when other imputation programs such as fastPHASE (224) were used.

Our recommended preference, where possible, would be to use a haplotype imputation method. If this is not possible, due to uncertainty in haplotype ascertainment, we would suggest using the multiple imputations method, with the latent variable method as a sensitivity analysis for the effect of omitting the phenotype and outcome from the imputation model.

The WinBUGS code for the general purpose multiple imputation methods used is available online (225) and as an appendix to this chapter.

7.5.1 Key points from chapter

- Use of multiple instruments in Mendelian randomization leads to more precise estimates of causal association. Sporadic missing genetic data can offset this gain, but missing data methods can recover the full sample size.
- Out of the four proposed methods in this chapter, the haplotype imputation method is recommended where the genetic variation in the population can be summarized by a set of haplotypes, and the SNP imputation method otherwise with the latent variable method as a sensitivity analysis.

Appendix: WinBUGS code

Bayesian method incorporating correlation

```
model {
 alpha0 ~ dnorm(0, 0.000001) # priors for regression parameters
 beta   ~ dnorm(0, 0.000001)
 beta0  ~ dnorm(0, 0.000001)
 xtau  <- pow(xsd, -2) # priors for variance parameters
 xsd    ~ dunif(0, 20)
 ytau  <- pow(ysd, -2)
 ysd    ~ dunif(0, 20)
```

```
 tauy  <- ytau/(1-pow(rho,2)) # conditional precision given x[i]
 rho    ~ dunif(-1, 1) # prior for correlation
 for(k in 1:K) { # index across IVs
  alpha[k] ~ dnorm(0, 0.000001) # prior for IV effects
  }
 for (i in 1:N) { # index across individuals
  xi[i] <- alpha0 + inprod(alpha[1:K], gene[i, 1:K])
    # phenotype regression in additive model across IVs
  x[i] ~ dnorm(xi[i], xtau) # normal model of phenotype
  muy[i] <- eta[i] + sqrt(xtau/ytau)*rho*(x[i]-xi[i])
    # conditional mean given x[i]
  y[i] ~ dnorm(muy[i], tauy) # normal model of outcome
  eta[i] <- beta0 + beta * xi[i] # unconditional mean of outcome
  } # beta is causal parameter of interest
 }
```

Multiple imputations method

```
model {
 alpha0 ~ dnorm(0, 0.000001)
 beta   ~ dnorm(0, 0.000001)
 beta0  ~ dnorm(0, 0.000001)
 xsig   ~ dunif(0, 20)
 xtau  <- pow(xsig, -2)
 ysig   ~ dunif(0, 20)
 ytau  <- pow(ysig, -2)
 tauy  <- ytau/(1-pow(rho,2))
 rho    ~ dunif(-1, 1)
 r      ~ dpick(1,10) # r indexes imputations
 for (j in 1:K) {
  alpha[k] ~ dnorm(0, 0.000001)
  }
 for (i in 1:N) {
  xi[i] <- alpha0 + alpha[1]*gene[i, 1, r] + alpha[2]*gene[i, 2, r]
           + alpha[3]*gene[i, 3, r] # phenotype regression uses current imputation
  x[i] ~ dnorm(xi[i], xtau)
  muy[i] <- eta[i] + sqrt(xtau/ytau)*rho*(x[i]-xi[i])
  y[i] ~ dnorm(muy[i], tauy)
```

```
  eta[i] <- beta0 + beta * xi[i]
  } }
```

SNP imputation method

```
model {
 alpha0 ~ dnorm(0, 0.000001)
 beta   ~ dnorm(0, 0.000001)
 beta0  ~ dnorm(0, 0.000001)
 xsig   ~ dunif(0, 20)
 xtau  <- pow(xsig, -2)
 ysig   ~ dunif(0, 20)
 ytau  <- pow(ysig, -2)
 tauy  <- ytau/(1-pow(rho,2))
 rho    ~ dunif(-1, 1)
 for (k in 1:K) {
  alpha[k] ~ dnorm(0, 0.000001)
  }
 for (i in 1:N) {
  for (k in 1:K} {
   gene[i, k] ~ dcat(geneprobs[i, k, 1:3]
   } # geneprobs are posterior probabilities from genetic imputation
  xi[i] <- alpha0 + inprod(alpha[1:K], gene[i, 1:K])
  x[i] ~ dnorm(xi[i], xtau)
  muy[i] <- eta[i] + sqrt(xtau/ytau)*rho*(x[i]-xi[i])
  y[i] ~ dnorm(muy[i], tauy)
  eta[i] <- beta0 + beta * xi[i]
 } }
```

Multivariate latent variable model

```
model {
 mu[1:K] ~ dmnorm(mu0[1:K], Sigma0[1:K, 1:K])
 Sigma[1:K, 1:K] ~ dwish(Sigma1[1:K, 1:K], K)
  # priors for the haplotype distributions:
 alpha0 ~ dnorm(0, 0.000001)
 beta   ~ dnorm(0, 0.000001)
 beta0  ~ dnorm(0, 0.000001)
```

```
xsig   ~ dunif(0, 20)
xtau  <- pow(xsig, -2)
ysig   ~ dunif(0, 20)
ytau  <- pow(ysig, -2)
for (k in 1:K) {
 alpha[k] ~ dnorm(0, 0.000001)
 }
for (i in 1:N) {
 psi1[i, 1:K] ~ dmnorm(mu[1:K], Sigma[1:K, 1:K])
 psi2[i, 1:K] ~ dmnorm(mu[1:K], Sigma[1:K, 1:K])
 # psi1 and psi2 are drawn from the same multivariate distribution
 # and represent the two haplotypes
 for (k in 1:K) {
  gene[i, k] ~ dgene.aux(psi1[i, k], psi2[i,k])
  } # gene values when known are entered as data, when unknown as NA
    # missing data values are imputed from the multivariate haplotype model
 xi[i] <- alpha0 + inprod(alpha[1:K], gene[i, 1:K])
 x[i] ~ dnorm(xi[i], xtau)
 muy[i] <- eta[i] + sqrt(xtau/ytau)*rho*(x[i]-xi[i])
 y[i] ~ dnorm(muy[i], tauy)
 eta[i] <- beta0 + beta * xi[i]
} }
```

Chapter 8

Meta-analysis of Mendelian randomization studies of C-reactive protein and coronary heart disease

8.1 Introduction

In previous chapters, we have explored various statistical issues related to instrumental variable (IV) analysis and in particular to Mendelian randomization. Throughout, data has been used to illustrate findings. In this chapter, we perform a comprehensive analysis of these data to answer definitively the applied research question of interest: the causal effect of C-reactive protein (CRP) on coronary heart disease (CHD) based on the totality of the data available. We use data collected by the CRP CHD Genetics Collaboration (CCGC) (64; 82), which were introduced in Chapter 1. Although the methods in this chapter were developed for the CCGC, we believe that they cover a wide range of study designs and scenarios and will also be useful for meta-analysis of Mendelian randomization data in other contexts.

Typically, the variation in the phenotype explained by genetic variants is small, and so adequately powered Mendelian randomization studies usually require large sample sizes, demanding synthesis of evidence from multiple studies (40). Traditionally, meta-analysis is performed on summary data from already published sources (226). While meta-analysis of causal effects from Mendelian randomization in individual studies is possible, there are several reasons why this may not be a preferable option. Firstly, the distribution of the causal effect in a given study is not normal (89) (Chapter 3), and the estimate of standard error given by some methods underestimates the true level of uncertainty (Chapter 6), meaning that simple inverse variance weighting methods are not optimal. Moreover, some

study-specific estimates may have infinite variance. Secondly, the estimates in smaller studies may be biased due to weak instruments (2) (Chapter 3). Thirdly, there is a correlation between bias and precision, meaning that more biased studies are overweighted in a meta-analysis (212) (Chapter 3). Fourthly, not all studies may have data available on both the phenotype and outcome, meaning that a causal estimate cannot be estimated in these studies (71) (Chapter 5). Fifthly, the studies could be combined more efficiently by allowing inference on a joint model rather than limiting our attention to each study in turn (Chapter 5). Finally, individual participant data (IPD) enable overall assessment of the IV assumptions by the use of measured confounders.

We combat these problems by use of the Bayesian hierarchical model introduced in Chapters 5 and 6 (140). By making certain simplifying assumptions, which are fully detailed below, we demonstrate how a range of different designs of studies with binary outcomes can be analysed, and how these causal estimates can be combined in a hierarchical model. By exploiting correlation between single nucleotide polymorphisms (SNPs), and defining haplotype patterns in a way which allows for individuals with missing genetic data in certain SNPs to be included in the analysis, we show how studies measuring different genetic markers can be included in the same genetic association model. By pooling estimates of genetic association in a random-effects model from studies which have measured the same genetic variants, we strengthen the instrument and increase precision (Chapter 5). By using the random effects distribution as an implicit prior, we show how studies with no data on the phenotype or only providing tabular data, which have measured the same genetic variants as other studies in the collaboration, can be included in the analysis (Chapter 5). By including measured covariates, we can reduce weak instrument bias and improve efficiency in estimation (Chapter 3). By including both prevalent disease events (those reported at baseline) and incident events in prospective studies, we use all available data on disease outcomes.

The structure of this chapter is as follows: having discussed the genetic instruments available in each study (Section 8.2), the two-stage and Bayesian frameworks for analysis are recalled (Section 8.3). We show how these frameworks can be used to analyse a single study as a worked example (Section 8.4), then each study in the collaboration (Section 8.5), assessing the model assumptions by use of sensitivity analyses. Extensions are recalled which efficiently deal with issues of combining evidence across studies (Section 8.6), and then results are presented for the causal change in CHD due to CRP (Section 8.7). We conclude by discussing the interpretation and potential applications of these methods (Section 8.8).

8.2 The CRP CHD Genetics Collaboration

The CCGC is a collaboration of 47 epidemiological studies seeking to ascertain the causal role of C-reactive protein (CRP) in coronary heart disease (CHD) using a Mendelian randomization approach. In all analyses, we restrict attention to participants of European descent, excluding from analysis the four studies with no European descent participants. This is to ensure greater homogeneity of the study populations and to prevent violations of the IV assumptions due to population stratification (2). CRP is positively-skewed, and so we take log(CRP) as the phenotype. We use the term risk ratio as a generic term meaning hazard ratio, odds ratio or relative risk as appropriate.

8.2.1 Genetic data and choice of instrument

Genetic data measured in the collaboration were introduced in Chapter 1. We use g1, g2, g3, and g4 to represent the four SNPs (or proxies thereof) pre-specified for use as IVs in the protocol to the CCGC (82). Studies are divided into four patterns based on the SNPs available in that study: Pattern 4, where all four SNPs (or suitable proxies) are measured; Pattern 3, where all SNPs except g3 are measured; Pattern 2, where all SNPs except g4 are measured; and Pattern 1, where SNP g2 (and possibly other SNPs) is measured. The exception is study ISIS, which does not measure any of the pre-specified SNPs, where we use SNP rs2808628 as the single IV.

To find the most appropriate model of genetic association, we plot for each study the mean level of the phenotype log(CRP) by number of variant alleles against the number of alleles (Figure 8.1). In this chapter, we use the word "per allele" to refer to linearity of a model for different levels of a SNP, and "additive" to mean additivity across SNPs. If a per allele model is appropriate, we expect to see straight lines through the means of log(CRP) per number of variant alleles. If the per allele parameter is the same in each study, then we expect these lines to be parallel. These figures suggest visually that for each SNP an additive assumption with similar size effect across studies seems reasonable; a more principled analysis follows. Only individuals who have not suffered a prevalent event at time of blood draw, or who are not cases in a case-control study are included in analyses involving the phenotype, to minimize the possibility of reverse causation.

8.2.2 Linear versus factorial versus saturated genetic models

For each study, Table 8.1 gives the minor allele frequencies for each of the SNPs, the adjusted R^2 and F statistic for various models making different assumptions about the

data. The linear model (8.1) assumes an additive effect of each SNP per variant allele. The factorial model (8.2) is additive between SNPs and models each SNP as a three-level factor. The saturated model (8.3) takes each SNP as a linear covariate and includes all possible interactions between the SNPs. One coefficient is included for each genotype exhibited in the population. The phenotype x_i is expressed as a function of the number of variant alleles g_{ik} of each SNP k ($1 \leq k \leq K$) with residual term ϵ_i:

$$\text{Linear: } x_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} + \epsilon_i \tag{8.1}$$

$$\text{Factorial: } x_i = \alpha_0 + \sum_{k=1}^{K} \alpha_{k1} 1_{g_{ik}=1} + \sum_{k=1}^{K} \alpha_{k2} 1_{g_{ik}=2} + \epsilon_i \tag{8.2}$$

$$\text{Saturated: } x_i = \alpha_0 + \sum_{j_1=1}^{3} \dots \sum_{j_K=1}^{3} \alpha_{j_1 j_2 \dots j_K} + \epsilon_i \tag{8.3}$$

In Table 8.1, the factorial regression contains two terms for each SNP included in the model, except where there are no participants with a particular number of variant alleles for a SNP, when the number of terms reduces. The saturated regression contains one term for each complete genotype exhibited in the population. In principle, for example in Pattern 4 studies, $3^4 = 81$ genotypes are possible, though if the assumption that the data can be summarized by 5 haplotypes (Section 8.2.4) is true, only 10 genotypes should be exhibited. The adjusted R^2 statistic shows that the proportion of variation explained by the SNPs beyond chance remains similar in the studies under each model of association. The F statistics, given in each case for testing the model in question against the null model, for the linear model shows that the SNPs are associated with log(CRP) and so are potential instruments, with $p < 0.001$ in 26 out of the 33 studies (including all the studies with CRP measurements in over 1000 individuals). The p-values for analysis of variance (ANOVA) tests are displayed in Table 8.1 by formatting of the text for the linear model versus the null model, and the factorial and saturated models versus the linear model. Tests of the factorial and saturated models against the linear model give little evidence to favour either model except in studies with only one SNP. Out of the 30 studies measuring CRP and more than one SNP, evidence favouring the factorial model ($p < 0.05$) was found in one study (WHITE2, $p = 0.021$) and for the saturated model in no studies. This is not more than would be expected by chance.

Estimates of causal association from IV methods are biased in the direction of the observational confounded association when the association between the instrument and phenotype is not statistically strong (2). Generally, an F statistic of 10 or less is quoted

as a rule of thumb as to when weak instrument bias would be an issue (102). As the F statistic for several of the studies is below 10 and as there is little evidence for the factorial or saturated models, we use a linear model throughout as the most parsimonious model of the three which seems to be explaining a similar proportion of variation.

8.2.3 Common versus different per allele genetic parameter in each study

For each SNP, we fit a linear regression of log(CRP) on the number of variant alleles in each study where CRP was measured. Figure 8.2 gives the forest plots of these effects in all studies. The between-study heterogeneity, as measured by I^2 (227) is for g1, 58% (95% CI: 37–72%); for g2, 29% (95% CI: 0–54%); for g3, 14% (95% CI: 0– 51%); and for g4, 8% (95% CI: 0–41%). This indicates that there is considerable statistical heterogeneity between the study-specific estimates, although visual inspection of the forest plots suggests that there is a consistent direction of association with similar magnitude across studies.

8.2.4 Defining haplotypes

Rather than analyzing studies measuring different subsets of SNPs separately, we can use the SNPs to define haplotypes as listed in Table 8.2. For studies measuring all four pre-specified SNPs (Pattern 4), we use five candidate haplotypes, as defined in the study protocol paper (82) and as found in the data. For studies measuring three of the pre-specified SNPs g1, g2 and g4 (Pattern 3), we use four candidate haplotypes. For studies measuring three of the pre-specified SNPs g1, g2 and g3 (Pattern 2), we use four candidate haplotypes. For studies measuring two of the pre-specified SNPs g1 and g2, we use three candidate haplotypes. We note that restricting the possible haplotypes in this way does not allow any possibility of phase uncertainty. Over 99% of the European descent participants in the CCGC had a genotype corresponding to a pair of these haplotypes. Four studies (CAPS, HIFMECH, ISIS, WOSCOPS) which did not measure SNPs g1 and g2 are excluded from the haplotype-based analyses.

We see that haplotype 1 is tagged by alleles C in SNP rs1205 (g1) and T in SNP rs1138064 (g2). This means that, even with missing data on SNPs rs1800947 (g3) and rs3093077 (g4), due to correlation between SNPs (LD), this haplotype can be uniquely determined. Haplotypes 4 and 5 differ only in SNP rs1800947 (g3). We categorize an individual having haplotype 4 or 5 with a missing value for g3 as having haplotype 7. This means that haplotype 7 will be an amalgamated category, consisting of a combination of haplotypes 4 and 5. Similarly, haplotype 6 will consist of a combination of haplotypes

2 and 3. Hence, in addition to the five candidate haplotypes observable in the data and pre-specified in the protocol paper, we used two haplotype categories corresponding to haplotypes which could not be determined between two candidate haplotypes due to missing or unmeasured genetic data.

We divide the haplotypes into three groups as follows. Group I contains haplotype 1, and is defined by a C allele in SNP rs1205 (g1) and a T allele in SNP rs1130864 (g2). Group II contains haplotypes 2 and 3 plus category 6, and is defined by a C allele in SNP rs1205 (g1) and a C allele in SNP rs1130864 (g2). Group III contains haplotypes 4 and 5 plus category 7, and is defined by a T allele in SNP rs1205 (g1) and a C allele in SNP rs1130864 (g2). A summary of the SNPs corresponding to the haplotypes and groups, including question marks for categories 6 and 7 where data are missing, is given in Table 8.3. Although studies may have different proportions of haplotypes 6 and 7 due to different proportions of missing data, if the participants in different studies come from comparable populations, then the proportions of haplotypes in groups I, II and III should be the same across all studies. Figure 8.3 and Table 8.4 show the frequency of each haplotype and group of haplotypes within each study. Group I haplotypes are coloured green, group II blue and group III red. The similarity in proportions of haplotypes across studies supports our claims of homogeneity of European descent populations, use of proxy SNPs (in complete LD) and determination of haplotypes.

8.2.5 Equivalence of SNP and haplotype models

If the assumption that the genetic variation in a population can be summarized by the five haplotypes of Table 8.3, then the linear SNP-based model (8.1) is equivalent to a linear haplotype-based model, where there is one coefficient (γ_k) per haplotype. For an individual with haplotypes h_1 and h_2, we have

$$x_i = \gamma_{h_1} + \gamma_{h_2} + \epsilon_i \tag{8.4}$$

where ϵ_i is an error term as before. This is because each haplotype is identified by the presence of a variant allele in one particular SNP (except haplotype 2, which is identified by no variant alleles in any of the SNPs). Such a combination of SNPs is known as a tagging set of SNPs. This is not a coincidence: the pre-specified SNPs were chosen precisely because they tag the five main haplotypes in European descent populations, so that no redundant genetic information need be measured.

The linear SNP-based model (8.1) and linear haplotype-based model (8.4) are trivially the same under this restriction, as there is a linear transformation (reparameterization)

taking the coefficients from one model to the other. Hence validity of the linear SNP-based model implies validity of the linear haplotype-based model.

8.2.6 Phenotype and outcome data

For each prospective (cohort) study in the collaboration, we show the quantile plot of the distribution of log(CRP) against quantiles of the standard normal distribution (Figure 8.4) and the piecewise constant estimate of hazard function (number of CHD events per participant-year) for each year of follow-up (Figure 8.5). Apart from for low levels of CRP, where assays are not sensitive enough to determine between small values, the distribution of log(CRP) can be approximated by a normal distribution. In most of the studies, the hazard function appears to be a smooth function of time. In later sections, we will investigate the sensitivity of regression of the outcome in cohort studies on parametric assumptions, and on ignoring variable follow-up. Although there are anomalous results in some of the studies (such as BRHS and CCHS), it seems that these assumptions may not severely misrepresent the data.

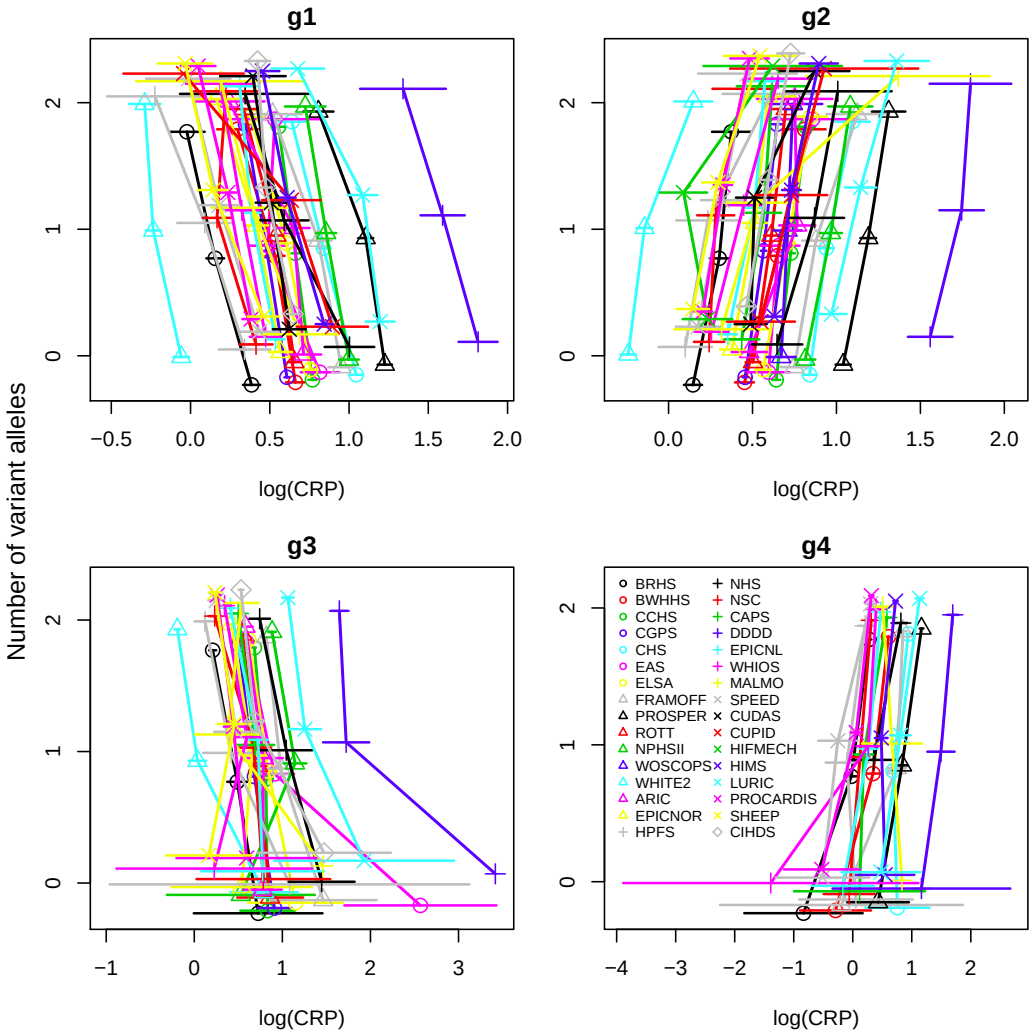


Figure 8.1: Studies with CRP measured – mean level of log(CRP) with 95% confidence interval in all non-diseased participants (control participants in case-control studies) with different numbers of variant alleles in each SNP. (Studies are separated vertically for visual purposes; some points omitted due to ≤ 1 participant with that number of variant alleles)

	Study	Minor allele frequencies g1	g2	g3	g4 [2]	Adjusted R^2 Linear	Factorial	Saturated	F statistic (degrees of freedom) [1] Linear	Factorial	Saturated
Pattern 4	BRHS	0.330	0.302	0.065	0.051	0.021	0.021	0.021	***19.73*** (4, 3511)	10.19 (8, 3507)	7.84 (11, 3504)
	DDDD	0.349	0.295	0.074	0.054	0.016	0.012	0.015	**3.45** (4, 609)	1.95 (8, 605)	1.91 (10, 603)
	EPICNL	0.324	0.322	0.063	0.056	0.019	0.019	0.018	***16.5*** (4, 3198)	8.75 (8, 3194)	6.37 (11, 3191)
	FRAMOFF	0.328	0.306	0.056	0.069	0.016	0.017	0.021	***7.2*** (4, 1474)	4.22 (8, 1470)	4.10 (10, 1468)
	HPFS	0.319	0.307	0.070	0.065	0.029	0.025	0.022	**3.98** (4, 398)	2.30 (8, 394)	1.76 (12, 390)
	LURIC	0.341	0.321	0.067	0.058	0.019	0.021	0.017	***8.73*** (4, 1594)	5.19 (8, 1590)	3.48 (11, 1587)
	MALMO	0.333	0.323	0.070	0.043	0.037	0.031	0.059	2.31 (4, 134)	1.55 (8, 130)	1.96 (9, 129)
	NHS	0.325	0.295	0.061	0.059	0.045	0.039	0.056	***5.53*** (4, 382)	3.26 (7, 379)	3.08 (11, 375)
	NSC	0.330	0.333	0.086	0.055	0.020	0.022	0.027	***6.04*** (4, 964)	3.66 (8, 960)	3.22 (12, 956)
	PROCARDIS	0.330	0.303	0.063	0.057	0.016	0.015	0.017	***14.13*** (4, 3297)	7.28 (8, 3293)	6.85 (10, 3291)
	SPEED	0.355	0.303	0.045	0.055	0.037	0.032	0.043	***6.43*** (4, 559)	3.69 (7, 556)	3.54 (10, 553)
	WHIOS	0.328	0.313	0.062	0.065	0.011	0.011	0.009	***5.73*** (4, 1717)	3.33 (8, 1713)	2.49 (11, 1710)
	FHSGRACE	0.320	0.307	0.069	0.057						
	GISSI	0.321	0.297	0.089	0.055						
	HVHS	0.328	0.309	0.058	0.062						
	UCP	0.335	0.294	0.062	0.065	No CRP data			No CRP data		
	AGES	0.312	0.317	0.050	0.077						
	HEALTHABC	0.334	0.316	0.064	0.064						
	MONICAKORA	0.332	0.315	0.067	0.066						
	PENNCATH	0.337	0.297	0.062	0.057						
Pattern 3	ARIC	0.349	0.292	-	0.068	0.021	0.025	0.024	***7.12*** (3, 855)	4.60 (6, 852)	4.55 (6, 852)
	CCHS	0.343	0.311	-	0.051	0.010	0.010	0.009	***32.09*** (3, 9499)	16.12 (6, 9496)	13.86 (7, 9495)
	CGPS	0.337	0.314	-	0.049	0.012	0.012	0.012	***126.47*** (3, 30487)	63.53 (6, 30484)	54.39 (7, 30483)
	CIHDS	0.336	0.320	-	0.048	0.015	0.016	0.015	***23.98*** (3, 4411)	13.19 (6, 4408)	12.15 (6, 4408)
	EAS	0.329	0.288	-	0.068	0.022	0.027	0.023	***5.88*** (3, 640)	4.00 (6, 637)	3.54 (6, 637)
	ELSA	0.327	0.272	-	0.087	0.016	0.016	0.016	***25.49*** (3, 4500)	12.96 (6, 4497)	11.23 (7, 4496)
	EPICNOR	0.324	0.313	-	0.073	0.012	0.011	0.013	***9.46*** (3, 2122)	5.02 (6, 2119)	4.88 (7, 2118)
	NPHSII	0.339	0.292	-	0.048	0.014	0.015	0.014	***11.58*** (3, 2154)	6.68 (6, 2151)	5.45 (7, 2150)
	ROTT	0.326	0.293	-	0.088	0.011	0.010	0.010	***17.37*** (3, 4520)	9.00 (6, 4517)	7.78 (7, 4516)
	SHEEP	0.338	0.309	-	0.052	0.026	0.024	0.026	***10.47*** (3, 1079)	5.52 (6, 1076)	5.84 (6, 1076)
	WHITE2	0.333	0.316	-	0.035	0.015	0.016	0.015	***24.71*** (3, 4796)	*14.04* (6, 4793)	11.2 (7, 4792)
	CHAOS	0.330	0.310	-	0.061	No CRP data			No CRP data		
Pattern 2	BWHHS	0.325	0.301	0.065	-	0.015	0.015	0.016	***16.34*** (3, 2966)	8.59 (6, 2963)	7.77 (7, 2962)
	CHS	0.339	0.304	0.069	-	0.018	0.018	0.018	***26.45*** (3, 4047)	13.67 (6, 4044)	11.53 (7, 4043)
	HIMS	0.334	0.307	0.060	-	0.019	0.018	0.018	***20.45*** (3, 3073)	10.56 (6, 3070)	10.41 (6, 3070)
	PROSPER	0.330	0.297	0.060	-	0.017	0.018	0.017	***29.1*** (3, 4872)	15.53 (6, 4869)	12.8 (7, 4868)
	INTERHEART	0.350	0.314	0.069	-	No CRP data			No CRP data		
Pattern 1	CAPS	-	0.329	0.076	0.051	0.019	0.017	0.021	***5.86*** (3, 753)	3.63 (5, 751)	3.65 (6, 750)
	CUDAS	0.349	0.293	-	-	0.006	0.009	0.006	*4.15* (2, 974)	3.16 (4, 972)	2.98 (3, 973)
	CUPID	0.342	0.298	-	-	0.061	0.060	0.060	***7.27*** (2, 190)	4.06 (4, 188)	5.04 (3, 189)
	HIFMECH	-	0.291	-	-	-0.002	0.0107		0.23 (1, 493)	**3.66** (2, 492)	
	WOSCOPS	-	0.315	-	-	-0.000	-0.001		0.66 (1, 1332)	0.33 (2, 1331)	
	ISIS [3]	-	-	-	-	-0.000	0.018		0.58 (1, 1235)	***12.21*** (2, 1234)	

Table 8.1: All studies – Minor allele frequencies, adjusted R^2 and F statistics (with degrees of freedom in regression model) for linear, factorial and saturated models of phenotype regressed on SNPs in non-diseased, non-cases

[1] F statistics are for linear, factorial or saturated model versus null model. Text formatting indicates p-value in ANOVA test of linear versus null, factorial versus linear, or saturated versus linear models. ***Bold-italic:*** $p < 0.001$, **Bold** $0.001 < p < 0.01$, *Italic:* $0.01 < p < 0.05$, Normal: $p > 0.05$.

[2] g1 = rs1205, g2 = rs1130864, g3 = rs1800947, g4 = rs3093077.

[3] SNP rs2808628 used as instrument.

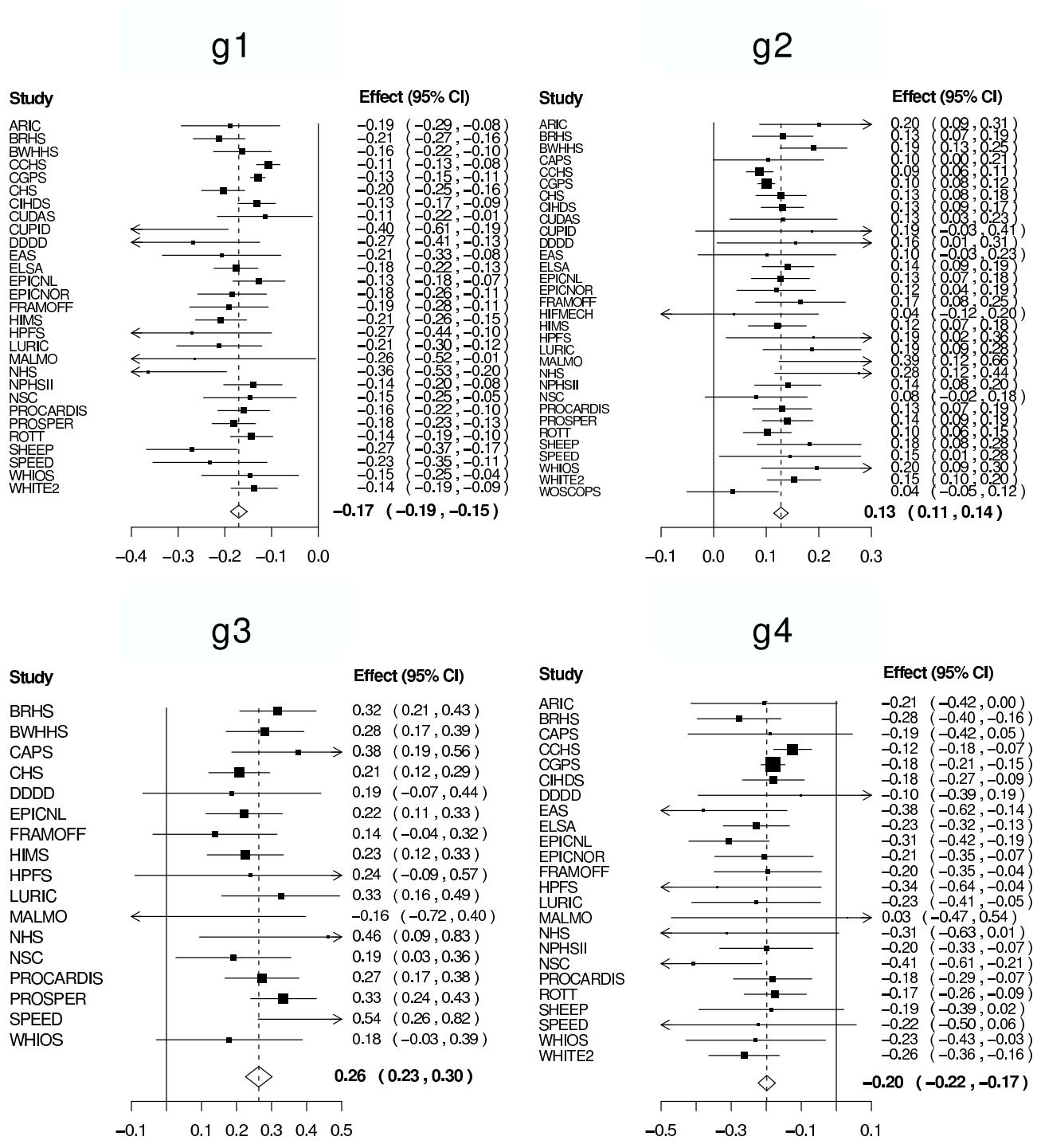


Figure 8.2: Studies with CRP measured – Forest plots for per allele effect of SNPs in univariate regression of log(CRP) on each SNP in non-diseased, non-cases. Pooled effects calculated using two-step random-effects meta-analysis

Haplotype (group)	rs1205 (g1)	rs1130864 (g2)	rs1800947 (g3)	rs3093077 (g4)
Pattern 4: 4 SNPs measured – g1, g2, g3, g4				
1 (I)	**C**	**T**	G	T
2 (II)	**C**	**C**	G	**T**
3 (II)	**C**	**C**	G	**G**
4 (III)	**T**	**C**	**G**	T
5 (III)	**T**	**C**	**C**	T
Pattern 3: 3 SNPs measured – g1, g2, g4				
1 (I)	**C**	**T**		T
2 (II)	**C**	**C**		**T**
3 (II)	**C**	**C**		**G**
7 (III) = 4+5	**T**	**C**		T
Pattern 2: 3 SNPs measured – g1, g2, g3				
1 (I)	**C**	**T**	G	
4 (III)	**T**	**C**	**G**	
5 (III)	**T**	**C**	**C**	
6 (II) = 2+3	**C**	**C**	G	
2 SNPs measured – g1, g2				
1 (I)	**C**	**T**		
6 (II) = 2+3	**C**	**C**		
7 (III) = 4+5	**T**	**C**		

Table 8.2: Candidate haplotypes used as instruments for each combination of SNPs measured. SNPs in bold represent those used as minimal tagging SNPs used for that haplotype

Haplotype (group)	rs1205 (g1)	rs1130864 (g2)	rs1417938 (g2)	rs1800947 (g3)	rs2794521	rs3091244	rs3093068 (g4)	rs3093077 (g4)
1 (I)	**C**	**T**	A	G	T	T	C	T
2 (II)	**C**	**C**	T	G	C	C	C	**T**
3 (II)	**C**	**C**	T	G	T	A	G	**G**
4 (III)	**T**	**C**	T	**G**	T	C	C	T
5 (III)	**T**	**C**	T	**C**	T	C	C	T
6 (II)	**C**	**C**	T	G	?	?	?	**?**
7 (III)	**T**	**C**	T	**?**	T	C	C	T

Table 8.3: Candidate haplotypes used as instruments in all studies. Question marks denote unknown values either due to missing or unmeasured data. SNPs in bold represent those used as minimal tagging SNPs for that haplotype

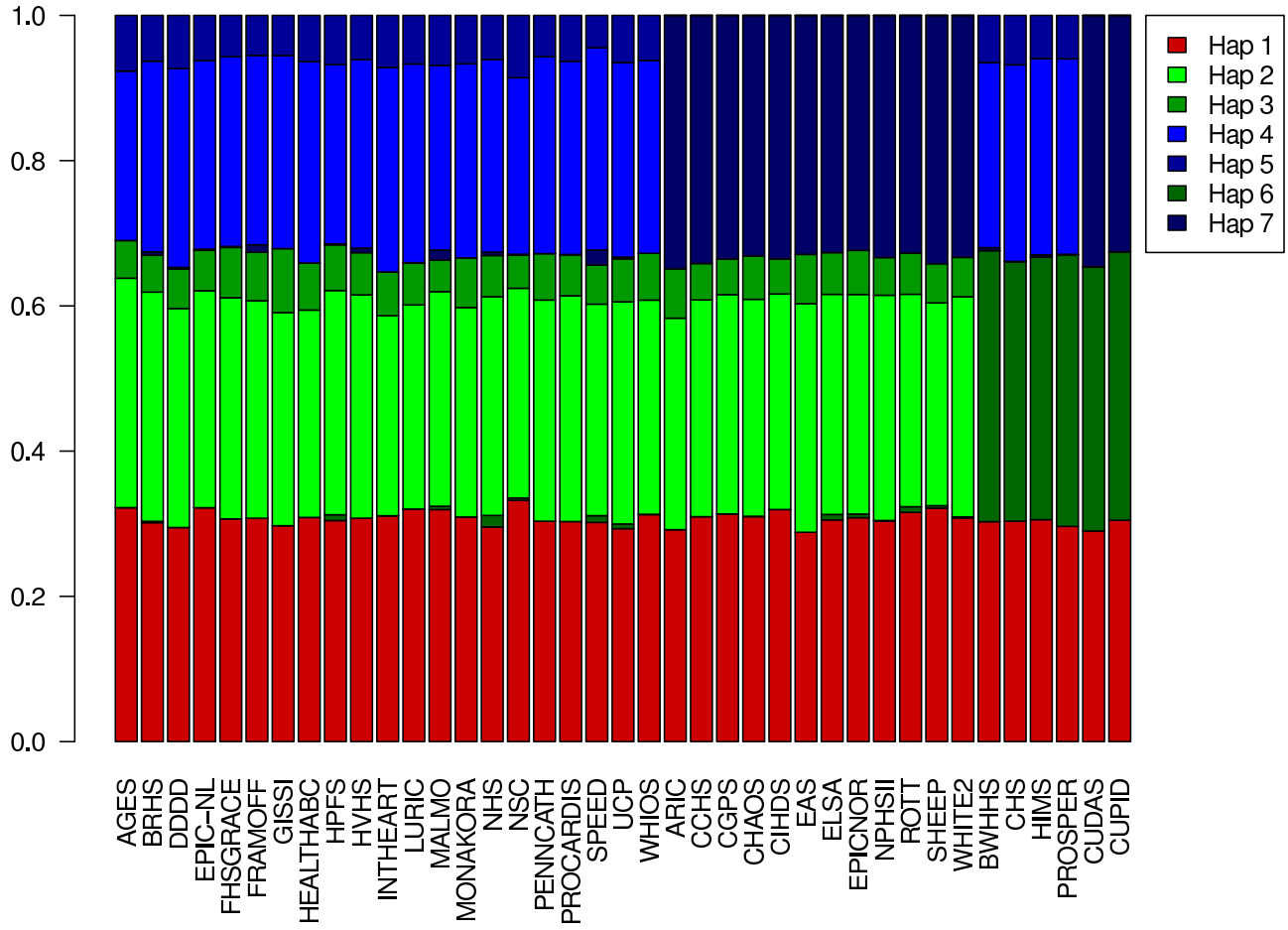


Figure 8.3: All studies with determinable haplotypes – Frequency of haplotypes in each study. Group I haplotypes (haplotype 1) are coloured green, Group II (haplotypes 2, 3 and haplotype category 6) are coloured blue and Group III (haplotypes 4, 5 and haplotype category 7) are coloured red

[1]The equivalent figure including Asian and African descent population groups separately (not displayed) shows that both African and Asian populations have different haplotype frequencies to European descent populations, but have similar frequencies for different study populations within each ethnic classification.

Study	N	Group I Haplo 1	Group II Total	Group II Haplo 2	Group II Haplo 3	Group III Total	Group III Haplo 4	Group III Haplo 5	Other
Pattern 4: Studies measuring 4 SNPs – g1, g2, g3, g4									
AGES	3219	0.322	0.368	0.316	0.052	0.310	0.233	0.077	0
BRHS	3824	0.301	0.368	0.316	0.051	0.330	0.263	0.063	2
DDDD	897	0.295	0.356	0.302	0.054	0.349	0.274	0.073	0
EPICNL	3478	0.322	0.354	0.299	0.056	0.324	0.259	0.062	6
FHSGRACE	4548	0.307	0.374	0.305	0.069	0.320	0.261	0.057	3
FRAMOFF	1680	0.308	0.366	0.300	0.067	0.326	0.261	0.055	2
GISSI	4034	0.297	0.382	0.294	0.088	0.321	0.266	0.055	0
HEALTHABC	1660	0.309	0.350	0.286	0.065	0.341	0.277	0.064	0
HPFS	737	0.304	0.380	0.309	0.063	0.316	0.247	0.068	3
HVHS	4407	0.308	0.366	0.308	0.058	0.327	0.259	0.061	6
INTHEART	4188	0.311	0.336	-	-	0.353	0.282	0.072	2
LURIC	2747	0.320	0.338	0.281	0.057	0.341	0.274	0.067	0
MALMO	2148	0.320	0.344	0.296	0.043	0.337	0.254	0.069	2
MONAKORA	1673	0.309	0.357	0.288	0.068	0.334	0.267	0.067	1
NHS	684	0.296	0.374	0.301	0.057	0.331	0.265	0.061	0
NSC	1673	0.332	0.338	0.289	0.046	0.330	0.243	0.086	0
PENNCATH	1509	0.304	0.368	0.305	0.064	0.328	0.271	0.057	0
PROCARDIS	6464	0.303	0.367	0.311	0.056	0.330	0.266	0.063	1
SPEED	854	0.302	0.354	0.291	0.054	0.344	0.279	0.045	1
UCP	3756	0.293	0.371	0.306	0.059	0.335	0.268	0.065	2
WHIOS	2011	0.313	0.360	0.295	0.064	0.328	0.266	0.062	7
Pattern 3: Studies measuring 3 SNPs – g1, g2, g4									
ARIC	2261	0.292	0.359	0.291	0.068	0.349	-	-	0
CCHS	10259	0.310	0.349	0.298	0.050	0.342	-	-	18
CGPS	32038	0.314	0.351	0.302	0.049	0.336	-	-	49
CHAOS	2475	0.310	0.359	0.299	0.060	0.331	-	-	1
CIHDS	6716	0.320	0.345	0.297	0.048	0.335	-	-	9
EAS	907	0.288	0.383	0.315	0.068	0.329	-	-	0
ELSA	5496	0.305	0.368	0.303	0.057	0.327	-	-	2
EPICNOR	3298	0.308	0.368	0.302	0.061	0.324	-	-	2
NPHSII	2282	0.304	0.363	0.310	0.052	0.333	-	-	8
ROTT	5406	0.316	0.357	0.293	0.057	0.327	-	-	1
SHEEP	2671	0.321	0.337	0.279	0.054	0.342	-	-	1
WHITE2	5515	0.308	0.359	0.303	0.054	0.333	-	-	5
Pattern 2: Studies measuring 3 SNPs – g1, g2, g3									
BWHHS	3771	0.303	0.373	-	-	0.324	0.255	0.065	3
CHS	4511	0.304	0.358	-	-	0.339	0.271	0.068	1
HIMS	3946	0.306	0.361	-	-	0.333	0.270	0.059	3
PROSPER	5777	0.296	0.374	-	-	0.329	0.269	0.060	4
Studies measuring 2 SNPs – g1, g2									
CUDAS	1107	0.290	0.364	-	-	0.346	-	-	5
CUPID	555	0.305	0.370	-	-	0.325	-	-	1

Table 8.4: All studies with determinable haplotypes – Proportion of seven haplotypes patterns in each of three groupings in each study, with total number of participants (N) and number omitted (other) due to not conforming to one of the seven candidate haplotype patterns

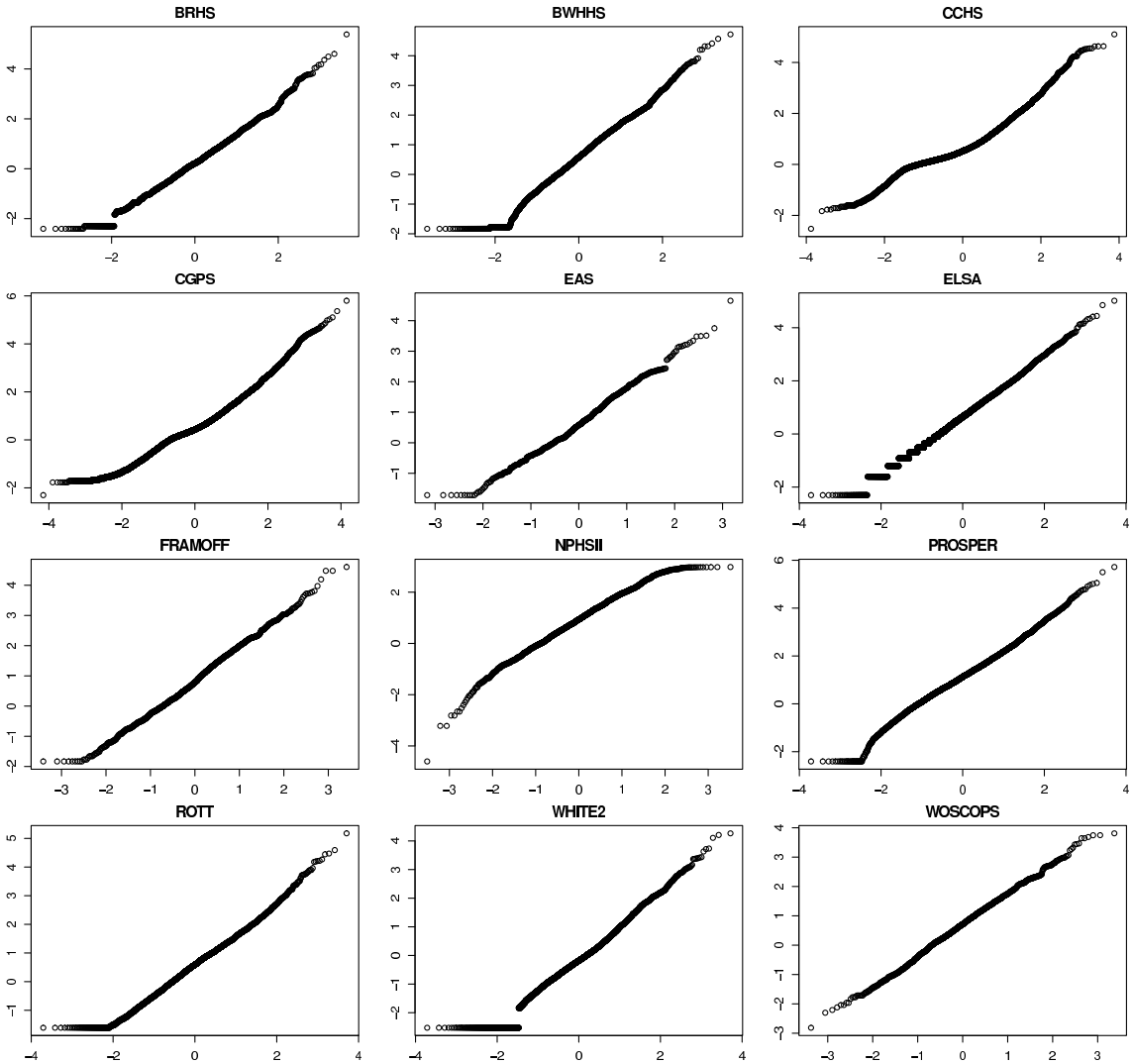


Figure 8.4: Cohort studies – Quantile plot of log(CRP) distribution against quantiles of a normal distribution

[1]CHS is displayed in Figure 8.7 and hence excluded from this figure.

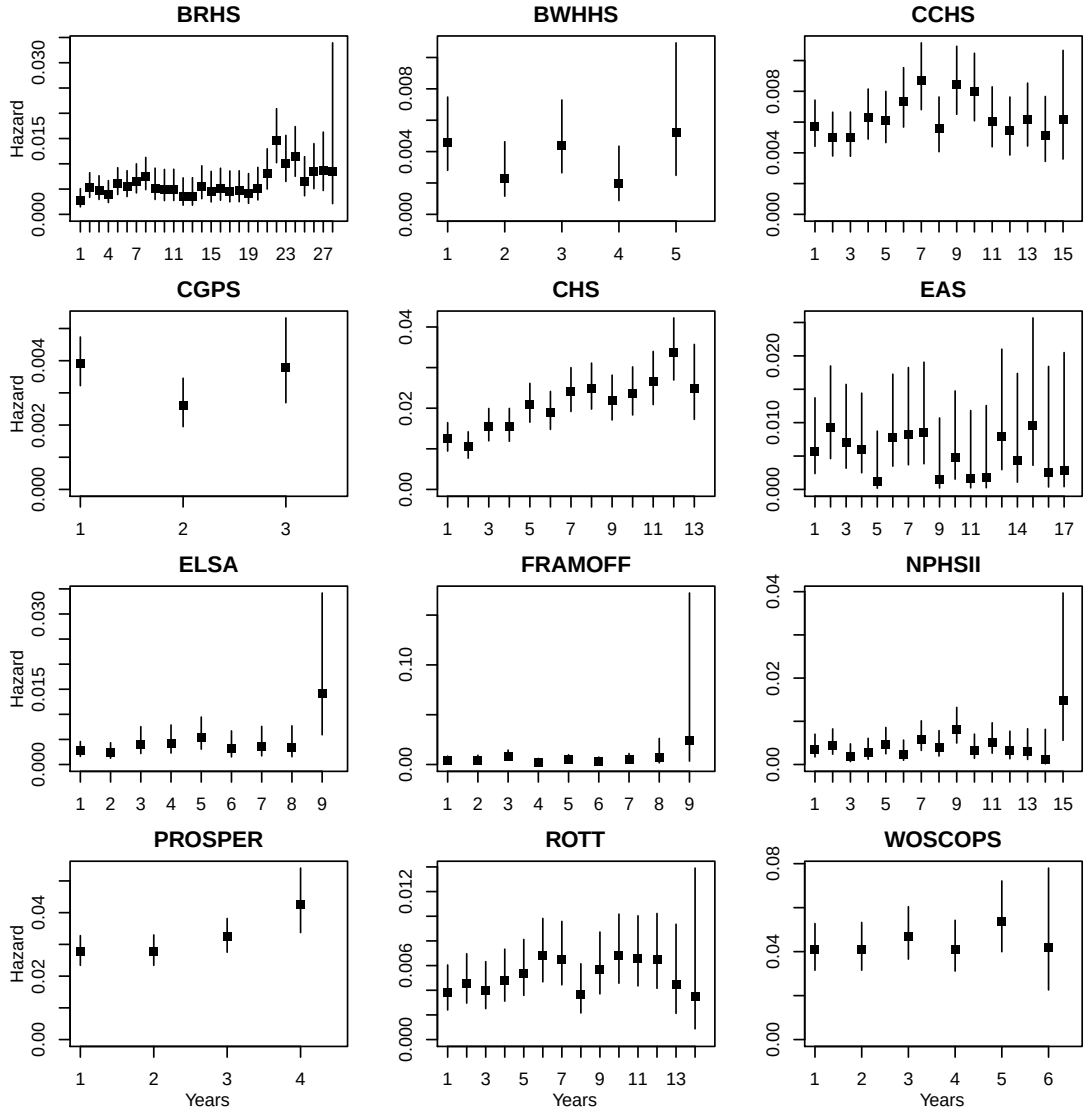


Figure 8.5: Cohort studies – Piecewise constant estimate of hazard function for each year of follow-up (lines are 95% confidence intervals from normal approximation for the log hazard)

[1]WHITE2 does not contain any incident CHD cases. It is therefore only analyzed retrospectively and is omitted from this figure.

8.3 Methods for instrumental variable analysis

In this section, we recall the two-stage method introduced in Chapter 2 and the Bayesian models of Chapters 5 and 6. We then discuss approaches to IV estimation with a survival outcome.

8.3.1 Two-stage methods

Two-stage methods, such as two-stage least squares (2SLS) (117) and two-stage predictor substitution (2SPS) (127), are so called because they can be calculated using two regression stages (93). The first stage (G-X regression) regresses X on G to give fitted values $\hat{X}|G$. The second stage (X-Y regression) regresses Y on the fitted values $\hat{X}|G$ from the first stage regression. In this chapter, we generally use a logistic second-stage regression, although we also use conditional logistic, Cox and Weibull regression. The standard error is taken from the second-stage regression with no correction for uncertainty in the first-stage regression.

We note that we use the term 'two-stage' to refer to a two-stage IV analysis and 'two-step' to a two-step meta-analysis based on combining summary estimates from individual studies. All two-step meta-analyses in this chapter use inverse-variance weighting and the DerSimonian–Laird method of moments to estimate heterogeneity in a random-effects model (228).

8.3.2 Bayesian models

We use a Bayesian framework with vague priors for our model. We divide our population using genetic information into subgroups, where a subgroup contains all individuals in a study with a certain genotype. For each subgroup j, we estimate the mean level of phenotype for the subgroup ξ_j assuming that, for each individual i in the subgroup j, the measured values of phenotype x_{ij} come from a normal distribution with mean ξ_j and variance σ^2, assumed to be common across subgroups. Assuming a logistic model of outcome on phenotype, we model the probability of an event π_j in subgroup j by assuming a binomial distribution of number of events n_j from total number at risk N_j. We use a logistic model and assume a linear relationship between the log-odds of event $\eta_j = \text{logit}(\pi_j)$ and mean level of phenotype ξ_j. The coefficient β_1, the increase in log-odds of event for unit increase in phenotype, is taken as our causal parameter of interest. As in the two-stage methods, we only use the phenotype values x_{ij} for individuals from the control population in a case-control study, and for individuals without previous history of

disease in a cohort study. Individuals with missing phenotype values are still included as cases or controls in the logistic regression.

$$\begin{aligned} X_{ij} &\sim \mathcal{N}(\xi_j, \sigma^2) \\ n_j &\sim \mathcal{B}(N_j, \pi_j) \\ \text{logit}(\pi_j) = \eta_j &= \beta_0 + \beta_1 \xi_j \end{aligned} \tag{8.5}$$

In a meta-analysis context, we combine estimates on the causal parameter across studies in a hierarchical model. In a fixed-effects model, the causal parameter β_1 is the same for each study $m = 1, \ldots, M$:

$$\begin{aligned} X_{ijm} &\sim \mathcal{N}(\xi_{jm}, \sigma_m^2) \\ n_{jm} &\sim \mathcal{B}(N_{jm}, \pi_{jm}) \\ \text{logit}(\pi_{jm}) = \eta_{jm} &= \beta_{0m} + \beta_1 \xi_{jm} \end{aligned} \tag{8.6}$$

In a random-effects model, the causal parameter is allowed to vary between studies, with a normal distribution imposed on the study-level causal parameters. Here, the causal parameter of interest μ_β is the mean causal effect across studies. We replace the final line from (8.6) with

$$\begin{aligned} \text{logit}(\pi_{jm}) = \eta_{jm} &= \beta_{0m} + \beta_{1m} \xi_{jm} \\ \beta_{1m} &\sim \mathcal{N}(\mu_\beta, \tau^2) \end{aligned} \tag{8.7}$$

where τ^2, the variance of the random-effects distribution, is a measure of the between-study heterogeneity in the β_{1m}.

Hence, unlike the two-stage method, the Bayesian analysis is performed in one stage, and the meta-analysis is performed in one step.

In a SNP-based approach, we model the phenotype additively across SNPs with a per allele model for each SNP. For each subgroup j comprising all people with g_{jk} variant allele copies for SNP k, where there are K total SNPs, we estimate the change in phenotype per allele α_k to give average levels of phenotype ξ_j for each subgroup:

$$\xi_j = \alpha_0 + \sum_{k=1}^{K} \alpha_k \, g_{jk} \tag{8.8}$$

Alternatively, we can model the phenotype additively across haplotypes as in model (8.9). For each subgroup j comprising all people with haplotypes h_{1j} and h_{2j}, we can estimate the mean phenotype contribution per haplotype γ_k:

$$\xi_j = \gamma_{h_{1j}} + \gamma_{h_{2j}} \tag{8.9}$$

We note that there is no intercept term γ_0, as each individual has exactly two haplotypes.

In each of the Bayesian analyses below, vague independent $\mathcal{N}(0, 1000^2)$ priors were placed throughout on all regression parameters and independent $\mathcal{U}(0, 20)$ priors on the all standard deviation parameters in normal distributions. Throughout, we use an additive per-allele SNP based model of genetic association (Model 8.1) using the pre-specified SNPs measured in each study. We regard the mean of the posterior distribution as the 'estimate' of the parameter of interest, the standard deviation of the posterior distribution as the 'standard error (SE)', and the 2.5th to the 97.5th percentile range as the '95% confidence interval'.

8.3.3 Survival regression models

Using the two-stage paradigm with survival outcomes, we perform second-stage Cox and Weibull regressions. It is not clear what the parameter estimated by such regressions represents (recalling the difficulty with binary outcomes in Chapter 4), and the results presented here are for comparative purposes only. We also convert the survival outcome into a binary outcome, ignoring variable follow-up, and use a logistic regression model.

In the Bayesian framework, we can use a Weibull distribution of survival times (Model 8.10), with shape parameter r and a log-linear model for the rate parameter μ_j for each individual i in genotypic group j with time-to-event t_{ij}.

$$\begin{aligned} X_{ij} &\sim \mathcal{N}(\xi_j, \sigma^2) \\ T_{ij} &\sim \mathcal{W}(r, \mu_j) \\ \log(\mu_j) = \eta_j &= \beta_0 + \beta_1 \xi_j \end{aligned} \tag{8.10}$$

If there is no event but an individual is right-censored, then we introduce a censoring indicator and use the likelihood contribution from the probability of not seeing an event until the time of censoring. A gamma distribution is used for the prior distribution of r with shape parameter 0.1 and rate parameter 0.1.

An alternative approach, not considered here, would be a Poisson regression model based on numbers of events and person-years of follow-up stratified by year of follow-up. This should estimate a relative rate which closely approximates the hazard ratio.

8.4 Worked example: Cardiovascular Health Study

We firstly analyse the Cardiovascular Health Study (CHS) (203) in detail as a worked example before considering the other studies.

8.4.1 Exploratory analyses

The CHS is an observational study of risk factors for cardiovascular disease in adults 65 years or older. We use cross-sectional baseline data for 4511 subjects of European descent from this study who have data for CRP, of whom 447 have a previous history of CHD, and survival data for the remaining 4064 subjects with no previous history of CHD. 793 of these subjects had an incident CHD event during the follow-up period.

Follow-up for participants ranges up to 13 years. The plot of hazard against year of follow-up shows an increasing risk of CHD event, as well as an increasing probability of censoring, for individuals during the follow-up period (Figure 8.6). 2365 participants have over 10 years of follow-up.

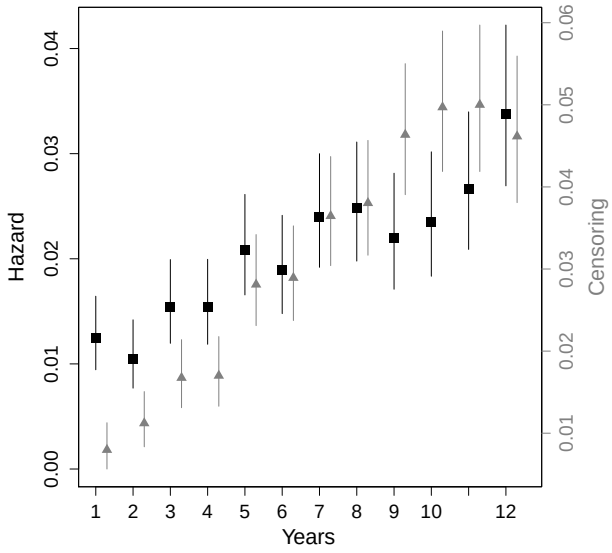


Figure 8.6: CHS – Piecewise constant estimate of hazard function (black squares) and probability of censoring (grey triangles) for each year of follow-up (lines are 95% confidence intervals from normal approximation for log rates)

The distribution of CRP is known to be skewed with large extreme values. It is usual to consider the log-transformed distribution of CRP. Figure 8.7 shows that, aside from extreme values of log(CRP), where the assay method is not sensitive enough to determine between small values, the log-transformed distribution of CRP is similar to a normal distribution.

The Kaplan-Meier curve (Figure 8.8) for CHD outcomes has a curved shape with survivor function decreasing more rapidly throughout the follow-up period. When the population is divided by quintile of CRP, we see separate lines for the survivor function

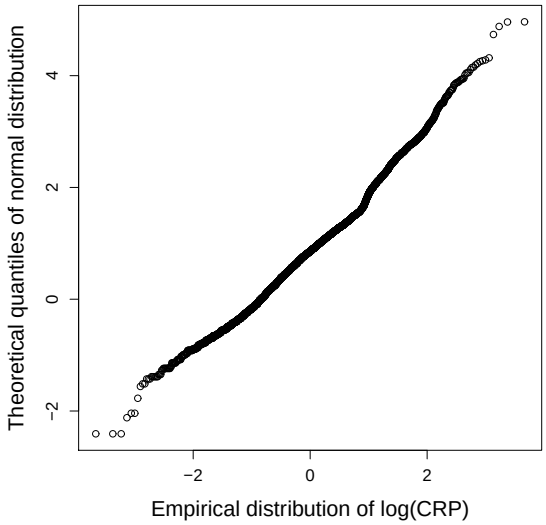


Figure 8.7: CHS – Quantile plot of log(CRP) distribution against quantiles of a normal distribution

with the survivor functions nearly coincident for the lowest two quintiles of CRP, but separated for higher quintiles of CRP with worse survival for increasing levels of CRP.

8.4.2 Observational analysis

We firstly analyse the study prospectively, fitting different survival models and a logistic model to the data to see how they differ in estimates of the association between outcome and CRP. We then analyse the study cross-sectionally (retrospectively) to estimate the observational CRP-CHD association using a logistic model. We adjust in all observational analyses for age at study entry.

We compare a Cox proportional hazards (PH) model, a Weibull PH model, and a logistic model. The Cox PH model is the most flexible, with a non-parametric baseline hazard. The Weibull model uses the Weibull distribution as a parametric baseline hazard function. To assess the suitability of a Weibull distribution, we plot the log cumulative hazard against the log of survival time (Figure 8.9, left pane). If the graph is a straight line, as is approximately the case, then the Weibull assumption is plausible (229). If the graphs when the population is divided into quintiles of CRP are parallel straight lines, as is approximately the case, then a Weibull PH model is appropriate (Figure 8.9, right pane) (229). We estimate cumulative hazard using the Kaplan-Meier estimator.

Alternatively, we can regard survival outcomes as binary data and use logistic regression, taking value 0 for no event and 1 for an event. This ignores the variable follow-up, and

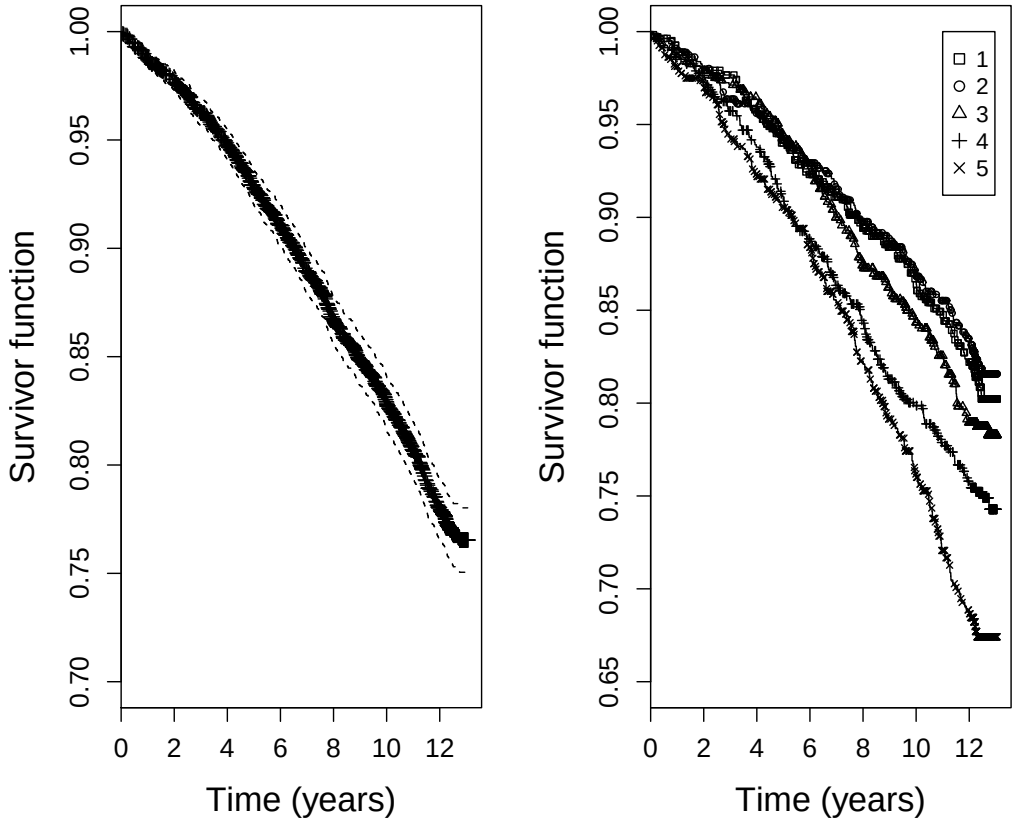


Figure 8.8: CHS – Kaplan-Meier plots, left - for all participants with 95% confidence interval (dashed line), right - divided by quintile of CRP (1 = lowest, 5 = highest)

so may result in a less precise estimate. However, under the assumption that individuals leave the study at random and the disease is rare, the estimates of association should be similar. Instead of a logistic model, we could use a log-linear model where the parameter of interest is a log-relative risk. Under the rare disease assumption, these parameters are approximately equal (230). However, the disease in this case does not seem to be rare, with 19.5% of the participants having a CHD event.

Each of the above models can be fitted in a classical and a Bayesian framework. For computational reasons, we do not present results from a Bayesian Cox PH model. As the two approaches are both based on likelihood, when vague priors are used we should obtain similar results from each method.

Results (Table 8.5) show that the estimates of log-hazard ratios using the Cox and Weibull models are very similar. The logistic model generally shows slightly lower estimates than the Cox or Weibull survival models. The larger standard errors reflect the

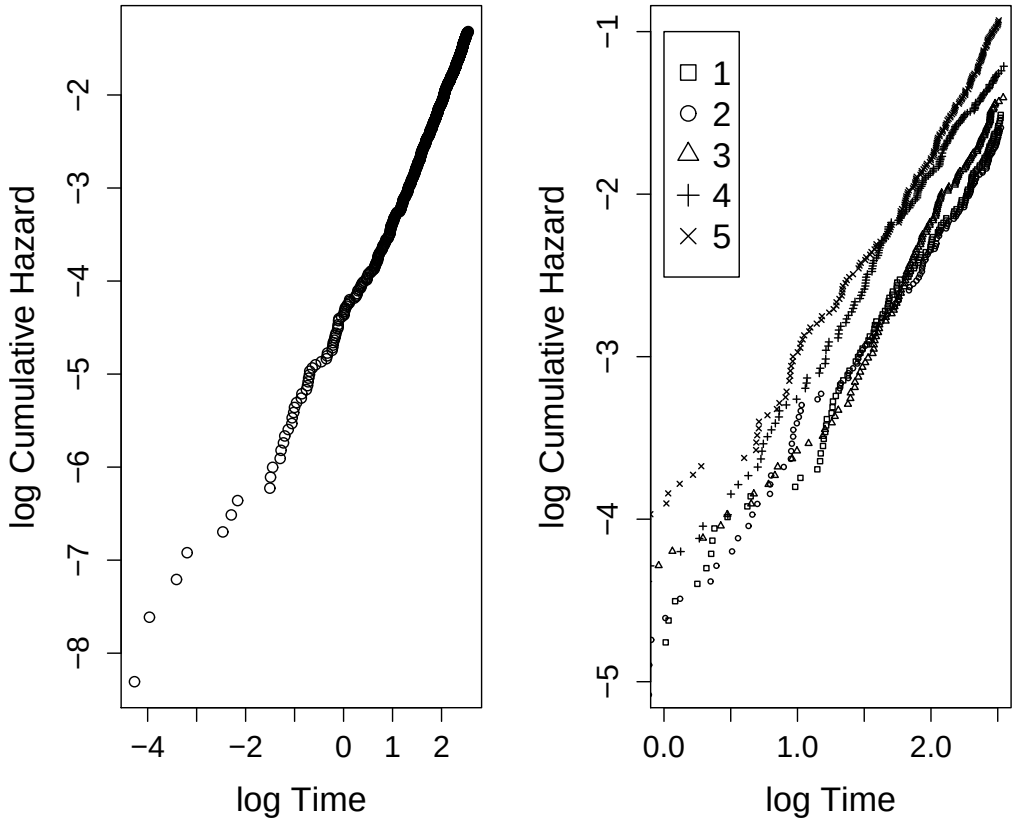


Figure 8.9: CHS – Assessing the Weibull baseline hazard assumption (left) and the proportional hazard assumptions (right)

loss of information in discarding the time-to-event data. The Bayesian estimates are very similar to the classical estimates with a similar degree of uncertainty. The shape parameter in the Weibull method is estimated at 1.372 (95% CI: 1.286 to 1.463) by the classical method and 1.375 (95% CrI: 1.288 to 1.469) by the Bayesian method.

In the cross-sectional (retrospective) analysis, only a logistic model is estimated. Bayesian and classical analyses give very similar results, and the log odds ratios estimated in the prospective and cross-sectional analyses are similar, especially log odds ratio per unit increase in log(CRP).

8.4.3 Causal analysis

To illustrate the instrumental variable method, we provide a visual representation of the causal analysis. We use four genetic variants as instruments: rs1205 (g1), rs1417938 (g2),

Prospective	Classical methods			Bayesian methods	
analysis	log-HR (SE)	log-HR (SE)	log-OR (SE)	log-HR (SE)	log-OR (SE)
Quintile	Cox model	Weibull model	Logistic model	Weibull model	Logistic model
Lowest	0 (reference)	0 (reference)	0 (reference)	0 (reference)	0 (reference)
2	-0.052 (0.124)	-0.052 (0.124)	-0.074 (0.136)	-0.053 (0.123)	-0.074 (0.136)
3	0.166 (0.119)	0.163 (0.119)	0.164 (0.132)	0.166 (0.120)	0.165 (0.131)
4	0.392 (0.115)	0.388 (0.115)	0.379 (0.128)	0.393 (0.116)	0.379 (0.127)
Highest	0.638 (0.111)	0.630 (0.112)	0.580 (0.125)	0.633 (0.113)	0.581 (0.124)
Per unit increase	0.250 (0.034)	0.247 (0.035)	0.227 (0.039)	0.248 (0.035)	0.227 (0.039)
Retrospective	Classical methods			Bayesian methods	
analysis			log-OR (SE)		log-OR (SE)
Quintile			Logistic model		Logistic model
Lowest			0 (reference)		0 (reference)
2			-0.170 (0.179)		-0.170 (0.180)
3			0.211 (0.166)		0.214 (0.166)
4			0.495 (0.159)		0.498 (0.159)
Highest			0.523 (0.158)		0.528 (0.158)
Per unit increase			0.230 (0.047)		0.229 (0.047)

Table 8.5: CHS – Observational log-risk ratio of CHD according to log(CRP) in prospective analysis (study viewed longitudinally with $n = 793$ events out of $N = 4064$ participants) and retrospective analysis (study viewed cross-sectionally with $n = 447$ baseline cases out of $N = 4511$ participants). Cox, Weibull, and logistic models of outcome regressed on quintile of CRP and on log(CRP), adjusting for age at study entry estimated using classical and Bayesian methods: log hazard ratios (HR) and log odds ratios (OR) with standard error (SE)

rs1800947 (g3) and rs2808630 (another SNP in the CRP coding region, here called g5). We divide the population up into genotypic subgroups using each of the genetic variants in turn, and then all of the variants together. For each group, we use bootstrap sampling to estimate the distribution of mean log(CRP) and log-odds of CHD within that group. Graphs are given separately for retrospectively (Figure 8.10) and prospectively assessed CHD (Figure 8.11). Using each of the SNPs individually, gives three subgroups which differ in mean CRP level. The gradient of the line passing through the centre of these distributions represents the causal association. The bottom two graphs in each figure use information from all the SNPs, taking the subgroups with greater than 400 participants, then with greater than 200 participants. Although the picture becomes less clear as more distributions are added, we see that the causal estimates should be more precise with multiple SNPs, as there are more subgroups.

We see from Table 8.6 that the results from different two-stage and Bayesian analyses are similar throughout. Different regression models give fairly similar results, though with some differences due to the different assumptions used for baseline hazard and follow-up, as discussed in the next section. The prior and posterior distributions of β_1 for the retrospective logistic analyses using SNPs g1, g2 and g3 separately are shown in Figure 8.12, and for g5 after 502 000 iterations (first 2000 discarded as 'burn-in') in Figure 8.13. We see that while the posterior distributions using g1, g2 and g3 are very different to the prior distribution, that in the case of g5, much of the information in the posterior distribution comes from the prior. The Markov chain in the MCMC process for g5 spent the majority of the time close to zero, but periodically "wandered off", as can be seen by the posterior distribution having long tails.

8.4.4 Differences between two-stage and Bayesian IV estimates in a single study

Although there is broad agreement between the Bayesian and two-stage IV results in Table 8.6, the differences are considerably greater than those between the classical and Bayesian observational analyses in Table 8.5. We discuss some possible reasons for the differences.

The Bayesian IV estimates in Table 8.6 are generally greater in magnitude than their two-stage counterparts, although p-values are very similar. The increase in size of effect may be due to random error in the mean phenotype estimates in genotypic groups leading to dilution of the regression coefficients in the second-stage regression and attenuation in

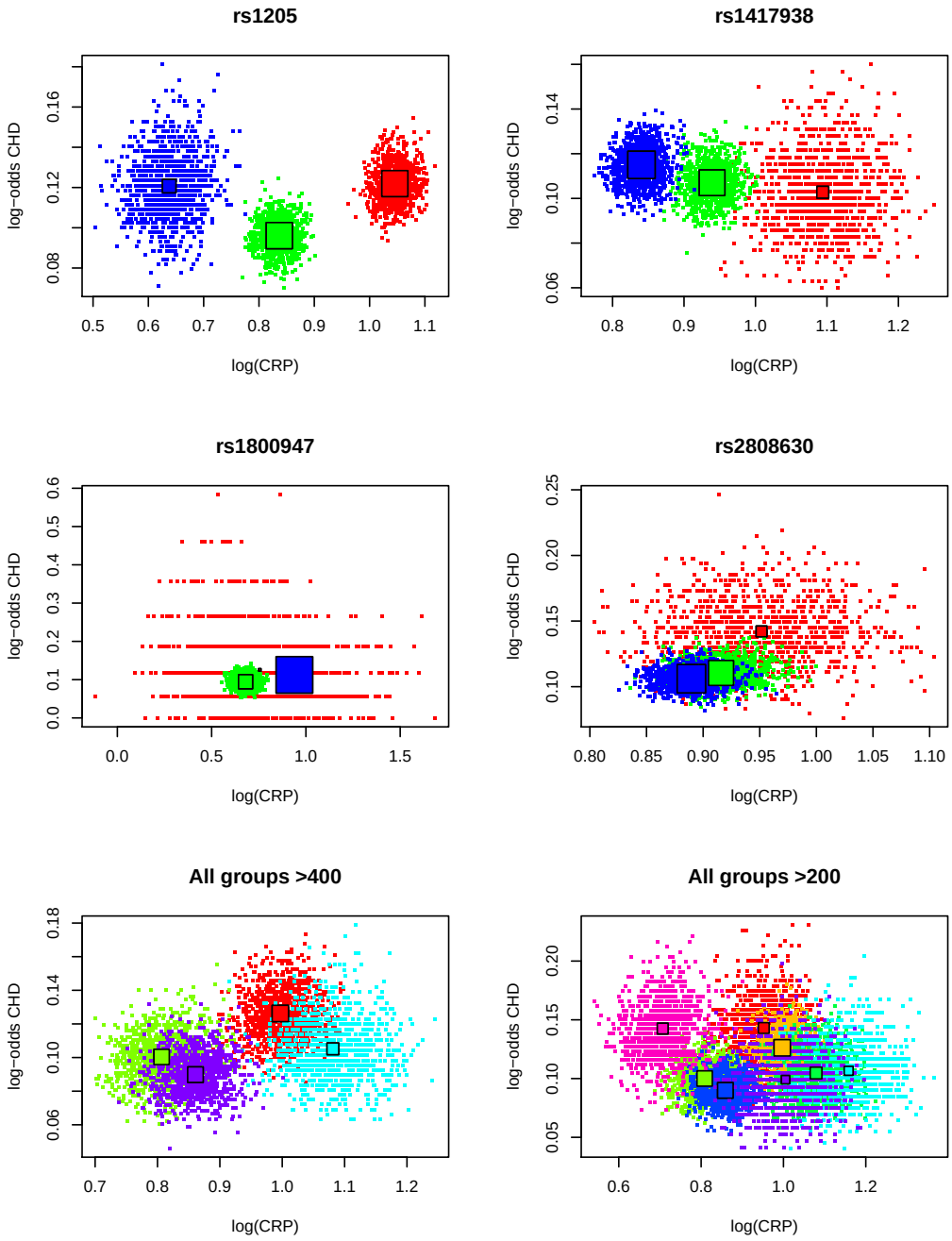


Figure 8.10: CHS – Bootstrap distributions of mean log(CRP) and log-odds of retrospectively assessed CHD within each genetically-defined subgroup with means (area of points is proportional to number of individuals in the group)

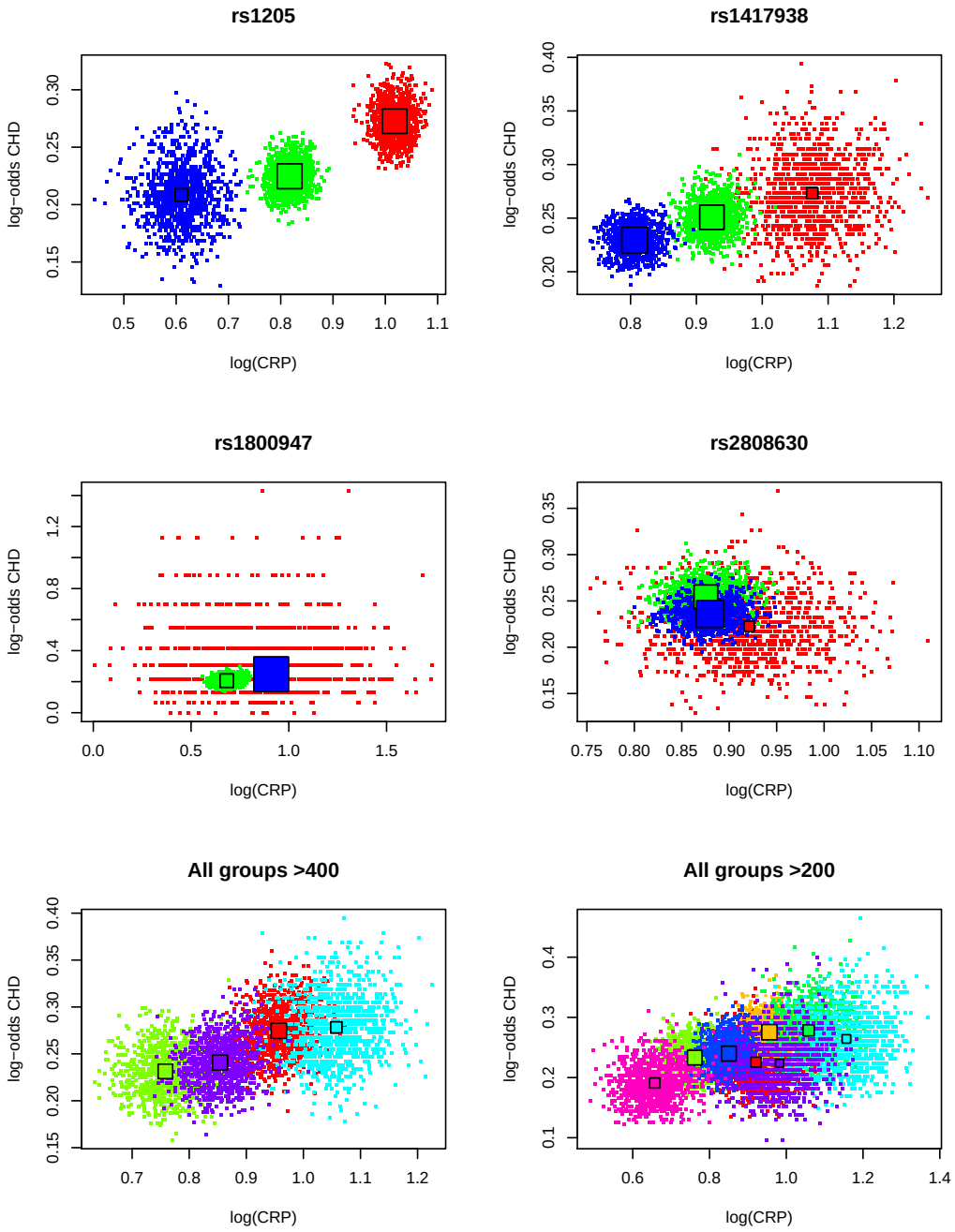


Figure 8.11: CHS – Bootstrap distributions of mean log(CRP) and log-odds of prospectively assessed CHD within each genetically-defined subgroup with means (area of points is proportional to number of individuals in the group)

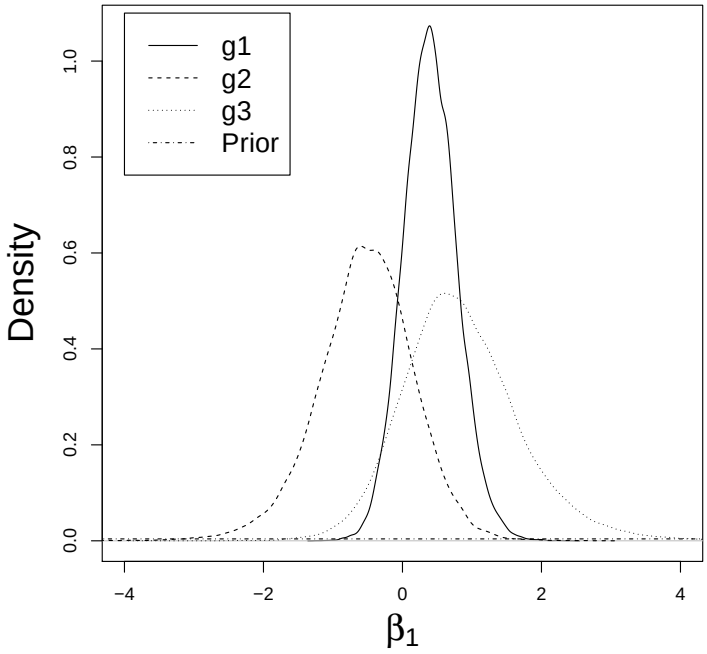


Figure 8.12: CHS – Prior and posterior distributions of β_1 for retrospective logistic analyses using SNPs g1, g2 and g3

the two-stage estimates (231). As the Bayesian analyses allow for error in X, the Bayesian estimates should be unaffected by regression dilution bias.

The Bayesian model estimates causal association in one stage, allowing for propagation of error and feedback throughout the model. In the two-stage model, there is no possibility of propagation of error or feedback from the second-stage to the first-stage regression.

The Bayesian analysis gives a posterior distribution rather than a single point estimate. When the posterior distribution cannot be well-approximated by a normal distribution, the mean and median of the posterior can be quite different, and neither may be an adequate summary of the posterior. The two-stage estimate may be closer to one of the posterior mean or median than the other.

With regards to the causal estimates using g5, Figures 8.10 and 8.11 show that the mean phenotype distributions in the subgroups defined by different numbers of variant alleles of g5 overlap substantially. Visually, the gradient joining the line through the mean phenotype and log odds ratio of the three subgroups in each case could plausibly be either horizontal or vertical. This is expressed in the two-stage method by a large standard error on the causal parameter, but expressed more accurately by the confidence interval in the ratio method from Fieller's Theorem, which covers the entire real line, or by the Bayesian

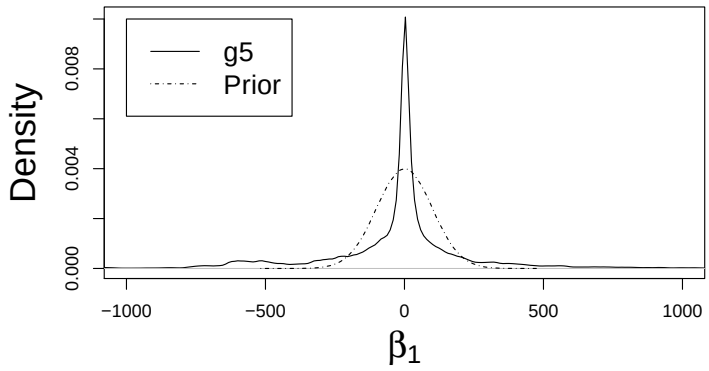


Figure 8.13: CHS – Prior and posterior distributions of β_1 for retrospective logistic analysis using SNP g5 (rs2808630)

method, where the posterior distribution fails to converge. Hence, failure to converge in the Bayesian method is not (necessary) a negative feature, but can be an indication that no proper posterior distribution reflects the uncertainty due to the weakness in the G-X association.

For these reasons, while we would expect the results from a Bayesian and two-stage IV analysis to be close for large studies, they may well give different estimates if the sample size is small, if there are few events, or if the IV is weak.

8.4.5 Summary of causal association in CHS

The estimates of causal association in the prospective analysis confirm the apparent positive causal association of the graphical representation (Figure 8.11). We see how using all of the SNPs as an IV rather than using the SNPs individually gives a more precise estimate of the causal association, synthesizing the individual causal estimates, which will be correlated if the SNPs are in LD. CHS was chosen as an example study as it shows a significant causal effect in some of the analyses: this is not representative of the totality of the data (Section 8.7).

Prospective	Two-stage methods			Bayesian methods [1]	
analysis	log-HR (SE)	log-HR (SE)	log-OR (SE)	log-HR (SE)	log-OR (SE)
Using IV	Cox model	Weibull model	Logistic model	Weibull model	Logistic model
g1	0.664 (0.264)	0.661 (0.265)	0.758 (0.295)	0.680 (0.283)	0.784 (0.320)
g2	0.681 (0.424)	0.673 (0.424)	0.671 (0.475)	0.725 (0.504)	0.728 (0.559)
g3	0.580 (0.505)	0.583 (0.506)	0.723 (0.556)	0.665 (0.621)	0.830 (0.704)
g5	1.525 (5.845)	1.583 (5.846)	1.889 (6.546)		
all	0.609 (0.225)	0.606 (0.226)	0.725 (0.252)	0.600 (0.233)	0.717 (0.264)
Retrospective	Two-stage methods			Bayesian methods [1]	
analysis			log-OR (SE)		log-OR (SE)
Using IV			Logistic model		Logistic model
g1			0.388 (0.366)		0.408 (0.382)
g2			-0.527 (0.671)		-0.531 (0.696)
g3			0.627 (0.620)		0.864 (0.893)
g5			3.521 (2.614)		
all			0.352 (0.322)		0.309 (0.326)

Table 8.6: CHS – Causal log odds ratio of CHD per unit increase in log(CRP) in prospective analysis (study viewed longitudinally with $n = 793$ events out of $N = 4064$ participants) and retrospective analysis (study viewed cross-sectionally with $n = 447$ baseline cases out of $N = 4511$ participants). Cox, Weibull, logistic and log-linear two-stage and Bayesian instrumental variable models: log hazard ratio (HR) and odds ratio (OR) estimates with standard error (SE)

[1]Posterior distribution of causal effect using g5 (rs2808630) did not converge.

8.5 Analysis of individual studies

Having discussed one particular study in detail, we return to the examine the other studies in the collaboration, which we consider in groups corresponding to different study designs. For each of the study designs in the CCGC, we desire to use a logistic model of disease association. This is for three reasons: first, to simplify calculations in the computationally intensive Bayesian framework; secondly, to aim to estimate the same target parameter in each of the studies; and thirdly, because there is an interpretation of the parameter in the logistic case (Chapter 4). In this section, we detail the conditions required for a logistic model to be valid for each study design and examine the difference between IV estimates based on different approaches (two-stage and Bayesian) and different models of association as a sensitivity analysis for the assumptions made in Section 8.7.

In cohort studies, where possible, two analyses are performed, as shown with the CHS analysis of Section 8.4 [1]. A retrospective analysis is performed by viewing the cohort at baseline as a cross-sectional study with cases taken as individuals with previous history of disease (prevalent cases) and controls as all non-diseased individuals. A prospective analysis excludes all prevalent cases and considers CHD events within the reporting period. An individual who is censored at the end of the follow-up period is taken as a control in both the retrospective and prospective analyses as they have two separate opportunities to become a case.

We look in turn at unmatched case-control studies and cohort studies viewed cross-sectionally (retrospectively), then matched case-control studies, and finally cohort studies viewed prospectively. In each case, we use both two-stage and Bayesian models to estimate a causal effect. For each study design, we estimate a pooled estimate from a meta-analysis across all the studies of that design.

8.5.1 Differences between two-stage and Bayesian IV estimates in a meta-analysis

As previously stated, the Bayesian model estimates causal association in one stage. Similarly, the Bayesian meta-analysis model estimates a pooled association in one step. In the Bayesian meta-analysis, the prior for the heterogeneity parameter ensures that the heterogeneity is always positive. In a two-step meta-analysis, the DerSimonian–Laird heterogeneity can be (and is often) zero. If there are not many studies or studies have

[1] We note that the results for the CHS study in this section are different to those in the previous section due to a different choice of instruments.

imprecise estimates, the DerSimonian–Laird estimate may be zero due to lack of evidence of heterogeneity, whereas the Bayesian one-step model sees a lack of information on the between-study variance, and the posterior for τ is similar to the prior. The point estimate changes as heterogeneity increases, as larger studies are down-weighted in comparison to small studies (232).

8.5.2 Unmatched case-control studies and cross-sectional analysis of cohort studies

In the case-control studies and cohort studies viewed cross-sectionally, we use a logistic model in the second stage regression. In both cases, this is the correct analysis, although with a cohort study, a log-linear model could also be used to estimate a relative risk, which is close to the odds ratio estimated by the logistic model under the rare-disease assumption. Table 8.7 shows that the two-stage and Bayesian methods give similar answers in most large studies. Some studies give less consistent results, especially ISIS and HIFMECH, where no results are given as the posterior distribution of the causal effect did not converge. In both of these studies, only one SNP is available and the F statistic in the additive model is less than 1, indicating that the IV explains less of the variation in the phenotype than would be expected by chance. As explained in Section 8.4.4, the Bayesian and two-stage estimates are not likely to agree in such a situation.

8.5.3 Analysis of matched case-control studies

In the matched case-control studies, in the two-stage approach, we use conditional and unconditional logistic models in the second stage regression. In a matched case-control trial, the effect size should be estimated using conditional logistic regression (233), although under certain assumptions about the matching variables, this should be equivalent to unconditional logistic regression. A sufficient condition is that the stratification variables (S) are either:

i. conditionally independent of the outcome given the phenotype ($S \perp\!\!\!\perp Y|X$)

ii. conditionally independent of the phenotype given the outcome ($S \perp\!\!\!\perp X|Y$)

Under this condition, both approaches asymptotically give the same estimates (234). Generally, the regression coefficient from unconditional logistic regression is conservatively biased compared to that from conditional logistic regression (235), but the bias is not generally very severe (188; 233)

In the Bayesian approach, we use an unconditional logistic model, due to issues of computational complexity and difficulty of Bayesian inference on a conditional likelihood. Table 8.8 shows that for most studies the two approaches give broadly similar estimates. The Bayesian and two-stage random-effects pooled results are quite different due to different assumptions about heterogeneity, as stated in Section 8.5.1. The lack of information on between-study heterogeneity due to the paucity of studies and diffuse prior on the heterogeneity parameter in the Bayesian approach gives a large estimate of τ. This conflict can be redressed by use of a more informative prior; two-stage and Bayesian fixed-effect meta-analyses (a point-mass prior for τ concentrated at 0) give much closer results.

8.5.4 Prospective analysis of cohort studies

In the cohort studies (viewed prospectively), as with the CHS analysis of Section 8.4 in the two-stage approach, we use Cox PH, Weibull, and logistic models in the second stage regression. In the Bayesian approach, we use a logistic model (8.5) and a Weibull model (8.10). For most studies, Table 8.9 shows that the approaches give similar estimates. There is a slight loss in precision in using a logistic model over a Cox or Weibull model, due to the loss of time-to-event information. We note that the Bayesian and two-stage analyses give similar inference throughout in studies, especially in studies with over 100 events. The standard error of the causal parameter in the Bayesian Weibull model is occasionally marginally larger than in the logistic model due to Monte Carlo error, despite dropping information on the time-to-event. As in Section 8.5.3, the random-effects meta-analysis results are different between the Bayesian and two-stage analyses, but the fixed-effect results are almost identical.

The correlation between the two-stage IV estimates in cohort studies viewed prospectively and cross-sectionally (using a logistic model in both analyses, similar results using a Cox or Weibull model) is 0.590 (10 studies). Figure 8.14 shows the estimates with 95% CIs from the two analyses for each study as a scatter plot.

8.5.5 Use of covariates

As mentioned in Chapter 3, use of covariates in IV analyses should help strengthen instruments and give more precise IV estimates. In a logistic regression, adjustment for covariates does not necessarily reduce the standard error of a coefficient, as the interpretation of the coefficient changes (Chapter 4), but the power to detect an effect should increase (195). Although adjustment for standard covariates such as age and sex is possible, we particularly want to adjust for other markers of inflammation as they typically

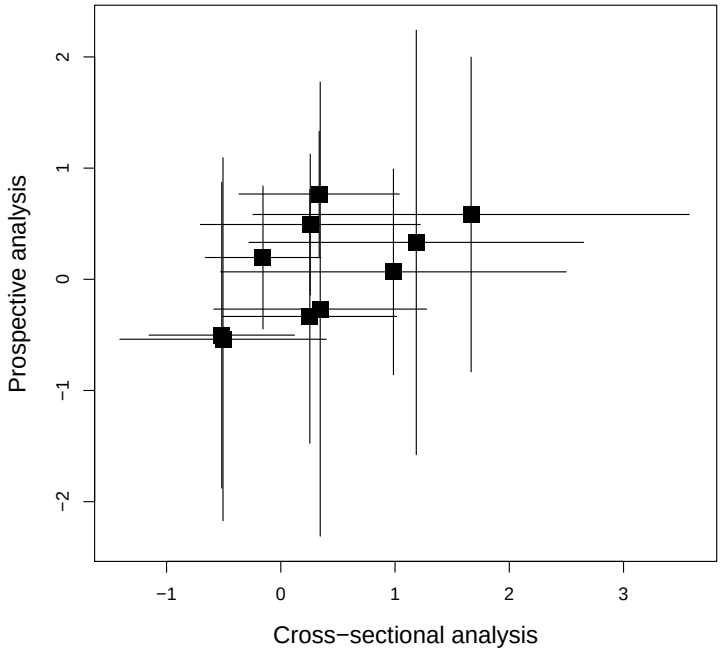


Figure 8.14: Cohort studies with prevalent cases – Scatter plot of two-stage IV log odds ratio estimates from cross-sectional and prospective analysis of each study in turn

explain a large proportion of variation in CRP. However, adjustment for covariates requires individual modelling of CRP, which means that a CRP measurement is needed from each individual. We therefore look at cohort studies viewed prospectively, excluding all participants with a prevalent disease event to avoid reverse causation. We take five cohort studies with measurements for fibrinogen (which should not be on the causal pathway from CRP to CHD, see Chapter 5) and the three cohort studies with measurements for fibrinogen and interleukin-6 (IL6), which was introduced and used as a covariate in Chapter 3. As individual modelling is necessary, we exclude from analysis all individuals with a missing value in CRP, fibrinogen, and IL6 (when adjusted for); hence results without adjustment will be slightly different to those in Section 8.5.4. As is conventional, we use log-transformed IL6, as the distribution of log(IL6) is closer to normal. We use two-stage and Bayesian approaches with a logistic model, adjusting for covariates in both the G-X and X-Y regression stages (118).

We see from Table 8.10 that, despite fibrinogen explaining 15–36% of the variance in log(CRP) and log(IL6) and fibrinogen together explaining 34–49% of the variance, the standard errors of the estimates of causal association did not consistently decrease. We note that the standard errors in Table 8.10 are typically wider than those in Table 8.9 due

to the decrease in sample size caused by restricting analysis to those with measured values of the phenotype and covariate(s). This is in contrast to the findings of Section 3.6.2. Although covariate adjustment increases the strength of the instrument, the uncertainty in the causal estimate is mainly due to the second-stage X-Y regression, not the first stage G-X regression. Adjustment for covariates in a logistic model changes the interpretation of the coefficients in the second-stage regression, generally leading to greater estimates and increased standard errors (191). It seems that covariate adjustment in the CCGC dataset is not a fruitful avenue to pursue. A further technical problem is that the Bayesian model takes longer to run, as each individual has a different level of disease risk in the model, as opposed to all individuals in a genotypic group having the same level of risk.

As inclusion of covariates explaining a large proportion of the variation in the phenotype does not make a great difference to the precision of the causal effect and sometimes increases the standard error while decreasing the available sample size, we conclude that adjustment for covariates is not worth performing in the overall analysis.

8.5.6 Summary of individual study analyses

To summarize Sections 8.4 and 8.5, we see that despite the logistic model relying on certain assumptions, the causal estimates are not particularly sensitive to these assumptions, and the loss of information in discarding survival outcomes is not great. We conclude that using a logistic model in all studies is a reasonable simplifying assumption.

The Bayesian and two-stage approaches make different assumptions in terms of feedback and propagation of errors between the regression stages, normality of the causal estimate, and heterogeneity in the random-effects models. We have seen that, where the number of cases is fairly large ($n > 100$), the sample size is large ($N > 1000$) and the instrument strength is moderate ($F > 5$), the Bayesian and two-stage analyses give similar inferences. In meta-analysis models, the fixed-effects two-stage and Bayesian analyses agree throughout, and the random-effects analyses agree when the number of studies is large (e.g. Table 8.7 with $M = 27$).

	Study	N	n	Two-stage analysis	Bayesian analyses
Case-control studies	ARIC	2261	632	0.249 (0.279)	0.248 (0.314)
	CAPS	1157	198	-0.292 (0.505)	-0.291 (0.600)
	CIHDS	6716	2236	-0.229 (0.225)	-0.240 (0.235)
	CUDAS	1107	56	-0.801 (1.392)	-1.012 (2.176)
	CUPID	555	340	0.250 (0.326)	0.276 (0.491)
	DDDD	897	269	-0.368 (0.446)	-0.517 (0.628)
	EPICNL	3478	426	-0.131 (0.340)	-0.134 (0.347)
	HIFMECH	1006	490	1.022 (2.508)	- [1]
	HIMS	3946	522	-0.461 (0.318)	-0.453 (0.333)
	ISIS	3618	2075	0.467 (1.480)	- [1]
	LURIC	2747	1137	-0.080 (0.212)	-0.086 (0.235)
	MALMO	2148	928	-0.111 (0.158)	-0.099 (0.194)
	PROCARDIS	6464	3126	0.033 (0.180)	0.032 (0.185)
	SHEEP	2671	1113	0.275 (0.216)	0.311 (0.250)
	SPEED	854	90	0.009 (0.488)	0.058 (0.608)
	WHIOS	3756	1725	0.017 (0.202)	0.017 (0.216)
Cohort studies	BRHS	3824	151	0.258 (0.491)	0.259 (0.500)
	BWHHS	3771	236	0.345 (0.475)	0.416 (0.531)
	CCHS	10259	241	0.986 (0.772)	0.988 (0.792)
	CGPS	32038	899	-0.517 (0.325)	-0.518 (0.326)
	CHS	4511	447	0.336 (0.358)	0.349 (0.375)
	EAS	907	28	1.666 (0.974)	1.726 (1.209)
	ELSA	5496	241	-0.506 (0.461)	-0.551 (0.496)
	FRAMOFF	1680	81	1.186 (0.747)	1.261 (0.852)
	PROSPER	5777	768	-0.156 (0.258)	-0.153 (0.261)
	ROTT	5406	614	0.254 (0.388)	0.271 (0.417)
	WHITE2	5515	31	0.535 (0.901)	1.289 (1.238)
	Pooled	122 565	18 900	-0.011 (0.061)	-0.008 (0.065)
	Heterogeneity			$I^2 = 0\%$ (0–33%)	$\hat{\tau} = 0.086$

Table 8.7: Case-control studies and cohort studies viewed cross-sectionally – Log odds ratio of (retrospectively assessed) CHD per unit increase in log(CRP) using two-stage and Bayesian IV methods, number of participants in study (N), number of events (n), pooled results from two-step inverse-variance weighted random-effects meta-analysis (two-stage) or hierarchical random-effects meta-analysis model (Bayesian), heterogeneity estimate (I^2 with 95% confidence interval for two-step method, $\hat{\tau}$ for hierarchical model): log odds ratio estimates with standard error

[1] Posterior distribution of causal effect did not converge.

			Two-stage analyses		Bayesian analyses
Study	N	n	Conditional logistic model	Unconditional logistic model	Logistic model
EPICNOR	3298	1074	0.102 (0.284)	0.125 (0.280)	0.139 (0.319)
HPFS	737	200	-0.372 (0.405)	-0.408 (0.362)	-0.572 (0.543)
NHS	684	196	-0.294 (0.327)	-0.204 (0.308)	-0.228 (0.374)
NSC	1673	577	0.326 (0.327)	0.258 (0.316)	0.245 (0.338)
Pooled (FE)	6392	2047	-0.019 (0.164)	-0.027 (0.156)	-0.031 (0.166)
Pooled (RE)	6392	2047	-0.019 (0.164)	-0.027 (0.156)	-0.063 (0.509)
Heterogeneity			$I^2 = 0\%$ (0–83%)	$I^2 = 0\%$ (0–82%)	$\hat{\tau} = 0.531$

Table 8.8: Matched case-control studies – Conditional and unconditional logistic models for causal log odds ratio of CHD per unit increase in log(CRP) using two-stage and Bayesian IV methods with standard error (SE), number of participants in study (N), number of events (n), pooled results from two-step inverse-variance weighted fixed-effects/random-effects (FE/RE) meta-analysis (two-stage) or hierarchical FE/RE meta-analysis model (Bayesian), heterogeneity estimate (I^2 with 95% confidence interval for two-step method, $\hat{\tau}$ for hierarchical model) from random-effects meta-analysis: log odds ratio estimates with standard error

			Two-stage analyses			Bayesian analyses	
			log-HR	log-HR	log-OR	log-HR	log-OR
Study	N	n	Cox model	Weibull model	Logistic model	Weibull model [1]	Logistic model
BRHS	3824	379	0.463 (0.305)	0.456 (0.306)	0.493 (0.323)	0.51 (0.33)	0.535 (0.351)
BWHHS	3771	43	-0.253 (1.034)	-0.255 (1.036)	-0.268 (1.042)	-0.17 (1.08)	-0.222 (1.085)
CCHS	10259	680	0.038 (0.457)	0.042 (0.457)	0.067 (0.472)	0.04 (0.48)	0.066 (0.482)
CGPS	32038	188	-0.460 (0.699)	-0.460 (0.700)	-0.502 (0.702)	-0.46 (0.71)	-0.516 (0.709)
CHS	4511	793	0.680 (0.258)	0.677 (0.259)	0.767 (0.288)	0.68 (0.28)	0.770 (0.307)
EAS	907	61	0.626 (0.689)	0.629 (0.692)	0.583 (0.722)	0.67 (0.84)	0.611 (0.891)
ELSA	5496	71	-0.487 (0.828)	-0.480 (0.829)	-0.539 (0.833)	-0.46 (0.85)	-0.554 (0.857)
FRAMOFF	1680	46	0.398 (0.965)	0.363 (0.965)	0.332 (0.974)	0.51 (1.21)	0.430 (1.204)
NPHSII	2282	99	-1.729 (0.815)	-1.727 (0.830)	-1.755 (0.837)	-1.94 (0.97)	-2.014 (1.008)
PROSPER	5777	476	0.252 (0.311)	0.237 (0.312)	0.196 (0.328)	0.25 (0.32)	0.205 (0.337)
ROTT	5406	259	-0.313 (0.564)	-0.313 (0.565)	-0.334 (0.582)	-0.35 (0.61)	-0.374 (0.635)
WOSCOPS	1451	279	-0.380 (2.539)	-0.429 (2.540)	-1.287 (2.806)	- [2]	
Pooled (FE)	77402	3374	0.266 (0.137)	0.262 (0.137)	0.251 (0.145)	0.26 (0.13)	0.252 (0.145)
Pooled (RE)	77402	3374	0.208 (0.159)	0.214 (0.156)	0.175 (0.175)	0.14 (0.23)	0.114 (0.139)
Heterogeneity			$I^2 = 14\%$ (0–54)	$I^2 = 12\%$ (0–51)	$I^2 = 19\%$ (0–57)	$\hat{\tau} = 0.38$	$\hat{\tau} = 0.419$

Table 8.9: Cohort studies – Cox, Weibull and logistic models for causal log risk ratio of CHD per unit increase in log(CRP) using two-stage and Bayesian IV methods with standard error (SE), number of participants in study (N), number of events (n), pooled results from two-step inverse-variance weighted fixed-effects/random-effects (FE/RE) meta-analysis (two-stage) or hierarchical FE/RE meta-analysis model (Bayesian), heterogeneity estimate (I^2 with 95% confidence interval for two-step method, $\hat{\tau}$ for hierarchical model): log-hazard ratio (HR) and odds ratio (OR) estimates with standard error

[1] The Weibull models were slower to run and mixed poorly, so results are only given to 2 decimal places due to Monte Carlo random error.

[2] Posterior distributions of causal effect did not converge.

Study	N	n	Two-stage analyses		Bayesian analyses [1]	
			Not adjusted	Adjusted	Not adjusted	Adjusted
			Adjustment for fibrinogen only			
BWHHS	3005	43	-0.261 (1.031)	-0.251 (1.011)	-0.15 (1.09)	-0.14 (1.04)
CCHS	8217	644	-0.182 (0.509)	-0.113 (0.489)	-0.31 (0.54)	-0.21 (0.53)
ELSA	4234	50	-1.081 (0.970)	-0.984 (0.865)	-1.10 (1.00)	-0.98 (0.89)
NPHSII	2153	99	-1.749 (0.834)	-1.998 (0.851)	-1.99 (0.99)	-2.19 (1.01)
ROTT	1775	94	-0.626 (0.834)	-0.436 (0.667)	-0.65 (0.92)	-0.44 (0.70)
			Adjustment for fibrinogen and log(IL6)			
CHS	3728	708	0.666 (0.301)	0.712 (0.314)	0.67 (0.32)	0.71 (0.33)
EAS	612	40	0.709 (0.867)	0.498 (1.063)	0.81 (1.13)	0.54 (1.25)
FRAMOFF	1471	43	0.374 (0.987)	0.166 (1.048)	0.56 (1.34)	0.24 (1.24)

Table 8.10: Cohort studies measuring fibrinogen – Causal log odds ratio of CHD per unit increase in log(CRP) using two-stage and Bayesian IV methods and logistic model with standard error (SE) without and with and adjustment for fibrinogen or fibrinogen and log-transformed interleukin-6 (log(IL6)), number of participants in study (N), number of events (n)

[1] Results given to two decimal places due to Monte Carlo error.

8.6 Dealing with issues of evidence synthesis

In this section, we recall some of the problems and solutions of combining evidence from heterogenous sources. These extensions in the Bayesian framework were first introduced in Section 5.6 and are briefly summarized here.

8.6.1 Cohort studies

We would like to include up to two outcomes for participants in cohort studies in the analysis, one in the study viewed retrospectively and one prospectively. However, the individual's phenotype should only be included once. Additionally, the same parameter should be estimated in both analyses. In the Bayesian model of Section 5.6.1, this is achieved by modelling two regression equations simultaneously. In the two-stage method, we calculate the causal effect separately using prospectively and retrospectively assessed events, combine the two estimates using an inverse-variance weighted fixed-effect meta-analysis, and take the result of this as the study-specific effect. This assumes, incorrectly, that the two estimates are independent; such an assumption is not made in the Bayesian method. Although in this case the phenotype data is used twice, the main source of uncertainty in the causal estimates comes from the second-stage regression, and so inclusion of the phenotype data twice should not add undue precision to the overall pooled result.

8.6.2 Common SNPs and haplotypes

In the Bayesian model, where studies have measured the same SNPs or have measured SNPs identifying the same haplotypes, the parameters of genetic association can be pooled across studies using a random-effects distribution as stated in Sections 5.6.2 and 5.6.3. Where two sets of studies have measured some of the same SNPs, we have not been able to pool parameters of association due to linkage disequilibrium (LD) between SNPs leading to correlation of the parameters. In SNP-based meta-analyses pooling parameters of genetic association, four parameter distributions were used in Pattern 4 studies, three in Pattern 3 and Pattern 2 studies, and one in Pattern 1 studies, leading to a total of eleven parameter distributions. In haplotype-analyses, only six parameter distributions were needed to cover all of the studies (except the four studies where sufficient genetic data to determine haplotypes were not available).

8.6.3 No phenotype data or tabular genetic data

As stated in Section 5.6.4, we can use the random-effects distributions of the genetic association parameters as a predictive distribution or implicit prior to enable inclusion of the 10 studies in the collaboration without phenotype data or with only tabular data in the Bayesian analysis. As no study-specific causal estimate can be obtained for these studies using a two-stage method, they are omitted from the two-step meta-analyses of the two-stage results.

8.7 Meta-analysis

We apply the methods of the previous sections to the CCGC data. Firstly, we look at estimation of the causal effect using a single instrument; then we present overall meta-analyses results from summary two-stage estimates and from Bayesian hierarchical models.

8.7.1 Using instruments one at a time

The forest plot of Figure 8.15 shows the results for the G-Y associations in all of the studies using each SNP in turn. In each case, we use the "correct" regression model: logistic regression for matched case-control studies and cross-sectional analysis of cohort studies, conditional logistic regression for unmatched case-control studies and Cox regression for cohort studies. Prospective and cross-sectional analyses of cohort studies have been combined in fixed-effect meta-analyses to give a single study-specific estimate. We see that the estimates are all close to null.

Using the method of Thompson et al. (71) (see Section 2.12), we calculate causal estimates using each instrument in turn. Confidence intervals are constructed assuming the within-study correlation between G-X and G-Y association is zero, as recommended in the Thompson paper. Results for the G-X and G-Y associations, as well as the causal X-Y association are given in Table 8.11.

The causal estimates from each SNP are similar; heterogeneity of estimates would be evidence against the validity of one or more of the instruments (117). As the causal estimates are derived from the same data and are correlated, they cannot be combined. As none of these analyses uses the totality of the genetic data, a two-stage or Bayesian approach would be preferred.

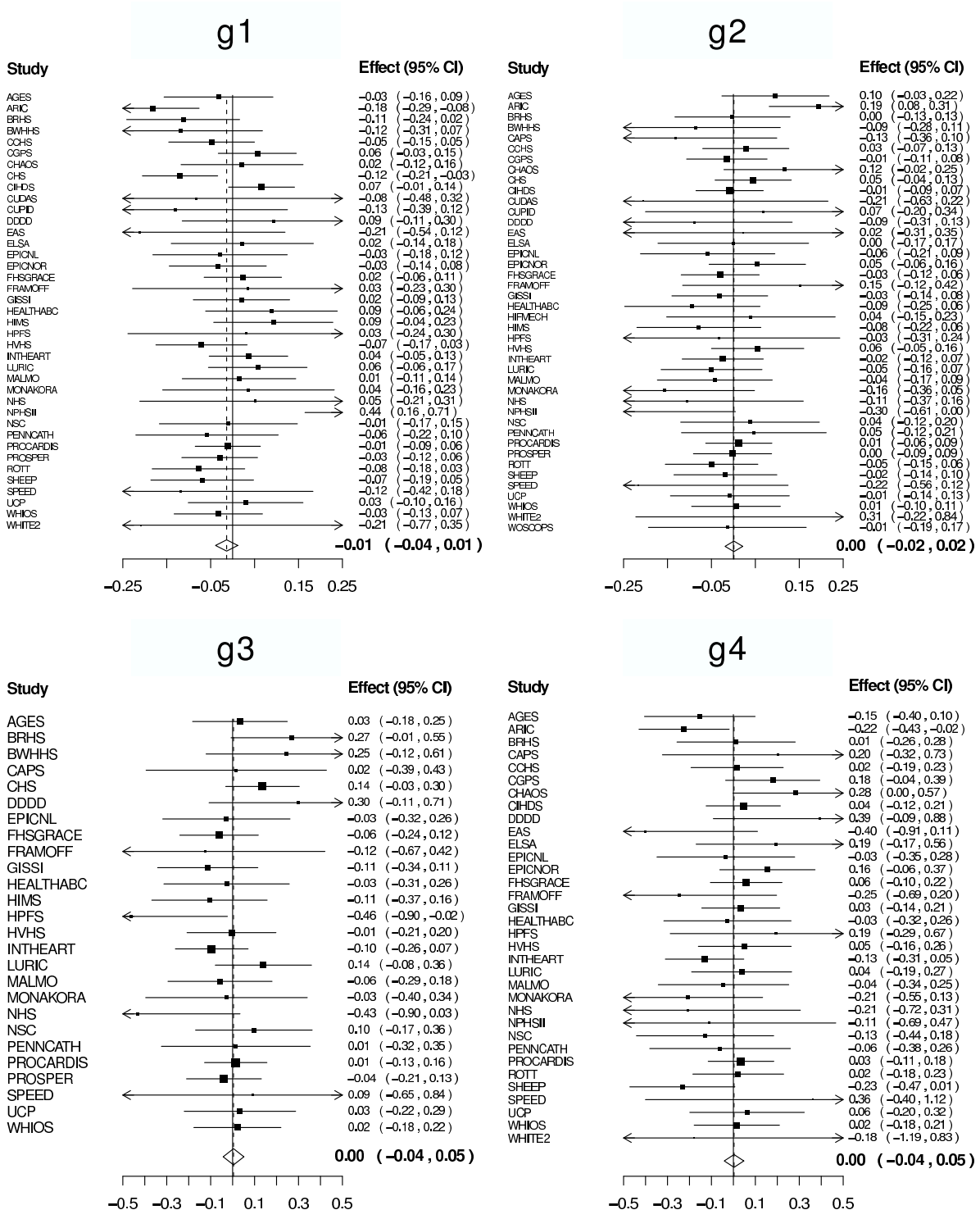


Figure 8.15: All studies – Forest plots for per allele log odds ratio of CHD for each SNP in univariate regression of CHD on the SNP using correct regression. Pooled effects calculated using two-step random-effects meta-analysis

[1]Using logistic regression for unmatched case-control studies and cross-sectional analysis of cohort studies, conditional logistic regression for matched case-control studies and Cox regression for prospective analysis of cohort studies. Prospective and cross-sectional analyses of cohort studies combined in fixed-effect meta-analysis to give single study-specific estimate.

	SNP	Number of studies	Pooled effect (SE)	p-value	Heterogeneity (I^2 and 95% CI)
G-X	g1	29	-0.1703 (0.0097)	2×10^{-77}	58% (37–72%)
	g2	32	0.1281 (0.0070)	2×10^{-75}	29% (0–54%)
	g3	17	0.2635 (0.0194)	3×10^{-42}	14% (0–51%)
	g4	24	-0.1985 (0.0125)	6×10^{-57}	8% (0–41%)
G-Y	g1	39	-0.0136 (0.0129)	0.29	31% (0–54%)
	g2	42	0.0012 (0.0105)	0.91	2% (0–37%)
	g3	26	0.0041 (0.0241)	0.86	0% (0–41%)
	g4	34	0.0030 (0.0227)	0.90	4% (0–32%)
	SNP	Number of studies	Causal estimate (95% CI)		
X-Y	g1	39	0.150 (-0.011, 0.310)		
	g2	42	0.007 (-0.178, 0.191)		
	g3	26	0.122 (-0.109, 0.353)		
	g4	34	-0.033 (-0.315, 0.248)		

Table 8.11: Pooled estimates from two-step inverse-variance weighted random-effects meta-analysis of per allele effect on log(CRP) (*G-X* association) and log odds of CHD (*G-Y* association) in regression on each SNP in turn, heterogeneity estimate; causal estimates (*X-Y* association) of log odds ratio of CHD per unit increase in log(CRP) from meta-analysis using method of Thompson et al. (71)

8.7.2 Using all instruments

Table 8.12 shows the pooled estimates of association using two-stage and Bayesian methods. We used an additive genetic model throughout with all the pre-specified SNPs available in each study as the IV. Figure 8.16 gives a forest plot of the two-stage causal estimates in each study using a logistic model. In the Bayesian analyses, we used either SNPs according to the four patterns or haplotypes. In the haplotype models, we used the seven defined haplotypes (Table 8.3) as instruments. When using a pooled model (Model (5.17) for SNPs, Model (5.18) for haplotypes), studies where CRP has not been measured have been included, resulting in a narrowed confidence interval, and the causal estimate is further from the confounded association, as would be anticipated due to reduction in weak instrument bias if the true causal effect were null.

We see that the causal effect is close to null. The results for the two-stage analysis using logistic regression throughout and the Bayesian analysis without pooling are the most directly comparable, as they use the same data and the same model of association. The point estimates in these analyses are very similar and the 95% CIs are of similar

width, with the Bayesian interval slightly wider. The pooled analyses based on the same data here give a slight reduction in precision, but the pooling enables the inclusion of studies without phenotype data, whence the precision of the causal effect increases.

The prediction interval (236), which represents the the range of values in which the true value of the causal effect for an additional study would be expected to lie with 95% certainty is −0.319 to 0.283. This is calculated from the SNP-based method using data from all the studies. The prediction interval is wider than the pooled estimate due to between-study heterogeneity.

These analyses rule out even a small causal effect of long-term CRP levels on CHD risk, with the upper bound of the 95% CI in the SNP-based pooled analyses using the totality of the data available corresponding to an odds ratio of 1.1 for a unit increase in log(CRP) (which is close to a 1 standard deviation increase in log(CRP) (64)).

Two-stage analyses					
IV used	Studies	N	n	Causal estimate	Heterogeneity
SNPs - Correct regression [1]	33	129777	24135	0.030 (-0.086 to 0.146)	$I^2 = 14\%$ (0–44%)
SNPs - Logistic regression	33	129777	24135	0.024 (-0.092 to 0.140)	$I^2 = 13\%$ (0–43%)
Bayesian analyses					
SNPs - unpooled	33	129777	24135	0.016 (-0.114 to 0.146)	$\hat{\tau} = 0.132$
SNPs - pooled (same studies)	33	129777	24135	0.009 (-0.134 to 0.150)	$\hat{\tau} = 0.153$
SNPs - pooled (all studies)	43	159207	36463	-0.013 (-0.115 to 0.094)	$\hat{\tau} = 0.106$
Haplotypes - unpooled	29	123120	21228	0.023 (-0.094 to 0.146)	$\hat{\tau} = 0.126$
Haplotypes - pooled	39	152678	33589	0.008 (-0.095 to 0.112)	$\hat{\tau} = 0.099$

Table 8.12: All studies measuring CRP (all studies if noted) – Causal estimate of log odds ratio of CHD per unit increase in log(CRP) using all available pre-specified SNPs (unpooled and pooled) or haplotypes (unpooled and pooled) as instruments in random-effect meta-analyses: number of studies, participants (N) and events (n) included in analysis, estimate of causal association (95% confidence interval), heterogeneity estimate (I^2 with 95% confidence interval for two-step method, $\hat{\tau}$ for hierarchical model)

[1] Using logistic regression for unmatched case-control studies and cross-sectional analysis of cohort studies, conditional logistic regression for matched case-control studies and Cox regression for prospective analysis of cohort studies.

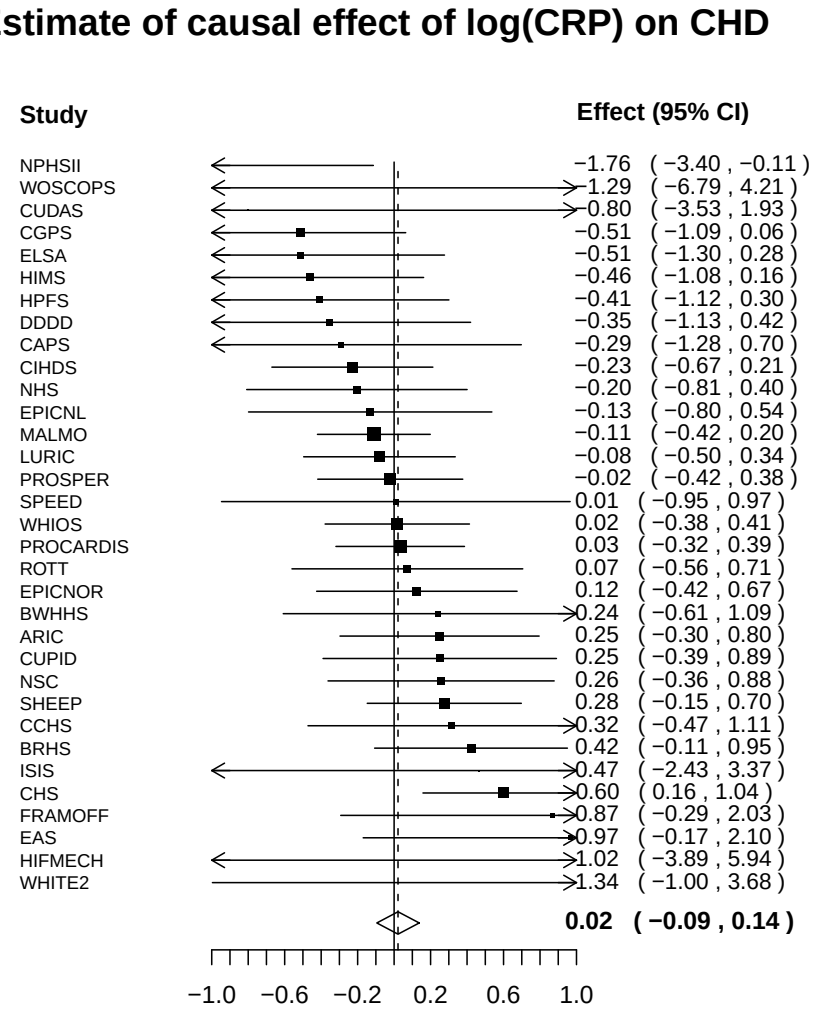


Figure 8.16: All studies measuring CRP – Forest plot for causal estimate of log odds ratio of CHD per unit increase in log(CRP) from two-stage method using logistic regression in each study. Studies ordered by size of causal association. Pooled effect calculated using two-step inverse-variance weighted random-effects meta-analysis

8.8 Discussion

This chapter illustrates methods for synthesis of Mendelian randomization data comprising a variety of study designs and measuring a variety of instruments. Studies with differing design can be analyzed separately and combined in a summary effect meta-analysis, or analyzed together in an individual participant data (IPD) meta-analysis using a Bayesian hierarchical design. Genetic variants can be used as instruments in SNP-based or haplotype-based models. In a Bayesian analysis, genetic effects can be pooled across studies by imposing a random-effects distribution on the study-specific genetic effects. This enables studies without phenotype data to be included in the meta-analysis, with genetic effect estimates drawn from the random-effects distribution. Heterogeneity in the Bayesian model is acknowledged when combining studies by the use of hierarchical models.

8.8.1 Precision of the causal estimate

To obtain a precise estimate of the causal association, one needs to have a precise estimate of both the genotype–phenotype and genotype–outcome associations. A precise estimate of the genotype–phenotype association comes from a study with many participants without a prevalent event, such as a cohort study. A precise estimate of the genotype–disease association comes from a study with many participants with events, such as a case-control study. The proposed Bayesian method borrows strength across all studies to provide a precise estimate of the genetic association in all studies, and therefore obtains a more precise estimate of the causal association. This is illustrated by the width of the 95% confidence interval of the causal parameter reducing from 0.321, 0.369, 0.462 and 0.563 using a single SNP as the instrument (Table 8.11), 0.232 or 0.260 (two-stage method or Bayesian method without pooling, Table 8.12) using all the pre-specified SNPs in an additive model with logistic regression throughout, down to 0.209 or 0.207 (SNP-based and haplotype-based) in the Bayesian method with pooling (Table 8.12) due to the borrowing of information across studies and inclusion of studies without measured phenotype levels. The use of the pooled Bayesian method represents approximately a 136% gain in efficiency compared to the single SNP analyses of Table 8.11, and a 26% gain compared to the two-stage estimate. This compares to the increase in sample size from the two-stage analysis of just under 30 000, and an increase in number of events of around 50%.

8.8.2 Non-collapsibility and heterogeneity

The statistical interpretation of the causal parameter from a Mendelian randomization analysis is an unconfounded population-averaged (marginal) effect. This is not the same estimate as the conditional association in a fully adjusted model (99; 153), due to non-collapsibility of the logistic model (110; 126). In logistic regression, the estimate of association obtained changes when we marginalize over a covariate, even if this covariate is not a confounder (33) (Chapter 4). Although the estimate gives a true test of the null hypothesis, the effect size will be attenuated compared to the association estimated in a fully adjusted model. Further, the IV estimate will be different for populations with different confounder distributions even when the underlying individual change in log-odds of event for unit change in phenotype is the same, giving rise to statistical heterogeneity.

The advantages of using logistic regression in each study is that the estimate in each study has an interpretation, and further each estimate has the same interpretation. Additionally, the computational time needed to obtain precise estimates is not prohibitive. The disadvantages are the assumptions made in ignoring matching and variable follow-up in some studies, and that the estimates will be different in each study as explained above. We prefer to use logistic regression and allow for heterogeneity through the use of random-effects models.

8.8.3 Comparison of two-stage and Bayesian methods

The Bayesian method is known to perform better than the two-stage method in terms of bias and coverage in the presence of weak instruments (237) (Chapter 6). Bias due to finite sample non-zero correlation between the IV and confounders gives rise to a bias in the causal effect in the direction of the observational association (102). As noted in Chapter 3, this is especially evident in a two-step meta-analysis. The Bayesian method is less biased with weak instruments and does not suffer from the problems of underestimated coverage also associated with asymptotic inference in the two-stage method (237) (Chapter 6). In our results, we see that the two-stage method gives slightly narrower confidence intervals than the Bayesian method when the same data is analysed, and the meta-analysis results show estimates from the two-stage method closer to the observational association than the Bayesian results. By using the Bayesian methodology, we can be more certain that our estimate is unbiased and that the true uncertainty of the estimate is expressed. Differences in the way in which uncertainty in between-study heterogeneity also give rise to greater standard errors of the pooled effect in the Bayesian meta-analysis model, although these could be reduced by use of a more realistic prior on the heterogeneity parameter. Despite

this, using the extensions to the Bayesian method described in Chapter 5 and recalled here, a more precise estimate of the causal estimate can be obtained in the Bayesian framework than that given by the two-stage approach, due to the flexibility of the Bayesian framework to make inference on all the data available.

8.8.4 Advantages of individual participant data meta-analysis

Individual participant data (IPD) present a number of advantages to the researcher. Several of the features of this analysis, especially those listed in Section 8.8.5, could not be attempted using summary statistics from each study. This is true both in the two-stage and Bayesian approaches. Specifically in Mendelian randomization, IPD enables the IV assumptions to be assessed carefully in each study as much as is possible (see Appendix F), increasing the plausibility of a causal interpretation from the IV estimate (64). The Bayesian method proposed is also able to incorporate the tabular data from studies which did not share IPD with the collaboration.

8.8.5 Novelty

Several aspects of this analysis are believed to be novel. This is one of the first meta-analyses in Mendelian randomization conducted using IPD, and the first to use a one-step or Bayesian model. The pooling of genetic parameters across studies is novel, as is the inclusion of studies where the phenotype has not been measured. The use of amalgam haplotype categories to represent genetic data across all the studies is novel, as is the inclusion of simultaneous prospective and cross-sectional analyses of cohort studies.

8.8.6 Conclusion

Our methods provide a way of synthesizing heterogenous studies measuring different genetic variants to give a single causal estimate corresponding to a population intervention in long-term phenotype levels based on the totality of available data. By combining all the evidence in this way, we here obtain an estimate precise enough to rule out even a moderate causal effect of CRP on CHD.

8.8.7 Key points from chapter

- The Bayesian and two-stage methods gave different results for single study analyses with small sample sizes and very weak instruments due to different modelling and inference assumptions. When the instruments were robustly associated with the

phenotype, and the number of participants and events were moderate to large ($N > 1000, n > 100$), Bayesian and two-stage methods gave similar results which were not especially sensitive to the modelling assumptions.

- Bayesian and two-stage fixed-effects meta-analyses gave similar results. With a moderate number of studies ($M = 12$), random-effects results differed due to different assumptions on between-study heterogeneity. With a large number of studies ($M = 27$), similar results were obtained.
- The standard error for the Bayesian analyses was greater than for the two-stage analyses, but this may better reflect the true uncertainty in the causal parameter.
- Pooling parameters of genetic association in the Bayesian method allows for inclusion of data from all the studies in the collaboration, leading to more precise estimates of causal association based on the totality of the data available.

Appendix: WinBUGS code for models

Meta-analysis with SNP-based logistic model of association without pooling

```
model {
 mubeta ~ dnorm(0, 0.000001) # prior for mubeta: the causal effect random-effects mean
 sigbeta ~ dunif(0, 20) # prior for sigbeta: the causal effect random-effects sd
 taubeta <- pow(sigbeta, -2) # taubeta: the causal effect random-effects precision
for(m in 1:T) { # m is study number, T is number of studies with CRP data
 for(k in 1:K[m]) { # k is SNP number, K[m] is number of SNPs in each study
  alpha[k,m] ~ dnorm(0, 0.000001) # alpha are study-specific SNP effects
  }
 alpha0[m] ~ dnorm(0, 0.000001) # alpha0 is intercept in G-X regression
 xsd[m] ~ dunif(0, 20) # xsd is standard deviation for phenotype distribution
 xtau[m] <- pow(xsd[m], -2) # xtau is precision for phenotype distribution
 beta[m] ~ dnorm(mubeta, taubeta) # beta is study-specific causal effect
 beta0[m] ~ dnorm(0, 0.000001) # beta0 is intercept in X-Y regression
 muxi[m] <- mean(xi[1:G[m], m]) # muxi is mean phenotype level
 for(j in 1:G[m]) { # j is genotypic group, G is number of groups in study
  for (i in 1:P[j,m]) { # P is number of individuals with phenotype measurement
   x[i, j, m] ~ dnorm(xi[j, m], xtau[m]) # G-X regression
   }
```

```
  xi[j, m] <- alpha0[m] + inprod(alpha[1:K[m], m], gene[j, m, 1:K[m]])
   # gene is number of alleles of each SNP in each genotypic group
  n[j, m] ~ dbin(pi[j, m], N[j, m]) # G-Y regression
  eta[j, m] <- beta0[m] + beta[m] * (xi[j, m] - muxi[m]) # eta is linear predictor
  pi[j, m] <- exp(eta[j, m])/(exp(eta[j, m])+1) # pi is event probability
 } }
for (m in 1:U) { # U is number of cohort studies
 beta0q[m] ~ dnorm(0, 0.000001) # beta0 is intercept in X-Y prospective regression
 for(j in 1:G[m]) {
  nq[j, m] ~ dbin(piq[j, m], Nq[j, m]) # G-Y prospective regression
  etaq[j, m] <- beta0q[m] + beta[m] * (xi[j, m] - muxi[m]) # etaq is linear predictor
  piq[j, m] <- exp(etaq[j, m])/(exp(etaq[j, m])+1) # piq is event probability
  } }
```

Meta-analysis with haplotype-based logistic model of association with pooling

```
model {
 mubeta ~ dnorm(0, 0.000001) # prior for mubeta: the causal effect random-effects mean
 sigbeta ~ dunif(0, 20) # prior for sigbeta: the causal effect random-effects sd
 taubeta <- pow(sigbeta, -2) # taubeta: the causal effect random-effects precision
for(k in 1:K) { # k is haplotype number, K is number of haplotypes in each study
 mugamma[m] ~ dnorm(0, 0.000001)
 } # prior for mugamma: the haplotype random-effects multivariate mean
 K1 <- K-1 # taubeta: the haplotype random-effects precision
 Tau[1:K1, 1:K1] ~ dwish(Tau1[1:K1, 1:K1], K1)
for(m in 1:M) { # M is total number of studies
 gamma[1, m] <- 0 # gamma1 is zero throughout for orthogonality
 gamma[2:K, k] ~ dmnorm(mugamma4[2:K], Tau[1:K1, 1:K1])
# gamma is study specific haplotype effect from random-effects distribution
 beta[m] ~ dnorm(mubeta, taubeta)  # beta is study-specific causal effect
 beta0[m] ~ dnorm(0, 0.000001)  # beta0 is intercept in X-Y regression
 muxi[m] <- mean(xi[1:G[m], m]) # muxi is mean phenotype level
 for(j in 1:G[m]) { # j is genotypic group, G is number of groups in study
  n[j, m] ~ dbin(pi[j, m], N[j, m]) # G-Y regression
  eta[j, m] <- beta0[m] + beta[m] * (xi[j, m] - muxi[m]) # eta is linear predictor
  pi[j, m] <- exp(eta[j, m])/(exp(eta[j, m])+1) # pi is event probability
  xi[j, m] <- gamma0[m] + gamma[h1[j,m], m] + gamma[h2[j,m], m]
   # h1 and h2 are haplotypes for individuals in that genotypic group
```

```
 } }
for(m in (T+1):M) {
 gamma0[m] <- 0 # gamma0 is intercept in G-X regression
 } # for studies with no CRP data, this is not identifiable
for (m in 1:T) { # m is study number, T is number of studies with CRP data
 gamma0[m] ~ dnorm(0, 0.000001)
 xsd[m] ~ dunif(0, 20) # xsd is standard deviation for phenotype distribution
 xtau[m] <- pow(xsd[m], -2) # xtau is precision for phenotype distribution
 for(j in 1:G[m]) { # j indexes genotypic groups, i indexes individuals
  for (i in 1:P[j,m]) { # P is number of individuals with phenotype measurement
   x[i, j, m] ~ dnorm(xi[j, m], xtau[m]) # G-X regression
   } } }
for (m in 1:U) { # U is number of cohort studies
 beta0q[m] ~ dnorm(0, 0.000001) # beta0q is intercept in X-Y prospective regression
 for(j in 1:G[m]) {
  nq[j, m] ~ dbin(piq[j, m], Nq[j, m]) # X-Y prospective regression
  etaq[j, m] <- beta0q[m] + beta[m] * (xi[j, m] - muxi[m]) # etaq is linear predictor
  piq[j, m] <- exp(etaq[j, m])/(exp(etaq[j, m])+1) # piq is event probability
  } } }
```

SNP-based Weibull model in a single study

```
model {
 xtau <- pow(xsig, -2); xsig ~ dunif(0, 20); muxi <- mean(xi[1:G])
 alpha0 ~ dnorm(0, 0.000001); beta0 ~ dnorm(0, 0.000001)
 beta ~ dnorm(0, 0.000001) # priors on parameters as above
 r ~ dgamma(0.1, 0.1) # shape parameter in Weibull distribution
for (k in 1:K) { # k indexes SNPs
 alpha[k] ~ dnorm(0, 0.000001)
 }
for (j in 1:G) { # j indexes genotypic groups
  xi[j] <- alpha0 + inprod(alpha[1:G], gene[j, 1:G])
  log(eta[j]) <- beta0 + beta * (xi[j] - muxi) # log-linear regression
 for (i in 1:P[j]) { # i indexes individuals
  x[i, j] ~ dnorm(xi[j], xtau)
  tx[i, j] ~ dweib(r, eta[j]) I(tc[i, j], ) # dweib is Weibull distribution
   # tx is event time (NA if no event),
   # tc is time at (right-)censoring (NA if no censoring)
 } } }
```

Chapter 9

Conclusions and future directions

9.1 Introduction

In this final chapter, we summarize the findings of each chapter of this dissertation, listing specific contributions and limitations of the work presented (Section 9.2). We then propose ideas for future research (Section 9.3). We finally discuss some general issues relating to Mendelian randomization and instrumental variable (IV) estimation (Section 9.4).

9.2 Summary of the dissertation

We recall each of the chapters of the dissertation in turn, summarizing the main findings, conclusions and limitations of the chapter.

9.2.1 Chapter 1

Chapter 1 provides an introduction to Mendelian randomization. The data to be used in the dissertation from the CRP CHD Genetics Collaboration (CCGC) are introduced, and are subsequently used throughout the dissertation to illustrate statistical findings, and specifically in Chapter 8 to address the question of the causal association of C-reactive protein (CRP) on coronary heart disease (CHD).

9.2.2 Chapter 2

Chapter 2 gives a literature review of statistical methods and issues relating to Mendelian randomization. This review comprises methods for IV analysis and the assumptions necessary for the methods to give valid answers, as well as the specific issues of weak instruments and meta-analysis. The general conclusion from the chapter is that although there has

been much research into IVs in econometric and epidemiological contexts, that this has not made a deep impact into applied Mendelian randomization analysis. This is due to: 1) problems of translation of concepts into an understandable language and a setting representative of a typical Mendelian randomization problem, 2) methodological gaps where concepts which are required in Mendelian randomization analysis are currently poorly understood in the literature, and 3) the lack of power of IV methods leading to reluctance to accept more robust estimation methods. These three strands form the motivation for the direction of the dissertation as a whole.

9.2.3 Chapter 3

Chapter 3 illustrates the problem of weak instrument bias. We demonstrate how IV estimates from finite samples typically have non-normal distributions and non-zero bias. The reasons for the bias are clearly explained, and the magnitude of the bias investigated in different scenarios. A novel measure, median relative bias, is introduced to compare different IV methods, some of which lack first moments. A bias–variance trade-off for the number of instruments used in an IV analysis is shown, and advice given as to how to choose the IV and method of analysis to minimize bias. A key finding is that *post hoc* choice of IVs can result in worse biases than use of the weak IVs themselves.

This chapter takes findings which are known in the econometrics literature about weak instruments and applying them to the context of Mendelian randomization. The conclusions reached in this chapter provide guidance to applied researchers in the planning and analysis of Mendelian randomization studies.

A major limitation of this chapter is that many of the findings rely on the results of simulation studies. However, the results did seem to be consistent under a range of parameter values, and they followed the known theoretical results on relative mean bias closely. A further limitation is that weak instrument analyses in practice are susceptible to bias due to violation of the IV assumptions (183). Although such biases are not unique to weak instrument scenarios, they are likely to be more pronounced with weak instruments as the instrument explains a small proportion of variance in the phenotype, and so any association of the IV with a confounder may be of similar magnitude to the association of the IV with the phenotype of interest. As opposed to the finite-sample problems of weak instruments caused by chance correlation with confounders, true correlation with confounders does not disappear even with increasingly large sample sizes, and leads to bias in IV estimates.

9.2.4 Chapter 4

Chapter 4 introduces the property of collapsibility. Of particular interest is the problem of IV estimation with binary outcomes and logistic regression, as the odds ratio is a non-collapsible measure of association across heterogeneous strata of the population, or across the distribution of a continuous phenotype. Different odds ratios are defined which represent the increase in risk corresponding to a unit increase in the phenotype for an individual or for a population. These odds ratios can be considered marginal or conditional on relevant covariates. We investigate a two-stage method for IV estimation, showing theoretically in a simple case, and by simulation in a more realistic case, how the method estimates a marginal population-averaged odds ratio. An adjustment to the two-stage method proposed in the literature gives estimates which are closer to the parameter estimated from a fully-adjusted logistic model, but which have no interpretation for a general model of confounded association.

This chapter builds on previous work, which focused on collapsibility across a covariate, by introducing the concept of collapsibility across the distribution of the variable intervened upon. Identification of the two-stage estimate as a marginal population-averaged odds ratio gives justification and interpretability to the two-stage method. This runs contrary to the perceived wisdom on "forbidden regressions" from the economics literature: that non-linear two-stage regressions give biased estimates and should be avoided. This chapter advances the debate on the bias of two-stage methods by identifying and defining the quantity estimated.

However the interpretation of the odds ratio from a two-stage analysis represents a limitation on the use of IVs with binary outcomes. The main motivating factor for the use of odds ratios and logistic models of association in conventional epidemiological methods is that the same odds ratio is estimated in a logistic regression analysis of the population and of a case-control sample from the same population. With the population odds ratios estimated by a two-stage approach, this property is not retained. The odds ratio estimated in an IV analysis is not the same in an analysis of the population and of a case-control sample, nor does it remain constant if the distribution of covariates or phenotype in the population changes. This is because the distribution of the covariates is different in the population and in the case-control sample. This phenomenon can also lead to between-study heterogeneity in a meta-analysis, as discussed in Section 8.8.2. However, this problem is not unique to IV estimation: the same objection could be made in a conventional logistic regression analysis for the misspecified individual odds ratio marginalized across a covariate estimated if not all relevant covariates are measured and adjusted for.

9.2.5 Chapter 5

Chapter 5 proposes a Bayesian framework for IV analysis. This framework is initially advocated as an alternative to the two-stage least squares (2SLS) method, giving similar results in analyses of the same data. It has advantages over the two-stage method, particularly in the analysis of data from multiple sources, where a natural extension allows each study to be analysed separately and combined in a hierarchical model on the causal parameter. The Bayesian framework can also be used to analyse binary outcomes with a logistic model, where it again gives results similar to a two-stage method. Several extensions are proposed to the Bayesian meta-analysis model, such as pooling of the parameters of genetic association across studies in a random-effects model, and inclusion of studies where phenotype data has not being measured but the same set of genetic variants has been measured in other studies. In this case, the random-effects distribution is used as an implicit prior for the parameters of genetic association, as there is no study-specific information on these parameters.

This chapter provides a novel framework for meta-analysis of causal associations in studies measuring multiple, possibly different, genetic variants. Chapter 8 shows that such a framework is able to include data from all of the studies in the CCGC in an efficient way.

A major limitation of the Bayesian methods proposed is a reliance on parametric assumptions and specification of error distributions. Due to the computationally intensive nature of the method, it is not always practical to assess departures from the parametric assumptions by sensitivity analyses. The less intensive two-stage method could be used for assessing sensitivity to these assumptions, which may be informative about the behaviour of the Bayesian methods under similar departures.

9.2.6 Chapter 6

Chapter 6 considers bias and coverage properties of various IV methods, primary among which are the two-stage and Bayesian methods. We bring together issues of weak instruments and non-collapsibility with the Bayesian methods introduced in the previous chapter. The chapter is divided into two parts. Firstly, with continuous outcomes, an adjustment is proposed in the Bayesian method to explicitly consider the observational correlation between phenotype and outcome, which leads to IV estimates which are free from bias with even moderately weak instruments. Secondly, with binary outcomes, the analogous adjustment is equivalent to the adjustment considered in Chapter 4. This leads

to estimates which are closer to the parameter usually considered the target of estimation, but which have no clear interpretation.

The Bayesian method also differs from other methods in that inference is based on the posterior distribution, rather than on an asymptotic estimate of the standard error. This leads to better coverage properties, especially with weak instruments, as the posterior distribution accurately expresses the true uncertainty in the causal estimate. Estimates using semi-parametric methods are also considered.

This chapter establishes an empirical justification for use of the Bayesian method, especially in the continuous outcome case. With binary outcomes, the Bayesian method performs similarly to the two-stage method, and so has the same interpretation of the causal effect of interest, leading to the same limitation that the estimand varies depending on the covariate and phenotype distributions.

A limitation of this chapter is that all of the simulations were undertaken with linear or logistic-linear models of association and normal error distributions. Although it is never possible to produce simulations for every possible scenario, no simulations are performed where the semi-parametric methods may be preferable, due to being more robust to departures from the strict distributional assumptions of the two-stage and Bayesian methods.

9.2.7 Chapter 7

Chapter 7 covers the problem of missing data in Mendelian randomization studies. Of particular interest is the problem of sporadic missing genetic data, as these are the hardest to impute and common in applied Mendelian randomization studies. Four methods to impute such missing data are proposed and implemented in a Bayesian model, which is also able to impute missing phenotype and outcome data. These methods are demonstrated to work well, giving improved precision compared to a complete-case analysis in simulations and with real data.

This chapter demonstrates the potential of the Bayesian framework to deal with different statistical issues. Missing data is an especially important issue where multiple genetic variants are measured, as the inclusion of all available genetic variants in a model may lead to loss of sample size due to missing data in particular genetic markers.

A limitation of the methods presented is computational intensity. While the methods work well with the datasets of several thousand, they would be more difficult to use in a meta-analysis context and impractical to use in the analysis of the entirety of data from the CCGC.

9.2.8 Chapter 8

Chapter 8 represents the culmination of the dissertation and the definitive analysis of the causal association between CRP and CHD based on data from the CCGC. Different assumptions of genetic association are made, including SNP-based and haplotype-based models. Different assumptions in the phenotype–outcome association model are made, and results are compared with different analysis models in two-stage and Bayesian frameworks. These are performed initially in one study, then in all studies of a particular design, and finally in all of the studies in the collaboration. Differences between results from the two-stage and Bayesian analyses are explained, and typically are small when the numbers of participants and cases are large, and the instruments are strong. The Bayesian methods lead to the most precise overall estimates as they are able to include data on almost 25% more participants and 50% additional CHD events compared to the two-stage analyses.

This chapter builds on the previous chapters, which have each included estimates of causal association based on individual studies, by combining evidence from all of the studies into a single estimate. This provides an answer to the question of applied research interest based on the totality of the evidence available.

As previously stated, the estimates from individual studies represent marginal population effects. The pooled estimate under a random-effects model does not represent the marginal population effect for any population. The prediction interval, which is calculated from the random-effects distribution, is more relevant if we are interested in the potential size of a causal effect in a new study population.

Several simplifying assumptions are made in the overall meta-analyses, which represent limitations to the analyses. One particular assumption was that all of the studies could be analysed using a logistic model of association. Although sensitivity analyses showed that results were similar for a wide range of assumptions, the results given rely on these assumptions. If studies with different designs were to be analysed using different assumptions, the assumption would then shift to the meta-analysis, to whether estimates of somewhat different parameters from studies of different designs can be combined in a single meta-analysis model.

9.3 Future work

We propose ideas for future work, both extensions to findings in this dissertation, and future directions for Mendelian randomization and genetic epidemiology in general. Most directly, there are several blood-based biomarkers similar to CRP for which consortia

similar to the CCGC can be established, to assess the causal effect of these biomarkers on a range of diseases including CHD.

The vast majority of applied Mendelian randomization analyses have included one single nucleotide polymorphism (SNP), one phenotype, and one disease. The majority of analyses have used the ratio estimate (60). Hence, translational work (eg. (44; 212)) is necessary to bring applied practice up to date with the current methodological state-of-the-art in terms of use of multiple IVs, more robust inference methods (such as the generalized method of moments (GMM), structural mean models (SMM)), and meta-analyses of causal associations.

Other areas which require methodological development are now discussed.

9.3.1 IV estimation using survival data

Although the methods of Section 8.3.3 provide IV estimates with survival data, they are fairly *ad hoc* and are only included in the dissertation for purpose of sensitivity analysis. Although the two-stage and Bayesian approaches proposed may lead to meaningful estimates, a more principled approach to estimation and interpretation with survival outcomes should be possible. One suggestion for such an approach is an accelerated failure-time model, as this has proved to be a good choice in other aspects of causal modelling (238).

9.3.2 Mendelian randomization with GWAS data

A genome-wide association study (GWAS) is an examination of the whole genome of group of individuals to discover genes associated with a particular trait or disease (239). Such studies present difficulties due to the sheer number of genetic variants which are tested for association. Stringent levels for p-values, such as $p < 10^{-7}$, have been used as a threshold for statistical significance to minimize the number of false-positive findings. However, such a stringent p-value means that the power to detect relevant variants may be low. Rather than testing many genetic variants, if there are several genes associated with a phenotype, a concordant relationship between the number of phenotype-increasing alleles across all these genetic variants and the trait of interest may be interpreted as evidence of a causal effect of the phenotype, even if none of the variants individually reaches the threshold for significance.

A Mendelian randomization approach adds an extra dimension to the interpretation of a GWAS. GWAS were designed to facilitate discovery of genetic variants which are associated with a disease. This is useful for prediction of disease and identification of individuals at elevated disease risk, but since the genetics of an individual cannot be changed,

the consequences of GWAS in terms of finding therapeutic targets can only come via a Mendelian randomization paradigm. Equally, the Mendelian randomization paradigm can be used in reverse, by searching for modifiable risk factors which are associated with SNPs which have already been shown to be robustly associated with disease.

If there are many known genetic variants associated with a phenotype, fitting a model of genetic association may require construction of a gene score or weighted gene score to avoid problems of multiple IVs such as weak instruments (44). It is not known what the impact of the assumptions necessary to construct a gene score are, or how it is best to combine information on a large number of genetic variants.

9.3.3 Hypothesis-free inference

The interpretation of GWAS above through the lens of Mendelian randomization requires prior knowledge and understanding of the function of genetic variants used as IVs. Where this is not available, the use of a genetic variant as an IV is questionable. It has been claimed (240) that a hypothesis-free approach to Mendelian randomization could be developed, where the association between genetic variant and disease, and genetic variant and phenotype are both considered simultaneously. In the spirit of the GWAS, no hypothesis is assumed and the data are allowed to speak for themselves. The idea is that pleiotropy and other violations of the IV assumptions are avoided by sheer weight of data. Suppose 1000 SNPs which are associated with the outcome are concordantly associated with the phenotype. Although it may be plausible for each of these associations individually to be due to, say, pleiotropy, it is implausible for all 1000 associations to be. We are then led to the conclusion that the phenotype is a true cause of variation in the outcome. This approach could be especially fruitful for multifactorial polygenic phenotypes, such as body mass index (BMI) or height, where genetic variants associated with the phenotype are found in many sites on many different chromosomes.

While such an approach is possible, it is currently unclear whether there are benefits over traditional Mendelian randomization, what is the impact of violation of the IV assumptions, or how to analyse such data efficiently.

9.3.4 Untangling multifactorial associations

Another possible extension of Mendelian randomization is to examine multiple risk factors simultaneously. In the inflammation pathway, for example, there are many other factors in addition to CRP that have an observational association with CHD risk. Looking at each of the markers in isolation is not the true goal of scientific inquiry, and leads to limited

conclusions. Examining several risk factors simultaneously would help to clarify the overall picture of disease aetiology. If a dataset has information on a number of phenotypes and genetic variants associated with each phenotype, then the causal effect of each phenotype on disease risk can simultaneously be estimated. Although additional assumptions about the causal pathway of disease may be necessary, simultaneous estimation of the effect of different phenotypes may improve the precision of causal estimates, due to a large proportion of the unmeasured confounders overlapping for each phenotype–disease association. Such analysis would require high-quality data and sensitivity analyses to assess the impact of assumptions about the causal pathways from genetic variants to phenotypes to disease.

9.3.5 Pathway analysis

Data on multiple phenotypes and genetic instruments collected in a cross-sectional sample of the population can be investigated in a Mendelian randomization setting. Here, the target of analysis would not be the causal association of a particular phenotype or set of phenotypes on disease, but the network of causal associations between phenotypes. Knowledge about such networks or pathways is informative about the underlying biological associations between risk factors, which may help to identify possible therapeutic targets (241). If the phenotypes vary over time, it may be necessary measure data at different time-points to investigate the temporal behaviour of the pathway.

9.4 Discussion

We finally discuss some general issues relating to Mendelian randomization and instrumental variable (IV) estimation, which have arisen as a result of this dissertation, but which do not fit neatly into any of the previous chapters.

9.4.1 Relevance of the dissertation to areas outside Mendelian randomization

The Bayesian framework introduced in this dissertation provides an alternative to the strong and sometimes misleading asymptotic assumptions necessary for inference in IV methods. This may have relevance in small sample IV problems, where the econometric literature currently lacks generally applicable methods robust to weak instruments (210).

Identification of the two-stage IV estimate for binary outcomes has relevance to the debate in the econometrics literature about the validity of such two-stage methods (127), and in the randomized trials literature, where the difference between marginal and conditional

estimates is recognized, but the interpretation of IV estimates addressing non-compliance is less well understood (128; 242).

Generally, although the focus of this dissertation has been Mendelian randomization, some of the issues discussed apply to a range of problems, including the use of IVs in other research contexts.

9.4.2 Differences between economic and epidemiological contexts

An issue which has been in the background throughout this dissertation has been the difference between the priorities in economic (or econometric) and epidemiological contexts. In translating methodology and findings between the established econometrics literature and the emerging field of Mendelian randomization, there are some differences in terminology between the two literatures. Once the researcher has become fluent in both languages, they realize that the applied problems faced by the fields are different. As applied problems tend to be the motivation for methodological research, this means that the two areas have evolved and specialized to deal with different issues.

One particular difference between the fields is most evident through the justification for using a candidate IV. In economics, to use a single instrument requires strong *a priori* belief in the validity of the IV assumptions, and to use multiple instruments without employing an overidentification test is anathema (118). In epidemiology, the belief for validity of a genetic IV comes via biological knowledge, with empirical justification from testing the association of the IV with various known risk factors. In economics, there are so few points agreed on by all economists that *a priori* belief and external knowledge are not to be relied on. Hence, much of the economics literature revolves around a barrage of tests for the validity of IVs and IV estimates. Although there is some justification for their use in an epidemiological context (30), the priorities for epidemiologists usually lie elsewhere.

9.4.3 Mendelian randomization and conventional epidemiological methods

Part of the scientific backdrop to this methodological dissertation is a controversy about the role of CRP in atherosclerosis and cardiovascular disease, with evidence from a randomized trial that statins reduce the risk of cardiovascular disease to a greater level than expected in a population with low lipid levels and elevated CRP levels (243).

The concept of causation has different meanings to different people. For example, to a biochemist, the question of causality is one of function. The question "Is CRP causally

implicated in atherosclerosis?" can be seen as equivalent to "In the absence of CRP, can atherosclerosis take place?". If the presence of CRP is necessary for the formation of atherosclerotic plaques then, on a biochemical level, CRP is causal for CHD. However, the epidemiological interpretation of the causal question of interest is: "What is the impact of an increase in CRP on CHD risk?". This is the relevant aetiological question from a clinical point of view where the primary concern is public health and patient risk. It may be that the amount of CRP necessary for the formation of atherosclerotic plaques is so small that no intervention can reduce CRP to a level where the CHD risk is eliminated. The biochemical notion of causation does not necessarily inform about the consequences of an intervention targeted at CRP.

In a randomized controlled trial (RCT), a population is chosen and intervened upon at a specific point in their disease progression. In Mendelian randomization, the genetic natural experiment occurs at conception. This means that the estimate from a RCT represents the answer to the question: "What is the effect on the study population of an intervention in usual CRP levels starting today?". A Mendelian randomization estimate typically represents the answer to the question: "What is the effect on the study population of an intervention in usual CRP levels across the life course?". It is conceivable that an intervention across the life course may have more impact than a targeted intervention even at a critical stage of disease development. Results also typically differ due to the choice of study population, which in a RCT is often recruited from a clinical context, whereas in Mendelian randomization is usually chosen from a population-based cohort or case-control study.

The estimate given by a statistical analysis should always be thought of as an answer to a question. When the question changes, we should also expect the answer to change. Hence incompatibility of estimates from different methodological approaches may not represent an antinomy, and assessing the reasons for the apparent contradiction requires *a priori* knowledge and reasoning, not statistical testing alone.

9.4.4 Conclusion

In conclusion, we recall that the aim of this dissertation was to help "bridge the gap" between statistical methodology and applied practice. While there remain many problems to address, we hope that the explanations, interpretations and methodological tools provided in this dissertation will help to bring the two research communities closer, facilitating better collaboration and leading to research which is more effective, efficient and credible.

References

[1] Khoury M, Beaty T, Cohen B. *Fundamentals of genetic epidemiology*. Oxford University Press, USA, 1993. 1

[2] Lawlor D, Harbord R, Sterne J, 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. 1, 2, 5, 7, 10, 11, 19, 21, 26, 29, 30, 39, 40, 41, 43, 97, 101, 123, 130, 181, 182, 183

[3] Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 2003; **32**(1):1–22, doi:10.1093/ije/dyg070. 1, 3, 4, 5, 9, 26, 73, 97

[4] Sheehan N, Didelez V, Burton P, Tobin M. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Medicine* 2008; **5**(8):e177, doi:10.1371/journal.pmed.0050177. 1, 26

[5] Holy Bible. *New international version*. Zondervan, 1984. 3

[6] Darwin C. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. Murray, London, 1859. 3

[7] Darwin C. *The descent of man and selection in relation to sex*. Murray, London, 1871. 3

[8] Mendel G. Versuche über Pflanzen-hybriden. *Verhandlungen des naturforschenden Vereines in Brünn [Proceedings of the Natural History Society of Brünn]* 1866; **4**:3–47. 3

[9] Fisher R. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 1918; **52**:399–433. 3

[10] Pauling L, Itano H, Singer S, Wells I. Sickle cell anemia, a molecular disease. *Science* 1949; **110**:543–548. 3

[11] Watson J, Crick F. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 1953; **171**(4356):737–738, doi:10.1038/171737a0. 3

[12] Roberts L, Davenport R, Pennisi E, Marshall E. A history of the Human Genome Project. *Science* 2001; **291**(5507):1195, doi:10.1126/science.291.5507.1195. 3

[13] The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* 2001; **409**(6822):934–941, doi:10.1038/35057157. 3

[14] Shendure J, Ji H. Next-generation DNA sequencing. *Nature Biotechnology* 2008; **26**(10):1135–1145, doi:10.1038/nbt1486. 3

[15] Rubin D. For objective causal inference, design trumps analysis. *Annals of Applied Statistics* 2008; **2**(3):808–840, doi:10.1214/08-aoas187. 4

[16] Taubes G, Mann C. Epidemiology faces its limits. *Science* 1995; **269**(5221):164–169. 4

[17] Khaw K, Bingham S, Welch A, Luben R, Wareham N, Oakes S, Day N. Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study: a prospective population study. *The Lancet* 2001; **357**(9257):657–663, doi:10.1016/s0140-6736(00)04128-3. 4

[18] Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20536 high-risk individuals: a randomised placebo-controlled trial. *The Lancet* 2002; **360**(9326):23–33, doi:10.1016/s0140-6736(02)09328-5. 4

[19] Peto R, Doll R, Buckley J, Sporn M. Can dietary beta-carotene materially reduce human cancer rates? *Nature* 1981; **290**:201–208, doi:10.1038/290201a0. 4

[20] Hennekens C, Buring J, Manson J, Stampfer M, Rosner B, Cook N, Belanger C, LaMotte F, Gaziano J, Ridker P, Willett W, Peto R. Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *New England Journal of Medicine* 1996; **334**(18):1145–1149. 4

[21] Hooper L, Ness A, Davey Smith G. Antioxidant strategy for cardiovascular diseases. *The Lancet* 2001; **357**(9269):1705–1706, doi:10.1016/s0140-6736(00)04876-5. 4

[22] Writing Group for the Women's Health Initiative Investigators. 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* 2002; **288**(3):321–333, doi:10.1001/jama.288.3.321. 4

[23] Million Women Study Collaborators. Breast cancer and hormone-replacement therapy in the Million Women Study. *The Lancet* 2003; **362**(9382):419–427, doi:10.1016/s0140-6736(03)14065-2. 4

[24] Ebrahim S, Davey Smith G. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics* 2008; **123**(1):15–33, doi:10.1007/s00439-007-0448-6. 5, 8, 10, 12

[25] Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**(434):444–455. 5, 7, 23, 24

[26] Zohoori N, Savitz D. Econometric approaches to epidemiologic data: Relating endogeneity and unobserved heterogeneity to confounding. *Annals of Epidemiology* 1997; **7**(4):251–257, doi:10.1016/s1047-2797(97)00023-9. 5, 60

[27] Davidson R, MacKinnon J. *Estimation and inference in econometrics. Chapter 18: Simultaneous equation models*. Oxford University Press, USA, 1993. 5, 33, 56, 131

[28] Greenland S. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* 2000; **29**(4):722–729, doi:10.1093/ije/29.4.722. 5, 7, 22, 24, 29

[29] Sussman J, Hayward R. An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. *British Medical Journal* 2010; **340**:c2073, doi:10.1136/bmj.c2073. 5, 25

[30] Wehby G, Ohsfeldt R, Murray J. "Mendelian randomization" equals instrumental variable analysis with genetic instruments. *Statistics in Medicine* 2008; **27**(15):2745–2749, doi:10.1002/sim.3255. 7, 38, 39, 44, 169, 239

[31] Lawlor D, Windmeijer F, Davey Smith G. Is Mendelian randomization "lost in translation": Comments on "Mendelian randomization equals instrumental variable analysis with genetic instruments" by Wehby et al. *Statistics in Medicine* 2008; **27**(15):2750–2755, doi:10.1002/sim.3308. 7, 43, 44, 121

[32] Wheatley K, Gray R. Commentary: Mendelian randomization–an update on its use to evaluate allogeneic stem cell transplantation in leukaemia. *International Journal of Epidemiology* 2004; **33**(1):15–17, doi:10.1093/ije/dyg313. 7, 8

[33] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 2007; **16**(4):309–330, doi:10.1177/0962280206077743. 7, 10, 21, 23, 25, 26, 28, 30, 73, 126, 141, 152, 158, 225

[34] Martens E, Pestman W, de Boer A, Belitser S, Klungel O. Instrumental variables: application and limitations. *Epidemiology* 2006; **17**(3):260–267, doi:10.1097/01.ede.0000215160.88317.cb. 7, 26, 29, 46

[35] Thomas D, Conti D. Commentary: the concept of 'Mendelian Randomization'. *International Journal of Epidemiology* 2004; **33**(1):21–25, doi:10.1093/ije/dyh048. 7, 25

[36] Bochud M, Chiolero A, Elston R, Paccaud F. A cautionary note on the use of Mendelian randomization to infer causation in observational epidemiology. *International Journal of Epidemiology* 2008; **37**(2):414–416, doi:10.1093/ije/dym186. 7, 25

[37] Davey Smith G, Lawlor D, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Medicine* 2007; **4**(12):e352, doi:10.1371/journal.pmed.0040352. 7

[38] Ridker P, Paynter N, Danik J, Glynn R. Interpretation of Mendelian randomization studies and the search for causal pathways in atherothrombosis: the need for caution. *Metabolic Syndrome and Related Disorders* 2010; **8**(6):465–469, doi:10.1089/met.2010.0071. 7, 25

[39] Davey Smith G. Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. *Genes & Nutrition* 2010; **6**(1):27–43, doi:10.1007/s12263-010-0181-y. 7

[40] Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* 2004; **33**(1):30–42, doi:10.1093/ije/dyh132. 7, 8, 9, 10, 11, 13, 25, 39, 156, 180

[41] Lewis S, Davey Smith G. Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiology Biomarkers & Prevention* 2005; **14**(8):1967–1971, doi:10.1158/1055-9965.epi-05-0196. 8, 42

[42] Schatzkin A, Abnet C, Cross A, Gunter M, Pfeiffer R, Gail M, Lim U, Davey Smith G. Mendelian randomization: how it can – and cannot – help confirm causal relations between nutrition and cancer. *Cancer Prevention Research* 2009; **2**(2):104–113, doi:10.1158/1940-6207.capr-08-0070. 8, 9

[43] Chen L, Davey Smith G, Harbord R, Lewis S. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLoS Medicine* 2008; **5**(3):e52, doi:10.1371/journal.pmed.0050052. 8, 12

[44] Pierce B, Ahsan H, VanderWeele T. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology* 2011; **40**(3):740–752, doi:10.1093/ije/dyq151. 8, 19, 40, 64, 71, 156, 236, 237

[45] Palmer T, Lawlor D, Harbord R, Sheehan N, Tobias J, Timpson N, Davey Smith G, Sterne J. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research* 2011; doi:10.1177/0962280210394459. 8, 21, 64, 156

[46] Davey Smith G, Ebrahim S. What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures? *British Medical Journal* 2005; **330**(7499):1076–1079, doi:10.1136/bmj.330.7499.1076. 8, 9

[47] Davey Smith G. Randomised by (your) god: robust inference from an observational study design. *Journal of Epidemiology and Community Health* 2006; **60**(5):382–388, doi:10.1136/jech.2004.031880. 8, 10, 96, 156

[48] Nitsch D, Molokhia M, Smeeth L, DeStavola B, Whittaker J, Leon D. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *American Journal of Epidemiology* 2006; **163**(5):397–403, doi:10.1093/aje/kwj062. 8

[49] Fisher R. *Statistical methods and scientific inference*. Oliver and Boyd: Edinburgh, 1935. 9

[50] Keavney B. Commentary: Katan's remarkable foresight: genes and causality 18 years on. *International Journal of Epidemiology* 2004; **33**(1):11. 9, 10

[51] Hernán M, Robins J. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006; **17**(4):360–372, doi:10.1097/01.ede.0000222409.00878.37. 9, 25, 26

[52] Terwilliger J, Weiss K. Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Annals of Medicine* 2003; **35**(7):532–544, doi:10.1080/07853890310015181. 9

[53] Pai J, Mukamal K, Rexrode K, Rimm E. C-reactive protein (CRP) gene polymorphisms, CRP levels, and risk of incident coronary heart disease in two nested case-control studies. *PLoS ONE* 2008; **3**(1):e1395, doi:10.1371/journal.pone.0001395. 9

[54] Tobin M, Minelli C, Burton P, Thompson J. Commentary: Development of Mendelian randomization: from hypothesis test to 'Mendelian deconfounding'. *International Journal of Epidemiology* 2004; **33**(1):26–29, doi:10.1093/ije/dyh016. 10, 26, 29, 142

[55] Brennan P. Commentary: Mendelian randomization and gene-environment interaction. *International Journal of Epidemiology* 2004; **33**(1):17–21, doi:10.1093/ije/dyh033. 10

[56] Pearl J. *Causality: models, reasoning, and inference. Chapter 6: Simpson's paradox, confounding and collapsibility*. Cambridge University Press, 2000. 10

[57] Thomas D. *Statistical methods in genetic epidemiology*. Oxford University Press, USA, 2004. 10

[58] Sleiman P, Grant S. Mendelian randomization in the era of genomewide association studies. *Clinical Chemistry* 2010; **56**(5):723–728, doi:10.1373/clinchem.2009.141564. 10, 26

[59] Ziegler A, König I, Pahlke F. *A statistical approach to genetic epidemiology: concepts and applications, with an e-learning platform*. Wiley-VCh, 2010. 10

[60] Bochud M, Rousson V. Usefulness of Mendelian randomization in observational epidemiology. *International Journal of Environmental Research and Public Health* 2010; **7**(3):711–728, doi:10.3390/ijerph7030711. 11, 12, 236

[61] Timpson N, Lawlor D, Harbord R, Gaunt T, Day I, Palmer L, Hattersley A, Ebrahim S, Lowe G, Rumley A, Davey Smith G. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. *The Lancet* 2005; **366**(9501):1954–1959, doi:10.1016/S0140-6736(05)67786-0. 12, 43

[62] Kivimäki M, Lawlor D, Davey Smith G, Kumari M, Donald A, Britton A, Casas J, Shah T, Brunner E, Timpson N, Halcox J, Miller M, Humphries S, Deanfield J, Marmot M, Hingorani A. Does high C-reactive protein concentration increase atherosclerosis? The Whitehall II Study. *PLoS ONE* 2008; **3**(8):e3013, doi:10.1371/journal.pone.0003013. 12, 43

[63] Elliott P, Chambers J, Zhang W, Clarke R, Hopewell J, Peden J, Erdmann J, Braund P, Engert J, Bennett D, Coin L, Ashby D, Tzoulaki I, Brown I, Mt-Isa S, McCarthy M, Peltonen L, Freimer N, Farrall M, Ruokonen A, Hamsten A, Lim N, Froguel P, Waterworth D, Vollenweider P, Waeber G, Jarvelin M, Mooser V, Scott J, Hall A, Schunkert H, Anand S, Collins R, Samani N, Watkins H, Kooner J. Genetic loci associated with C-reactive protein levels and risk of coronary heart disease. *Journal of the American Medical Association* 2009; **302**(1):37–48, doi:10.1001/jama.2009.954. 12, 43

[64] CRP CHD Genetics Collaboration. Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *British Medical Journal* 2011; **342**:d548, doi:10.1136/bmj.d548. 12, 13, 17, 46, 67, 144, 157, 180, 222, 226

[65] Allin K, Nordestgaard B, Zacho J, Tybjærg-Hansen A, Bojesen S. C-reactive protein and the risk of cancer: a Mendelian randomization study. *Journal of the National Cancer Institute* 2010; **102**(3):202–206, doi:10.1093/jnci/djp459. 12, 37, 44

[66] Casas J, Bautista L, Smeeth L, Sharma P, Hingorani A. Homocysteine and stroke: evidence on a causal link from Mendelian randomisation. *The Lancet* 2005; **365**(9455):224–232, doi:10.1016/s0140-6736(05)17742-3. 12

[67] Ding E, Song Y, Manson J, Hunter D, Lee C, Rifai N, Buring J, Gaziano J, Liu S. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. *New England Journal of Medicine* 2009; **361**(12):1152–1163, doi:10.1056/NEJMoa0804381. 12, 43

[68] Kamstrup P, Tybjaerg-Hansen A, Steffensen R, Nordestgaard B. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *Journal of the American Medical Association* 2009; **301**(22):2331–2339, doi:10.1001/jama.2009.801. 12, 43

[69] Voight B, *et al.*. A Mendelian randomization study for plasma high-density lipoprotein cholesterol. *Rejected by New England Journal of Medicine (twice), Proceedings of the National Academy of Sciences, currently submitted to the Lancet* 2011; . 12, 97

[70] Trompet S, Jukema J, Katan M, Blauw G, Sattar N, Buckley B, Caslake M, Ford I, Shepherd J, Westendorp R, de Craen A. Apolipoprotein E genotype, plasma cholesterol, and cancer: a Mendelian randomization study. *American Journal of Epidemiology* 2009; **170**(11):1415–1421, doi:10.1093/aje/kwp294. 12

[71] Thompson J, Minelli C, Abrams K, Tobin M, Riley R. Meta-analysis of genetic studies using Mendelian randomization – a multivariate approach. *Statistics in Medicine* 2005; **24**(14):2241–2254, doi:10.1002/sim.2100. 12, 29, 34, 39, 40, 67, 122, 181, 219, 221

[72] Kivimäki M, Lawlor D, Eklund C, Davey Smith G, Hurme M, Lehtimaki T, Viikari J, Raitakari O. 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* 2007; **27**(4):978–979, doi:10.1161/01.atv.0000258869.48076.14. 12

[73] Timpson N, Sayers A, Davey Smith G, Tobias J. How does body fat influence bone mass in childhood? A Mendelian randomization approach. *Journal of Bone and Mineral Research* 2009; **24**:522–533, doi:10.1359/jbmr.081109. 12

[74] Mumby H, Elks C, Li S, Sharp S, Khaw K, Luben R, Wareham N, Loos R, Ong K. Mendelian randomisation study of childhood BMI and early menarche. *Journal of Obesity* 2011; doi:10.1155/2011/180729. 12

[75] von Hinke Kessler Scholder S, Davey Smith G, Lawlor D, Propper C, Windmeijer F. Genetic markers as instrumental variables: An application to child fat mass and academic achievement. *The Centre for Market and Public Organisation 10/229*, Department of Economics, University of Bristol, UK 2010. 12, 29

[76] Lewis S, Ebrahim S, Davey Smith G. Meta-analysis of MTHFR 677C - T polymorphism and coronary heart disease: does totality of evidence support causal role for homocysteine and preventive potential of folate? *British Medical Journal* 2005; **331**(7524):1053, doi:10.1136/bmj.38611.658947.55. 12, 42, 122

[77] Almon R, Álvarez-Leon E, Engfeldt P, Serra-Majem L, Magnuson A, Nilsson T. Associations between lactase persistence and the metabolic syndrome in a cross-sectional study in the Canary Islands. *European Journal of Nutrition* 2010; **49**(3):141–146, doi:10.1007/s00394-009-0058-2. 12

[78] Bech B, Autrup H, Nohr E, Henriksen T, Olsen J. Stillbirth and slow metabolizers of caffeine: comparison by genotypes. *International Journal of Epidemiology* 2006; **35**(4):948–953, doi:10.1093/ije/dyl116. 12

[79] Irons D, McGue M, Iacono W, Oetting W. Mendelian randomization: A novel test of the gateway hypothesis and models of gene–environment interplay. *Development and Psychopathology* 2007; **19**(04):1181–1195, doi:10.1017/s0954579407000612. 12

[80] Ding W, Lehrer S, Rosenquist J, Audrain-McGovern J. The impact of poor health on academic performance: new evidence using genetic markers. *Journal of Health Economics* 2009; **28**(3):578–597, doi:10.1016/j.jhealeco.2008.11.006. 12

[81] Botto L, Yang Q. 5, 10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. *American Journal of Epidemiology* 2000; **151**(9):862–877. 12

[82] CRP CHD Genetics Collaboration. Collaborative pooled analysis of data on C-reactive protein gene variants and coronary disease: judging causality by Mendelian randomisation. *European Journal of Epidemiology* 2008; **23**(8):531–540, doi:10.1007/s10654-008-9249-z. 13, 15, 16, 46, 68, 106, 110, 123, 168, 180, 182, 184

[83] Emerging Risk Factors Collaboration. The Emerging Risk Factors Collaboration: analysis of individual data on lipid, inflammatory and other markers in over 1.1 million participants in 104 prospective studies of cardiovascular diseases. *European Journal of Epidemiology* 2007; **22**(12):839–869, doi:10.1007/s10654-007-9165-7. 13

[84] Emerging Risk Factors Collaboration. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *The Lancet* 2010; **375**(9709):132–140, doi:10.1016/S0140-6736(09)61717-7. 13, 17, 67, 106

[85] Casas J, Shah T, Cooper J, Hawe E, McMahon A, Gaffney D, Packard C, O'Reilly D, Juhan-Vague I, Yudkin J, Tremoli E, Margaglione M, Di Minno G, Hamsten A, Kooistra T, Stephens J, Hurel S, Livingstone S, Colhoun H, Miller G, Bautista L, Meade T, Sattar N, Humphries S, Hingorani A. Insight into the nature of the CRP–coronary event association using Mendelian randomization. *International Journal of Epidemiology* 2006; **35**(4):922–931, doi:10.1093/ije/dyl041. 13, 42, 97

[86] Casas J, Shah T, Hingorani A, Danesh J, Pepys M. C-reactive protein and coronary heart disease: a critical review. *Journal of Internal Medicine* 2008; **264**(4):295–314, doi:10.1111/j.1365-2796.2008.02015.x. 13

[87] Danesh J, Pepys M. C-reactive protein and coronary disease: is there a causal link? *Circulation* 2009; **120**(21):2036–2039, doi:10.1161/circulationha.109.907212. 13

[88] Keavney B. C reactive protein and the risk of cardiovascular disease. *British Medical Journal* 2011; **342**:d144, doi:10.1136/bmj.d144. 13

[89] Minelli C, Thompson J, Tobin M, Abrams K. An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. *American Journal of Epidemiology* 2004; **160**(5):445–452, doi:10.1093/aje/kwh228. 15, 30, 39, 122, 134, 152, 180

[90] Johnson A, Handsaker R, Pulit S, Nizzari M, O'Donnell C, de Bakker P. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 2008; **24**(24):2938–2939, doi:10.1093/bioinformatics/btn564. 15, 16

[91] Goetghebeur E. Commentary: To cause or not to cause confusion vs transparency with Mendelian randomization. *International Journal of Epidemiology* 2010; **39**(3):918, doi:10.1093/ije/dyq100. 17

[92] Stock J, Wright J, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 2002; **20**(4):518–529, doi:10.1198/073500102288618658. 19, 40, 41, 56, 106, 151

[93] Bowden R, Turkington D. *Instrumental variables*. Cambridge University Press, 1990. 22, 31, 130, 195

[94] Palmer T, Thompson J, Tobin M, Sheehan N, Burton P. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *International Journal of Epidemiology* 2008; **37**(5):1161–1168, doi:10.1093/ije/dyn080. 23, 30, 32, 77, 90, 100, 126, 142, 151

[95] Dawid A. Influence diagrams for causal modelling and inference. *International Statistical Review* 2002; **70**(2):161–189, doi:10.1111/j.1751-5823.2002.tb00354.x. 23

[96] Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**(5):688–701. 23, 27

[97] Clarke P, Windmeijer F. Instrumental variable estimators for binary outcomes. *The Centre for Market and Public Organisation 10/239*, Centre for Market and Public Organisation, University of Bristol, UK 2010. 24, 34, 35, 37, 126

[98] Pearl J. Causal diagrams for empirical research. *Biometrika* 1995; **82**(4):669–688, doi:10.1093/biomet/82.4.669. 24

[99] Didelez V, Meng S, Sheehan N. Assumptions of IV methods for observational epidemiology. *Statistical Science* 2010; **25**(1):22–40, doi:10.1214/09-sts316. 24, 29, 30, 32, 37, 97, 225

[100] Lawlor D, Harbord R, Timpson N, Lowe G, Rumley A, Gaunt T, Baker I, Yarnell J, Kivimäki M, Kumari M, Norman P, Jamrozik K, Hankey G, Almeida O, Flicker L, Warrington N, Marmot M, Ben-Shlomo Y, Palmer L, Day I, Ebrahim S, Davey Smith G. The association of C-reactive protein and CRP genotype with coronary heart disease: Findings from five studies with 4,610 cases amongst 18,637 participants. *PLoS ONE* 2008; **3**(8):e3011, doi:10.1371/journal.pone.0003011. 25, 30, 122, 130, 156

[101] Bound J, Jaeger D, Baker R. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 1995; **90**(430):443–450. 25, 30, 41, 46, 51, 52, 53, 59, 126, 130, 131

[102] Staiger D, Stock J. Instrumental variables regression with weak instruments. *Econometrica* 1997; **65**(3):557–586. 25, 41, 54, 55, 56, 126, 130, 184, 225

[103] Thanassoulis G, O'Donnell C. Mendelian randomization: nature's randomized trial in the post-genome era. *Journal of the American Medical Assocation* 2009; **301**(22):2386–2388, doi:10.1001/jama.2009.812. 25

[104] Verduijn M, Siegerink B, Jager K, Zoccali C, Dekker F. Mendelian randomization: use of genetics to enable causal inference in observational studies. *Nephrology Dialysis Transplantation* 2010; **25**(5):1394–1398, doi:10.1093/ndt/gfq098. 25, 156

[105] Ogbuanu I, Zhang H, Karmaus W. Can we apply the Mendelian randomization methodology without considering epigenetic effects? *Emerging Themes in Epidemiology* 2009; **6**(1):3, doi:10.1186/1742-7622-6-3. 26

[106] Spirtes P, Glymour C, Scheines R. *Causation, prediction, and search*. The MIT Press, 2000. 26

[107] Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1):41–55, doi:10.1093/biomet/70.1.41. 27

[108] Cox D. *Planning of experiments. Section 2: Some key assumptions*. Wiley New York, 1958. 27

[109] Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987; **125**(5):761–768. 27, 73, 79, 126

[110] Greenland S, Robins J, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**(1):29–46, doi:10.2307/2676645. 28, 75, 126, 141, 225

[111] Wald A. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics* 1940; **11**(3):284–300. 29, 97, 130

[112] Angrist J, Graddy K, Imbens G. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies* 2000; **67**(3):499–527, doi:10.1111/1467-937x.00141. 29, 30, 32

[113] Manski C. Nonparametric bounds on treatment effects. *The American Economic Review* 1990; **80**(2):319–323. 30

[114] Fieller E. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1954; **16**:175–185. 30, 130, 134

[115] Buonaccorsi J. *Encyclopedia of Biostatistics*, chapter Fieller's theorem. Wiley, 2005; 1951–1952. 30

[116] Thomas D, Lawlor D, Thompson J. Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. *Annals of Epidemiology* 2007; **17**(7):511–513, doi:10.1016/j.annepidem.2006.12.005. 31

[117] Baum C, Schaffer M, Stillman S. Instrumental variables and GMM: Estimation and testing. *Stata Journal* 2003; **3**(1):1–31. 31, 38, 39, 41, 42, 56, 60, 68, 133, 195, 219

[118] Angrist J, Pischke J. *Mostly harmless econometrics: an empiricist's companion. Chapter 4: Instrumental variables in action: sometimes you get what you need.* Princeton University Press, 2009. 31, 32, 33, 38, 54, 55, 73, 93, 131, 142, 212, 239

[119] Mikusheva A, Poi B. Tests and confidence sets with correct size when instruments are potentially weak. *Stata Journal* 2006; **6**(3):335–347. 31, 130

[120] Murphy KM, Topel RH. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics* 1985; **3**(4):370–379. 31

[121] Hardin J, Carroll R. Variance estimation for the instrumental variables approach to measurement error in generalized linear models. *Stata Journal* 2003; **3**(4):342–350. 31, 130

[122] Imbens G, Rosenbaum P. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2005; **168**(1):109–126, doi:10.1111/j.1467-985x.2004.00339.x. 31, 130, 151

[123] Hardin J, Schmiediche H, Carroll R. Instrumental variables, bootstrapping, and generalized linear models. *Stata Journal* 2003; **3**(4):351–360. 31, 42

[124] Kinal T. The existence of moments of k-class estimators. *Econometrica* 1980; **48**(1):241–249. 31, 56, 130

[125] Rassen J, Schneeweiss S, Glynn R, Mittleman M, Brookhart M. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *American Journal of Epidemiology* 2009; **169**(3):273–284, doi:10.1093/aje/kwn299. 32, 126

[126] Foster E. Instrumental variables for logistic regression: an illustration. *Social Science Research* 1997; **26**(4):487–504, doi:10.1006/ssre.1997.0606. 32, 34, 145, 225

[127] Terza J, Basu A, Rathouz P. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics* 2008; **27**(3):531–543, doi:10.1016/j.jhealeco.2007.09.009. 32, 91, 94, 195, 238

[128] Cai B, Small D, Ten Have T. Two-stage instrumental variable methods for estimating the causal odds ratio: Analysis of bias. *Statistics in Medicine* 2011; **30**(15):1809–1824, doi:10.1002/sim.4241. 32, 91, 94, 239

[129] Wooldridge J. *Econometric analysis of cross section and panel data*. The MIT Press, 2002. 32, 73, 142

[130] Ten Have T, Joffe M, Cary M. Causal logistic models for non-compliance under randomized treatment with univariate binary response. *Statistics in Medicine* 2003; **22**(8):1255–1283, doi:10.1002/sim.1401. 32

[131] Nagelkerke N, Fidler V, Bernsen R, Borgdorff M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine* 2000; **19**(14):1849–1864. 32, 90, 142

[132] Hayashi F. *Econometrics*. Princeton University Press, 2000. 33, 131

[133] Hahn J, Hausman J, Kuersteiner G. Estimation with weak instruments: accuracy of higher-order bias and MSE approximations. *Econometrics Journal* 2004; **7**(1):272–306, doi:10.1111/j.1368-423x.2004.00131.x. 33, 56, 59, 131

[134] Kleibergen F, Zivot E. Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics* 2003; **114**(1):29–72, doi:10.1016/S0304-4076(02)00219-1. 34, 121, 137

[135] Conley T, Hansen C, McCulloch R, Rossi P. A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics* 2008; **144**(1):276–305. 34, 121

[136] Imbens G, Rubin D. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics* 1997; **25**(1):305–327, doi:10.1214/aos/1034276631. 34

[137] Sims C. Bayesian methods in applied econometrics, or, why econometrics should always and everywhere be Bayesian. *Technical Report*, Princeton University 2007. Accessed at http://sims.princeton.edu/yftp/EmetSoc607/AppliedBayes.pdf on 1 June 2009. 34

[138] Palmer T, Thompson J, Tobin M. Meta-analysis of Mendelian randomization studies incorporating all three genotypes. *Statistics in Medicine* 2008; **27**(30):6570–6582, doi:10.1002/sim.3423. 34, 40, 122

[139] Verzilli C, Shah T, Casas J, Chapman J, Sandhu M, Debenham S, Boekholdt M, Khaw K, Wareham N, Judson R, Benjamin E, Kathiresan S, Larson M, Rong J, Sofat R, Humphries S, Smeeth L, Cavalleri G, Whittaker J, Hingorani A. Bayesian meta-analysis of genetic association studies with different sets of markers. *American Journal of Human Genetics* 2008; **82**(4):859–872, doi:10.1016/j.ajhg.2008.01.016. 34

[140] Burgess S, Thompson S, CRP CHD Genetics Collaboration. Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables. *Statistics in Medicine* 2010; **29**(12):1298–1311, doi:10.1002/sim.3843. 34, 68, 72, 127, 131, 143, 181

[141] McKeigue P, Campbell H, Wild S, Vitart V, Hayward C, Rudan I, Wright A, Wilson J. Bayesian methods for instrumental variable analysis with genetic instruments ("Mendelian randomization"): example with urate transporter SLC2A9 as instrumental variable for effect of urate levels on metabolic syndrome. *International Journal of Epidemiology* 2010; **39**(3):907–918, doi:10.1093/ije/dyp397. 34, 121, 131, 143

[142] Johnston K, Gustafson P, Levy A, Grootendorst P. 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* 2007; **27**(9):1539–1556, doi:10.1002/sim.3036. 34, 35, 145

[143] Hansen L. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1982; **50**(4):1029–1054. 35

[144] Greene W. *Econometric analysis*. 5th edition, Prentice-Hall, 2003. 35

[145] Robins J. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics – Theory and Methods* 1994; **23**(8):2379–2412. 35, 145

[146] Fischer-Lapp K, Goetghebeur E. Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials* 1999; **20**(6):531–546, doi:10.1016/S0197-2456(99)00027-6. 35, 145

[147] Pearl J. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000. 35

[148] Dunn G, Maracy M, Tomenson B. 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* 2005; **14**(4):369–395, doi:10.1191/0962280205sm403oa. 36

[149] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986; **7**(9-12):1393–1512, doi:10.1016/0270-0255(86)90088-6. 36, 145

[150] Greenland S, Lanes S, Jara M. Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. *Clinical Trials* 2008; **5**(1):5–13, doi:10.1177/1740774507087703. 36, 145

[151] Robins J. *Statistical models in epidemiology: the environment and clinical trials*, chapter Marginal structural models versus structural nested models as tools for causal inference. Springer, 1999; 95–134. 36

[152] Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2003; **65**(4):817–835, doi:10.1046/j.1369-7412.2003.00417.x. 36, 146

[153] Vansteelandt S, Bowden J, Babanezhad M, Goetghebeur E. On instrumental variables estimation of causal odds ratios. *Statistical Science* 2011; **26**(3):403–422, doi:10.1214/11-sts360. 36, 95, 97, 126, 145, 225

[154] Greenland S, Longnecker M. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology* 1992; **135**(11):1301–1309. 37

[155] Harbord R, Lawlor D. Genetically elevated C-reactive protein and vascular disease. *New England Journal of Medicine* 2009; **360**(9):935, doi:10.1056/nejmc082413. 37

[156] Shea J. Instrument relevance in multivariate linear models: A simple measure. *Review of Economics and Statistics* 1997; **79**(2):348–352, doi:10.1162/rest.1997.79.2.348. 38, 66

[157] Basmann R. On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association* 1960; **55**(292):650–659. 38

[158] Sargan J. The estimation of economic relationships using instrumental variables. *Econometrica* 1958; **26**(3):393–415. 38, 60, 106, 169

[159] Baum C, Schaffer M, Stillman S. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal* 2007; **7**(4):465–506. 40

[160] Nelson C, Startz R. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business* 1990; **63**(1):125–140. 41, 46, 51, 126, 130

[161] Stock J, Yogo M. Testing for weak instruments in linear IV regression. *SSRN eLibrary* 2002; **11**:T0284. 41, 69, 106, 126, 130, 170

[162] Donald S, Newey W. Choosing the number of instruments. *Econometrica* 2001; **69**(5):1161–1191, doi:10.1111/1468-0262.00238. 41

[163] Richardson D. The exact distribution of a structural coefficient estimator. *Journal of the American Statistical Association* 1968; **63**(324):1214–1226. 41, 46

[164] Sawa T. The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *Journal of the American Statistical Association* 1969; **64**(327):923–937. 41, 46

[165] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria 2011. URL `http://www.R-project.org`, ISBN 3-900051-07-0. 41

[166] StataCorp. *Stata Statistical Software: Release 11*. College Station, TX: StataCorp LP 2009. 41

[167] Fox J. Teacher's Corner: Structural Equation Modeling With the *sem* Package in R. *Structural Equation Modeling: A Multidisciplinary Journal* 2006; **13**(3):465–486. 42, 133

[168] Nichols A. IVPOIS: Stata module to estimate an instrumental variables Poisson regression via GMM. *Technical Report*, Boston College Department of Economics 2007. 42

[169] Drukker D. Generalized method of moments estimation in Stata 11. *Technical Report*, Stata Corp. 2009. 42

[170] Chaussé P. Computing generalized method of moments and generalized empirical likelihood with R. *Journal of Statistical Software* 2010; **34**(11):1–35. 42, 147

[171] Orsini N, Bellocco R, Greenland S. GLST: Stata module for trend estimation of summarized dose-response data. *Technical Report*, Karolinska Institutet 2005. 42

[172] Keavney B, Danesh J, Parish S, Palmer A, Clark S, Youngman L, Delepine M, Lathrop M, Peto R, Collins R, The International Studies of Infarct Survival (ISIS) Collaborators. Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'. *International Journal of Epidemiology* 2006; **35**(4):935–943, doi: 10.1093/ije/dyl114. 42, 122

[173] Bautista L, Smeeth L, Hingorani A, Casas J. Estimation of bias in nongenetic observational studies using "Mendelian triangulation". *Annals of Epidemiology* 2006; **16**(9):675–680, doi:10.1016/j.annepidem.2006.02.001. 43, 142

[174] Davey Smith G, Lawlor D, Harbord R, Timpson N, Rumley A, Lowe G, Day I, Ebrahim S. Association of C-reactive protein with blood pressure and hypertension: life course confounding and Mendelian randomization tests of causality. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2005; **25**(5):1051–1056, doi: 10.1161/01.atv.0000160351.95181.d0. 43, 122

[175] Zacho J, Tybjaerg-Hansen A, Jensen J, Grande P, Sillesen H, Nordestgaard B. Genetically elevated C-reactive protein and ischemic vascular disease. *New England Journal of Medicine* 2008; **359**(18):1897–1908, doi:10.1056/nejmoa0707402. 47, 157

[176] Fuller W. Some properties of a modification of the limited information estimator. *Econometrica* 1977; **45**(4):939–953. 56

[177] Hansen C, Hausman J, Newey W. Estimation with many instrumental variables. *Journal of Business and Economic Statistics* 2008; **26**(4):398–422, doi:10.1198/073500108000000024. 56

[178] Hall A, Rudebusch G, Wilcox D. Judging instrument relevance in instrumental variables estimation. *International Economic Review* 1996; **37**(2):283–298. 60, 64

[179] Hansson G. Inflammation, atherosclerosis, and coronary artery disease. *The New England Journal of Medicine* 2005; **352**(16):1685–1695, doi:10.1056/nejmra043430. 67

[180] Little R, Rubin D. *Statistical analysis with missing data (2nd edition)*. Wiley New York, 2002. 68, 157

[181] Cameron A, Trivedi P. *Microeconometrics: methods and applications*. Cambridge University Press, 2005. 70

[182] Baum C. *An introduction to modern econometrics using Stata*. Stata Corp, 2006. 71

[183] Small D, Rosenbaum P. War and wages: the strength of instrumental variables and their sensitivity to unobvserved biases. *Journal of the American Statistical Association* 2008; **103**(483):924–933, doi:10.1198/016214507000001247. 72, 231

[184] Bound J, Jaeger D, Baker R. The cure can be worse than the disease: A cautionary tale regarding instrumental variables. *Technical Report 137*, NBER 1993. 72

[185] Davey Smith G, Ebrahim S. Data dredging, bias, or confounding. *British Medical Journal* 2002; **325**(7378):1437, doi:10.1136/bmj.325.7378.1437. 73

[186] Zeger S, Liang K, Albert P. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**(4):1049–1060. 73, 79

[187] Hausman J. *Handbook of econometrics*, vol. 1, chapter Specification and estimation of simultaneous equation models. Elsevier, 1983; 391–448. See footnote 60. 73, 96

[188] Whittemore A. Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 1978; **40**(3):328–340. 75, 210

[189] Geng Z. Collapsibility of relative risk in contingency tables with a response variable. *Journal of the Royal Statistical Society: Series B (Methodological)* 1992; **54**(2):585–593. 75

[190] Ducharme G, LePage Y. Testing collapsibility in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 1986; **48**(2):197–205. 75

[191] Ford I, Norrie J, Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine* 1995; **14**(8):735–746, doi:10.1002/sim.4780140804. 75, 213

[192] Tan Z. Marginal and nested structural models using instrumental variables. *Journal of the American Statistical Association* 2010; **105**(489):157–169, doi:10.1198/jasa.2009.tm08299. 79

[193] Berntsen J, Espelid T, Genz A. An adaptive algorithm for the approximate calculation of multiple integrals. *ACM Transactions on Mathematical Software (TOMS)* 1991; **17**(4):437–451, doi:10.1145/210232.210233. 81, 141

[194] Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology* 2004; **160**(4):301–305, doi:10.1093/aje/kwh221. 83

[195] Ford I, Norrie J. The role of covariates in estimating treatment effects and risk in long-term clinical trials. *Statistics in Medicine* 2002; **21**(19):2899–2908, doi:10.1002/sim.1294. 91, 96, 211

[196] Stock J. Nonparametric policy analysis. *Journal of the American Statistical Association* 1989; **84**(406):567–575. 95

[197] Stukel T, Fisher E, Wennberg D, Alter D, Gottlieb D, Vermeulen M. Analysis of observational studies in the presence of treatment selection bias. *Journal of the American Medical Association* 2007; **297**(3):278, doi:10.1001/jama.297.3.278. 95, 96

[198] Robinson L, Jewell N. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 1991; **59**(2):227–240. 96

[199] Steyerberg E, Bossuyt P, Lee K. Clinical trials in acute myocardial infarction: Should we adjust for baseline characteristics? *American Heart Journal* 2000; **139**(5):745–751. 96

[200] Tybjarg-Hansen A, Steffensen R, Meinertz H, Schnohr P, Nordestgaard B. Association of mutations in the apolipoprotein B gene with hypercholesterolemia and the risk of ischemic heart disease. *The New England Journal of Medicine* 1998; **338**(22):1577–1584, doi:10.1056/nejm199805283382203. 97

[201] Carroll R, Ruppert D, Stefanski L, Crainiceanu C. *Measurement error in nonlinear models: a modern perspective*. CRC Press, 2006. 97

[202] Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**(4):325–337, doi:10.1023/A:1008929526011. 102

[203] Fried L, Borhani N, Enright P, Furberg C, Gardin J, Kronmal R, Kuller L, Manolio T, Mittelmark M, Newman A. The Cardiovascular Health Study: Design and rationale. *Annals of Epidemiology* 1991; **1**(3):263–276, doi:10.1016/1047-2797(91)90005-w. 106, 197

[204] Spiegelhalter D, Best N, Carlin B, Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **64**(4):583–639, doi:10.1111/1467-9868.00353. 112

[205] Angrist J, Krueger A. The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association* 1992; **87**(418):328–336. 120

[206] Burgess S, Thompson S. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine* 2011; **30**(11):1312–1323, doi:10.1002/sim.4197. 130, 131

[207] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS version 1.4 user manual. *Technical Report*, MRC Biostatistics Unit, Cambridge, UK 2003. 133, 158

[208] Bowden J, Vansteelandt S. Mendelian randomisation analysis of case-control data using structural mean models. *Statistics in Medicine* 2011; **30**(6):678–694, doi:10.1002/sim.4138. 145, 146, 152

[209] Zivot E, Startz R, Nelson C. Valid confidence intervals and inference in the presence of weak instruments. *International Economic Review* 1998; **39**(4):1119–1144. 151

[210] Mikusheva A. Robust confidence sets in the presence of weak instruments. *Journal of Econometrics* 2010; **157**(2):236–247, doi:10.1016/j.jeconom.2009.12.003. 151, 238

[211] Lunn D, Best N, Spiegelhalter D, Graham G, Neuenschwander B. Combining MCMC with 'sequential' PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics* 2009; **36**:19–38, doi:10.1007/s10928-008-9109-1. 152, 158

[212] Burgess S, Thompson S, CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology* 2011; **40**(3):755–764, doi:10.1093/ije/dyr036. 156, 181, 236

[213] Yau L, Little R. Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association* 2001; **96**(456):1232–1244. 157

[214] Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**(4):457–472, doi:10.1214/ss/1177011136. 158

[215] Browning S. Multilocus association mapping using variable-length Markov chains. *American Journal of Human Genetics* 2006; **78**(6):903–913. 158

[216] Browning S, Browning B. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* 2007; **81**(5):1084–1097, doi:10.1086/521987. 158, 161

[217] Meng X. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**(4):538–558, doi:10.1214/ss/1177010269. 159

[218] Lunn D, Whittaker J, Best N. A Bayesian toolkit for genetic association studies. *Genetic Epidemiology* 2006; **30**(3):231–247, doi:10.1002/gepi.20140. 159

[219] Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* 2010; **11**(7):499–511, doi:10.1038/nrg2796. 161

[220] White I. simsum: Analyses of simulation studies including Monte Carlo error. *Stata Journal* 2010; **10**(3):369–385. 165

[221] Willett W. An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. *Statistics in Medicine* 1989; **8**(9):1031–1040. 167

[222] Fu W, Wang Y, Wang Y, Li R, Lin R, Jin L. Missing call bias in high-throughput genotyping. *BMC Genomics* 2009; **10**(1):106, doi:10.1186/1471-2164-10-106. 172

[223] Little R. A note about models for selectivity bias. *Econometrica: Journal of the Econometric Society* 1985; **53**(6):1469–1474. 173

[224] Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* 2006; **78**(4):629–644, doi:10.1086/502802. 176

[225] Burgess S. WinBUGS code for imputation of missing data in Mendelian randomization studies. *Technical Report 2011/1*, MRC Biostatistics Unit 2011. Available from www.mrc-bsu.cam.ac.uk. 176

[226] Borenstein M, Hedges L, Higgins J, Rothstein H. *Introduction to meta-analysis. Chapter 34: Generality of the basic inverse-variance method*. Wiley, 2009. 180

[227] Higgins J, Thompson S, Deeks J, Altman D. Measuring inconsistency in meta-analyses. *British Medical Journal* 2003; **327**(7414):557–560, doi:10.1136/bmj.327.7414.557. 184

[228] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188, doi:10.1016/0197-2456(86)90046-2. 195

[229] Collett D. *Modelling survival data in medical research*. Chapman and Hall/CRC press, 2003. 199

[230] Collett D. *Modelling binary data*. Chapman and Hall/CRC press, 2003. 200

[231] Frost C, Thompson S. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2000; **163**(2):173–189, doi:10.1111/1467-985x.00164. 206

[232] Sutton A, Abrams K. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* 2001; **10**(4):277–303, doi:10.1177/096228020101000404. 210

[233] Breslow N, Day N. *The analysis of case-control studies. Chapter 7: Conditional logistic regression for matched sets*. International Agency for Research on Cancer, 1980. 210

[234] Bishop Y, Fienberg S, Paul W. *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, 1975. 210

[235] Armitage P. *Perspectives in probability and statistics*, chapter The use of the cross-ratio in aetiological surveys. Academic Press, New York-London, 1975; 349–355. 210

[236] Higgins J, Thompson S, Spiegelhalter D. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; **172**(1):137–159, doi:10.1111/j.1467-985x.2008.00552.x. 222

[237] Burgess S, Thompson S. Improvement of bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statistics in Medicine* 2011; Accepted manuscript. 225

[238] Robins J. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* 1992; **79**(2):311–334, doi:10.1093/biomet/79.2.321. 236

[239] Manolio T. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine* 2010; **363**(2):166–176. 236

[240] Davey Smith G. Random allocation in observational data: how small but robust effects could facilitate hypothesis-free causal inference. *Epidemiology* 2011; **22**(4):460–463, doi:10.1097/ede.0b013e31821d0426. 237

[241] Relton C, Davey Smith G. Epigenetic epidemiology of common complex disease: Prospects for prediction, prevention, and treatment. *PLoS Medicine* 2010; **7**(10):e1000 356, doi:10.1371/journal.pmed.1000356. 238

[242] Sheiner L, Rubin D. Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology & Therapeutics* 1995; **57**(1):6–15. 239

[243] Ridker P, Danielson E, Fonseca F, Genest J, Gotto A, Kastelein J, Koenig W, Libby P, Lorenzatti A, MacFadyen J, Nordestgaard B, Shepherd J, Willerson J, Glynn R, JUPITER Study Group. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *New England Journal of Medicine* 2008; **359**(21):2195–2207. 239