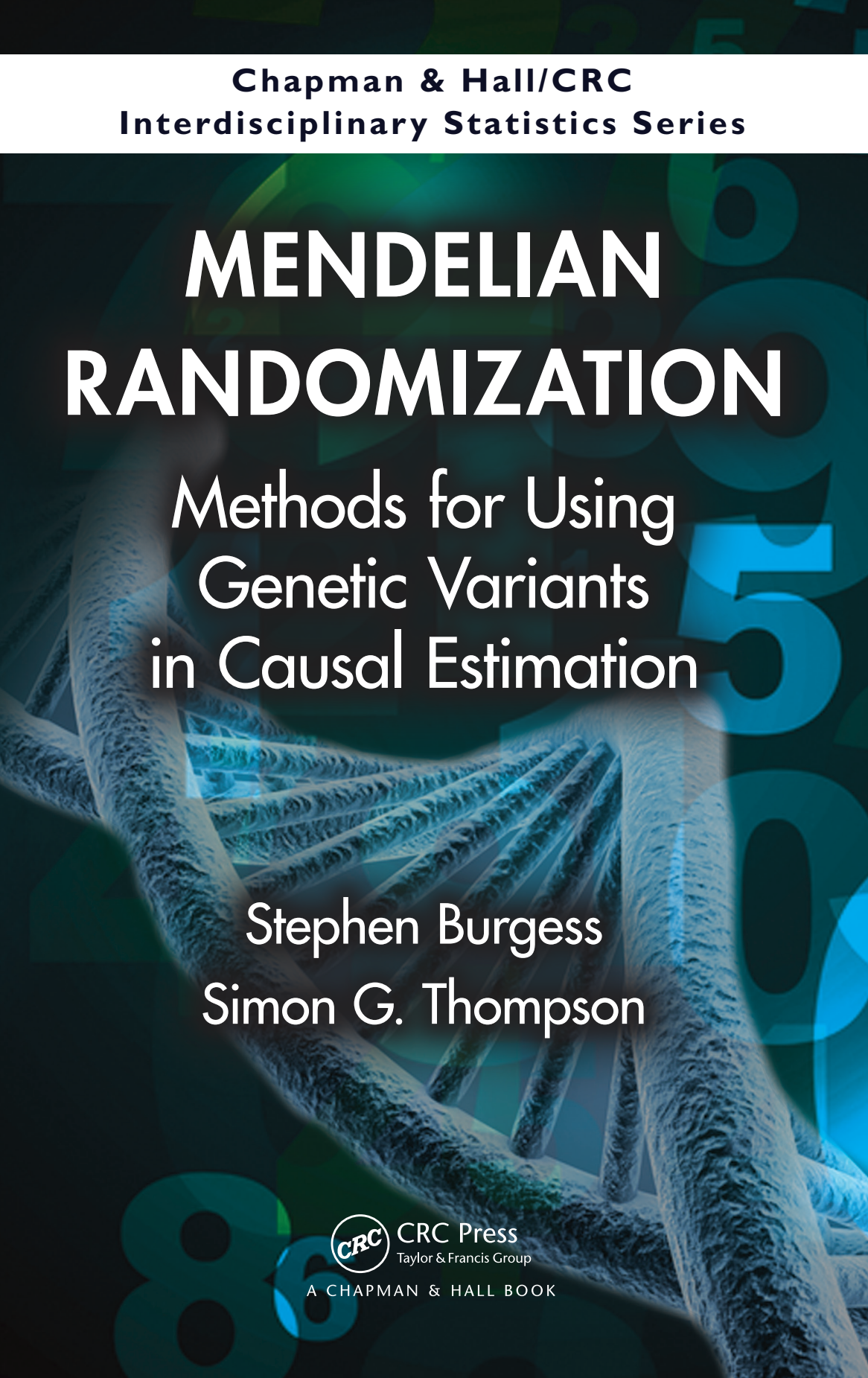
Chapman & Hall/CRC
Interdisciplinary Statistics Series

MENDELIAN RANDOMIZATION

Methods for Using Genetic Variants in Causal Estimation

Stephen Burgess
Simon G. Thompson



CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

MENDELIAN RANDOMIZATION

Methods for Using Genetic Variants in Causal Estimation

CHAPMAN & HALL/CRC

Interdisciplinary Statistics Series

Series editors: N. Keiding, B.J.T. Morgan, C.K. Wikle, P. van der Heijden

Published titles

AGE-PERIOD-COHORT ANALYSIS: NEW MODELS, METHODS, AND EMPIRICAL APPLICATIONS	Y. Yang and K. C. Land
ANALYSIS OF CAPTURE-RECAPTURE DATA	R. S. McCrea and B. J. T. Morgan
AN INVARIANT APPROACH TO STATISTICAL ANALYSIS OF SHAPES	S. Lele and J. Richtsmeier
ASTROSTATISTICS	G. Babu and E. Feigelson
BAYESIAN ANALYSIS FOR POPULATION ECOLOGY	R. King, B. J. T. Morgan, O. Gimenez, and S. P. Brooks
BAYESIAN DISEASE MAPPING: HIERARCHICAL MODELING IN SPATIAL EPIDEMIOLOGY, SECOND EDITION	A. B. Lawson
BIOEQUIVALENCE AND STATISTICS IN CLINICAL PHARMACOLOGY	S. Patterson and B. Jones
CLINICAL TRIALS IN ONCOLOGY, THIRD EDITION	S. Green, J. Benedetti, A. Smith, and J. Crowley
CLUSTER RANDOMISED TRIALS	R.J. Hayes and L.H. Moulton
CORRESPONDENCE ANALYSIS IN PRACTICE, SECOND EDITION	M. Greenacre
DESIGN AND ANALYSIS OF QUALITY OF LIFE STUDIES IN CLINICAL TRIALS, SECOND EDITION	D.L. Fairclough
DYNAMICAL SEARCH	L. Pronzato, H. Wynn, and A. Zhigljavsky
FLEXIBLE IMPUTATION OF MISSING DATA	S. van Buuren
GENERALIZED LATENT VARIABLE MODELING: MULTILEVEL, LONGITUDINAL, AND STRUCTURAL EQUATION MODELS	A. Skrondal and S. Rabe-Hesketh
GRAPHICAL ANALYSIS OF MULTI-RESPONSE DATA	K. Basford and J. Tukey
INTRODUCTION TO COMPUTATIONAL BIOLOGY: MAPS, SEQUENCES, AND GENOMES	M. Waterman

Published titles

MARKOV CHAIN MONTE CARLO IN PRACTICE	W. Gilks, S. Richardson, and D. Spiegelhalter
MEASUREMENT ERROR AND MISCLASSIFICATION IN STATISTICS AND EPIDEMIOLOGY: IMPACTS AND BAYESIAN ADJUSTMENTS	P. Gustafson
MEASUREMENT ERROR: MODELS, METHODS, AND APPLICATIONS	J. P. Buonaccorsi
MENDELIAN RANDOMIZATION: METHODS FOR USING GENETIC VARIANTS IN CAUSAL ESTIMATION	S. Burgess and S.G. Thompson
META-ANALYSIS OF BINARY DATA USING PROFILE LIKELIHOOD	D. Böhning, R. Kuhnert, and S. Rattanasiri
STATISTICAL ANALYSIS OF GENE EXPRESSION MICROARRAY DATA	T. Speed
STATISTICAL AND COMPUTATIONAL PHARMACOGENOMICS	R. Wu and M. Lin
STATISTICS IN MUSICOLOGY	J. Beran
STATISTICS OF MEDICAL IMAGING	T. Lei
STATISTICAL CONCEPTS AND APPLICATIONS IN CLINICAL MEDICINE	J. Aitchison, J.W. Kay, and I.J. Lauder
STATISTICAL AND PROBABILISTIC METHODS IN ACTUARIAL SCIENCE	P.J. Boland
STATISTICAL DETECTION AND SURVEILLANCE OF GEOGRAPHIC CLUSTERS	P. Rogerson and I. Yamada
STATISTICS FOR ENVIRONMENTAL BIOLOGY AND TOXICOLOGY	A. Bailer and W. Piegorsch
STATISTICS FOR FISSION TRACK ANALYSIS	R.F. Galbraith
VISUALIZING DATA PATTERNS WITH MICROMAPS	D.B. Carr and L.W. Pickle

Chapman & Hall/CRC
Interdisciplinary Statistics Series

MENDELIAN RANDOMIZATION

Methods for Using Genetic Variants in Causal Estimation

Stephen Burgess

Department of Public Health and Primary Care
University of Cambridge, UK

Simon G. Thompson

Department of Public Health and Primary Care
University of Cambridge, UK

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20141229

International Standard Book Number-13: 978-1-4665-7318-5 (eBook - PDF)

Contents

Preface

The quantity of research into the genetics of common diseases has exploded over the last 20 years. While many genetic variants related to various diseases have been identified, their usefulness may lie more in what they offer to our understanding of the biological mechanisms leading to disease rather than to, for example, predicting disease risk. To understand mechanisms, we need to separate the relationships of risk factors with diseases into those that are causal and those that are not. This is where Mendelian randomization can play an important role.

The technique of Mendelian randomization itself has undergone rapid development, mostly in the last 10 years, and applications now abound in current medical and epidemiological journals. Its basis is that of instrumental variable analysis, which has a much longer history in statistics and particularly in econometrics. Relevant papers on Mendelian randomization are therefore dispersed across the multiple fields of genetics, epidemiology, statistics and econometrics. The intention of this book is to bring together this literature on the methods and practicalities of Mendelian randomization, especially to help those who are relatively new to this area.

In writing this book, we envisage the target audience comprising two main groups, Epidemiologists and Medical Statisticians, who want to perform applied Mendelian randomization analyses or understand how to interpret their results. We therefore assume a familiarity with basic epidemiological terminology, such as prospective and case-control studies, and basic statistical methods, such as ordinary least squares and logistic regression. Meanwhile, we have tried to make the perhaps alien terminology of econometrics accessible to our intended readership.

While we hope that this book will be accessible to a wide audience, a geneticist may baulk at the simplistic explanations of Mendelian inheritance, a statistician may yearn for a deeper level of technical exposition, and an epidemiologist may wonder why we don't just cut to the chase of how to perform the analyses. Our hope is that enough detail is given for those who need it, references are available for those who want more, and a section can simply be glossed over by those for whom it is redundant.

While we have included relevant statistical methodology available up to the publication date of the book, our focus has been on methods and issues which are of practical relevance for applied Mendelian randomization analyses, rather than those which are of more theoretical interest, or 'cutting-edge' developments which may not stand the test of time. As such, to a research

statistician, the book will provide a background to current areas of methodological debate, but it will generally not offer opinions on controversial topics which are likely to become out-of-date quickly as further investigations are performed. Where possible, sections with technical content in the first part of the book are marked with asterisks (*), and are written in such a way that they can be omitted without interrupting the flow of the book.

A website to complement this book, as well as the authors' ongoing research on this topic, is available at www.mendelianrandomization.com. This contains chapter summaries, paper summaries, web-based applications, and software code for implementing some of the statistical techniques discussed in the book.

We would like to express our thanks to all those who commented on chapters of this book, whether in chapter or book form. We thank Frank Dudbridge, Brandon Pierce, Dylan Small, Maria Glymour, Stephen Sharp, Mary Schooling, Tom Palmer, George Davey Smith, Debbie Lawlor, John Thompson, Jack Bowden, Shaun Seaman, Lucas Tittmann, Daniel Freitag, Peter Willeit, Edmund Jones, Angela Wood and Adam Butterworth. Further individuals commented as anonymous referees, and so we cannot thank them by name. We also thank Rob Calver, our editor, for being knowledgeable, supportive, and open to our ideas. We are also grateful to the principal investigators of the studies in the CRP CHD Genetics Collaboration who have allowed us to use their data in this book, as well as to the study participants for giving their time and consent to participate in this research.

In short, while we realize that we will not be able to please all of our readers all of the time, we hope that this book will enable a wide range of people to better understand what is an important, but complex and multidisciplinary, area of research.

Stephen Burgess, Simon G. Thompson

University of Cambridge, UK

August 2014

Abbreviations

2SLS	two-stage least squares
ACE	average causal effect
BMI	body mass index
CCGC	CRP CHD Genetics Collaboration
CRP	C-reactive protein
CHD	coronary heart disease
CI	confidence interval
COR	causal odds ratio
CRR	causal risk ratio
DIC	deviance information criterion
DNA	deoxyribonucleic acid
FIML	full information maximum likelihood
FTO	a gene associated with obesity
GMM	generalized method of moments
GWAS	genome-wide association study (or studies)
HDL-C	high-density lipoprotein cholesterol
IL6	interleukin-6
IPD	individual participant data
IV	instrumental variable
LIML	limited information maximum likelihood
LIVAE	linear IV average effect
LD	linkage disequilibrium
LDL-C	low-density lipoprotein cholesterol
lp(a)	lipoprotein(a)
MAR	missing at random
MCMC	Monte Carlo Markov chain
MI	myocardial infarction
OLS	ordinary least squares
OR	odds ratio
RCT	randomized controlled trial
RNA	ribonucleic acid
SE	standard error
SMM	structural mean model
SNP	single nucleotide polymorphism
SUTVA	stable unit treatment value assumption

Abbreviations for the various studies in the CCGC are given in Table 10.1.

Notation

Throughout this book, we use the notation:

X	exposure: the risk factor (or protective factor, or intermediate phenotype) of interest
Y	outcome
U	(sufficient) confounder of the X–Y association
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
ρ	correlation parameter between X and Y
ρ_{GX}	correlation parameter between G and X
σ^2	variance parameter
τ^2	between-study heterogeneity variance parameter
F	F statistic from regression of X on G
i	subscript indexing individuals
j	subscript indexing genetic subgroups
J	total number of genetic subgroups
k	subscript indexing genetic variants (SNPs)
K	total number of genetic variants
m	subscript indexing studies in a meta-analysis
M	total number of studies
N	total number of individuals
n	total number of cases (individuals with a disease event)
$\mathcal{N}$	normal distribution
$\mathcal{N}_2$	bivariate normal distribution

We follow the usual convention of using upper-case letters for random variables and lower-case letters for data values with X, Y, U, and G.

Part I

Using genetic variants as instrumental variables to assess causal relationships

1

Introduction and motivation

This book concerns making inferences about causal effects based on observational data using genetic instrumental variables, a concept known as Mendelian randomization. In this chapter, we introduce the basic idea of Mendelian randomization, giving examples of when the approach can be used and why it may be useful. We aim in this chapter only to give a flavour of the approach; details about its conditions and requirements are reserved for later chapters. Although the examples given in this book are mainly in the context of epidemiology, Mendelian randomization can address questions in a variety of fields of study, and the majority of the material in this book is equally relevant to problems in different research areas.

1.1 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 important medical questions, such as to determine disease aetiology (what is the cause of a disease?), to assess the impact of a medical or public health intervention (what would be the result of a treatment?), to inform public policy, to prioritize healthcare resources, to advise clinical practice, or to counsel on the impact of lifestyle choices, then we have to answer questions of cause and effect. The optimal way to address these questions is by appropriate study design, such as the use of prospective randomized trials.

1.1.1 Randomized trials and observational studies

The 'gold standard' for the empirical testing of a scientific hypothesis in clinical research is a randomized controlled trial. This design involves the allocation of different treatment regimes at random to experimental units (usually individuals) in a population. In its simplest form, one 'active treatment' (for example, intervention on a risk factor) is compared against a 'control treat-

ment' (no intervention), and the average outcomes in each of the arms of the trial are contrasted. Here the risk factor (which we will often refer to as the "exposure" variable) is a putative causal risk factor. We seek to assess whether the risk factor is a cause of the outcome, and estimate (if appropriate) the magnitude of the causal effect.

While randomized trials are in principle the best way of determining the causal status of a particular risk factor, they have some limitations. Randomized trials are expensive and time-consuming, especially when the outcome is rare or requires a long follow-up period to be observed. Additionally, in some cases, a targeted treatment which has an effect only on the risk factor of interest may not be available. Moreover, many risk factors cannot be randomly allocated for practical or ethical reasons. For example, in assessing the impact of drinking red wine on the risk of coronary heart disease, it would not be feasible to recruit participants to be randomly assigned to either drink or abstain from red wine over, say, a 20-year period. Alternative approaches for judging causal relationships are required.

Scientific hypotheses are often assessed using observational data. Rather than by intervening on the risk factor, individuals with high and low levels of the risk factor are compared. In many cases, differences between the average outcomes in the two groups have been interpreted as evidence for the causal role of the risk factor. However, such a conclusion confuses correlation with causation. There are many reasons why individuals with elevated levels of the risk factor may have greater average outcome levels, without the risk factor being a causal agent.

Interpreting an association between an exposure and a disease outcome in observational data as a causal relationship relies on untestable and usually implausible assumptions, such as the absence of unmeasured confounding (see Chapter 2) and of reverse causation. This has led to several high-profile cases where a risk factor has been widely promoted as an important factor in disease prevention based on observational data, only to be later discredited when evidence from randomized trials did not support a causal interpretation [Taubes and Mann, 1995]. For example, observational studies reported a strong inverse association between vitamin C and risk of coronary heart disease, which did not attenuate on adjustment for a variety of risk factors [Khaw et al., 2001]. However, results of experimental data obtained from randomized trials showed a non-significant association in the opposite direction [Collins et al., 2002]. The confidence interval for the observational association did not include the randomized trial estimate [Davey Smith and Ebrahim, 2003]. Similar stories apply to the observational and experimental associations between β-carotene and smoking-related cancers [Peto et al., 1981; Hennekens et al., 1996], and between vitamin E and coronary heart disease [Hooper et al., 2001]. More worrying is the history of hormone-replacement therapy, which was previously advocated as being 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 [Rossouw et al., 2002; Beral

et al., 2003]. More robust approaches are therefore needed for assessing causal relationships using observational data. Mendelian randomization is one such approach.

1.2 The rise of genetic epidemiology

Genetic epidemiology is the study of the role of genetic factors in health and disease for populations. We sketch the history and development of genetic epidemiology, indicating why it is an important area of epidemiological and scientific research.

1.2.1 Historical background

Although the inheritance of characteristics from one generation to the next has been observed for millennia, the mechanism for inheritance was long unknown. When Charles Darwin proposed his theory of evolution in 1859, one of its major problems was the lack of an underlying mechanism for heredity [Darwin, 1871]. Gregor Mendel in 1866 proposed two laws of inheritance: the law of segregation, that when any individual produces gametes (sex cells), the two 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 [cell division]' [Mendel, 1866]. These laws are summarized by the term "Mendelian inheritance", and it is this which gives Mendelian randomization its name [Davey Smith and Ebrahim, 2003]. 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 [Fisher, 1918]. A specific connection 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 [Pauling et al., 1949]. 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 [Watson and Crick, 1953]. The Human Genome Project was established in 1990, leading to the publication of the entirety of the human genetic code by the early 2000s [Roberts et al., 2001; McPherson et al., 2001]. 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 [Shendure and Ji, 2008].

1.2.2 Genetics and disease

As the knowledge of the human genome has developed, the search for genetic determinants of disease has expanded from monogenic disorders (disorders which are due to a single mutated gene, such as sickle-cell anaemia), 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 coronary heart disease, tend to cluster within families, but also depend on modifiable risk factors, such as diet and blood pressure. Several genetic factors have been found which relate to these diseases, especially through the increased use of genome-wide association studies (GWAS), in which the associations of thousands or even millions of genetic variants with a disease outcome are tested. In some cases, these discoveries have added to the scientific understanding of disease processes and the ability to predict disease risk for individuals. Nevertheless, they are of limited immediate interest from a clinical perspective, as an individual's genome cannot generally be changed. However, genetic discoveries provide opportunities for Mendelian randomization: a technique for using genetic data to assess and estimate causal effects of modifiable (non-genetic) risk factors based on observational data.

1.3 Motivating example: The inflammation hypothesis

We introduce the approach of Mendelian randomization using an example. The 'inflammation hypothesis' is an important question in the understanding of cardiovascular disease. Inflammation is one of the body's response mechanisms to a harmful stimulus. It is characterized by redness, swelling, heat, pain and loss of function in the affected body area. Cases can be divided into acute inflammation, which refers to the initial response of the body, and chronic inflammation, which refers to more prolonged changes. Examples of conditions classified as inflammation include appendicitis, chilblains, and arthritis.

Cardiovascular disease is a term covering a range of diseases including coronary heart disease (in particular myocardial infarction or a 'heart attack') and stroke. It is currently the biggest cause of death worldwide. The inflammation hypothesis states that there is some aspect of the inflammation response mechanism which leads to cardiovascular disease events, and that intervening on this pathway will reduce the risk of cardiovascular disease.

1.3.1 C-reactive protein and coronary heart disease

As part of the inflammation process, several chemicals are produced by the body, known as (positive) acute-phase proteins. These represent the body's

first line of defence against infection and injury. There has been particular interest in one of these, C-reactive protein (CRP), and the role of elevated levels of CRP in the risk of coronary heart disease (CHD). It is known that CRP is observationally associated with the risk of CHD [Kaptoge et al., 2010], but, prior to robust Mendelian randomization studies, it was not known whether this association was causal [Danesh and Pepys, 2009]. The specific question in this example (a small part of the wider inflammation hypothesis) is whether long-term elevated levels of CRP lead to greater risk of CHD.

1.3.2 Alternative explanations for association

In our example, there are many factors that increase both levels of CRP and the risk of CHD. These factors, known as confounders, may be measured and accounted for by statistical analysis, for instance multivariable regression. However, it is not possible to know whether all such factors have been identified. Also, CRP levels increase in response to sub-clinical disease, giving the possibility that the observed association is due to reverse causation.

One of the potential confounders of particular interest is fibrinogen, a soluble blood plasma glycoprotein, which enables blood-clotting. It is also part of the inflammation pathway. Although CRP is observationally positively associated with CHD risk, this association was shown to reduce on adjustment for various conventional risk factors (such as age, sex, body mass index, and diabetes status), and to attenuate to near null on further adjustment for fibrinogen [Kaptoge et al., 2010]. It is important to assess whether elevated levels of CRP are causally related to changes in fibrinogen, since if so conditioning the CRP–CHD association on fibrinogen would represent an over-adjustment, which would attenuate a true causal effect.

1.3.3 Instrumental variables

To address the problems of confounding and reverse causation in conventional epidemiology, we introduce the concept of an instrumental variable. An instrumental variable is a measurable quantity (a variable) which is associated with the exposure of interest, but not associated with any other competing risk factor that is a confounder. Neither is it associated with the outcome, except potentially via the hypothesized causal pathway through the exposure of interest. A potential example of an instrumental variable for health outcomes is geographic location. We imagine that two neighbouring regions have different policies on how to treat patients, and assume that patients who live on one side of the border are similar in all respects to those on the other side of the border, except that they receive different treatment regimes. By comparing these groups of patients, geographic location acts like the random allocation to treatment assignment in a randomized controlled trial, influencing the exposure of interest without being associated with competing risk factors. It therefore is an instrumental variable, and gives rise to a natural experiment

in the population, from which causal inferences can be obtained. Other plausible non-genetic instrumental variables include government policy changes (for example, the introduction of a smoking ban in public places, or an increase in cigarette tax, which might decrease cigarette smoking prevalence without changing other variables) and physician prescribing preference (for example, the treatment a doctor chose to prescribe to the previous patient, which will be representative of the doctor's preferred treatment, but should not be affected by the current patient's personal characteristics or case history).

1.3.4 Genetic variants as instrumental variables

A genetic variant is a section of genetic code that differs between individuals. In Mendelian randomization, genetic variants are used as instrumental variables. Individuals in a population can be divided into subgroups based on their genetic variants. On the assumption that the genetic variants are 'randomly' distributed in the population, that is independently of environmental and other variables, then these genetic subgroups do not systematically differ with respect to any of these variables. Additionally, as the genetic code for each individual is determined before birth, there is no way that a variable measured in a mature individual can be a 'cause' of a genetic variant. Returning to our example, if we can find a suitable genetic variant (or variants) associated with CRP levels, then we can compare the genetically-defined subgroup of individuals with lower average levels of CRP to the subgroup with higher average levels of CRP. In effect, we are exploiting a natural experiment in the population, whereby nature has randomly given some individuals a genetic 'treatment' which increases their CRP levels. If individuals with a genetic variant, which is associated with elevated average levels of CRP and satisfies the instrumental variable assumptions, exhibit greater incidence of CHD, then we can conclude that CRP is a causal risk factor for CHD, and that lowering CRP is likely to lead to reductions in CHD rates. Under further assumptions about the statistical model for the relationship between CRP and CHD risk, a causal parameter can be estimated. Although Mendelian randomization uses genetic variants to answer inferential questions, these are not questions about genetics, but rather about modifiable risk factors, such as CRP, and their causal effect on outcomes (usually disease outcomes).

1.3.5 Violations of instrumental variable assumptions

It is impossible to test whether there is a causal relationship between two variables on the basis of observational data alone. All empirical methods for making causal claims by necessity rely on untestable assumptions. Instrumental variable methods are no exception. Taking the example of Section 1.3.3, if geographic location is associated with other factors, such as socioeconomic status, then the assumption that the distribution of the outcome would be the same for both populations under each policy regime would be violated.

Or if the genetic variant(s) associated with CRP levels used in a Mendelian randomization analysis were also independently associated with, say, blood pressure, the comparison of genetic subgroups would not be a valid test of the causal effect of CRP on CHD risk. The validity of the instrumental variable assumptions is crucial to the interpretation of a Mendelian randomization investigation, and is discussed at length in later chapters.

1.3.6 The CRP CHD Genetics Collaboration

The statistical methods and issues discussed in this book are illustrated using the example of the causal relationships of CRP on the outcomes CHD risk and fibrinogen. Data are taken from the CRP CHD Genetic Collaboration (CCGC), a consortium of 47 studies comprising cohort, case-control and nested case-control studies [CCGC, 2008]. Most of these studies recorded data on CRP levels, on incident CHD events (or history of CHD events in retrospective or cross-sectional studies), and on up to 20 genetic variants associated with CRP levels. Of these, we will focus on four, which were pre-specified as the variants to be used as instrumental variables in the main applied analysis from the collaboration and are located in and around the *CRP* gene region on chromosome 1. Some studies did not measure all four of these variants; others did not measure CRP levels in some or all participants. Several studies measured a range of additional covariates, including fibrinogen, many of which are potential confounders in the association between CRP and CHD risk. A full analysis of the data from the CCGC for the causal effect of CRP on CHD risk is given in Chapter 10. While the aim of the book is not to prove or disprove the causal role of CRP for CHD, the epidemiological implications of the analyses are explored.

1.4 Other examples of Mendelian randomization

Although the initial applications of Mendelian randomization were in the field of epidemiology [Youngman et al., 2000], the use of genetic instrumental variables is becoming widespread in a number of different fields. A systematic review of applied Mendelian randomization studies was published in 2010 [Bochud and Rousson, 2010]. A list of the exposures and outcomes of some causal relationships which have been assessed using Mendelian randomization is given in Table 1.1. The list includes examples from the fields of epidemiology, nutrition, sociology, psychology, and economics: the only limitation in the use of Mendelian randomization to assess the causal effect of an exposure on an outcome is the availability of a suitable genetic variant to use as the instrumental variable.

The reasons to use Mendelian randomization outside of epidemiology are similar to those in epidemiology. In many fields, randomized experiments are difficult to perform and instrumental variable techniques represent one of the few ways of assessing causal relationships in the absence of complete knowledge of confounders. Although the language and context of this book will generally be that of epidemiology, much applies equally to other areas of research. More detailed expositions of some examples of applied Mendelian randomization analyses are given in Chapter 5.

1.5 Overview of book

Although there has been much research into the use of instrumental variables in econometrics and epidemiology since they were first proposed [Wright, 1928], several barriers existed in applying this to the context of Mendelian randomization. These include differences in terminology, where the same concept is referred to in various disciplines by different names, and differences in theoretical concepts, particularly relating to the definition and interpretation of causal relationships. Additionally, several methodological issues have been posed by the use of genetic variants as instrumental variables that had not previously been considered in the instrumental variables literature, and required (and still require) methodological development. A major motivation in writing this book is to provide an accessible resource to those coming from different academic disciplines to understand issues relevant to the use of genetic variants as instrumental variables, and particularly for those wanting to undertake and interpret Mendelian randomization analyses.

1.5.1 Structure

This book is divided into three parts. The first part, comprising Chapters 1 to 6, is entitled "Using genetic variants as instrumental variables to assess causal relationships". This part contains the essential information for a practitioner interested in Mendelian randomization (Chapters 1 and 2), including definitions of causal relationships and instrumental variables (Chapter 3), and methods for the estimation of causal effects (Chapter 4). With the exception of some of the technical details about statistical methods marked as 'starred', these sections should be fully accessible to most epidemiologists. Issues surrounding the application of Mendelian randomization in practice are explored by presenting examples of Mendelian randomization investigations from the literature (Chapter 5). Also addressed is the question of how to interpret a Mendelian randomization estimate, and how it may compare to the effect of an intervention on the exposure of interest in practice (Chapter 6).

The second part, comprising Chapters 7 to 10, is entitled "Statistical issues with instrumental variable analysis in Mendelian randomization". This

Nature of exposure	Exposure	Outcome	Reference
Biomarker	apolipoprotein E	cancer	[1]
	CRP	insulin resistance	[2]
	CRP	CIMT	[3]
	CRP	cancer	[4]
	folate	blood pressure	[5]
	HDL-C	myocardial infarction	[6]
	homocysteine	stroke	[7]
	lipoprotein(a)	myocardial infarction	[8–9]
	SHBG	CHD	[10]
Physical characteristic	BMI	CIMT	[11]
	BMI	early menarche	[12]
	BMI	labour market outcomes	[13]
	fat mass	academic achievement	[14]
Dietary factor	alcohol intake	blood pressure	[15]
	caffeine intake	stillbirth	[16]
	milk intake	metabolic syndrome	[17]
Pathological behaviour	alcohol abuse	drug abuse	[18]
	ADHD	education	[19]
	depression	education	[19]
Inter-generational effect	interuterine folate	neural tube defects	[20]

TABLE 1.1
Examples of causal relationships assessed by Mendelian randomization in applied research.

Abbreviations:
CRP = C-reactive protein, CIMT = carotid intima-media thickness, CHD = coronary heart disease, SHBG = sex-hormone binging globulin, HDL-C = high-density lipoprotein cholesterol, BMI = body mass index, ADHD = attention deficit hyperactivity disorder.

References:
1. Trompet et al., 2009,
2. Timpson et al., 2005,
3. Kivimäki et al., 2008,
4. Allin et al., 2010,
5. Thompson et al., 2005,
6. Voight et al., 2012,
7. Casas et al., 2005,
8. Kamstrup et al., 2009,
9. Clarke et al., 2009,
10. Ding et al., 2009a,
11. Kivimäki et al., 2007,
12. Mumby et al., 2011,
13. Norton and Han, 2008,
14. Von Hinke et al., 2010,
15. Chen et al., 2008,
16. Bech et al., 2006,
17. Almon et al., 2010,
18. Irons et al., 2007,
19. Ding et al., 2009b,
20. Ebrahim and Davey Smith, 2008

consists of comparisons of methods for using instrumental variables to estimate a causal effect, and matters concerning the behaviour of instrumental variable estimates, such as potential biases. In particular, we consider the issue of weak instrument bias (Chapter 7), and the problems of estimating a single causal effect using data on multiple instrumental variables (Chapter 8) and data from multiple studies (Chapter 9). Estimates from instrumental variable methods typically have wide confidence intervals, often necessitating the synthesis of evidence from multiple sources to obtain an estimate precise enough to be clinically relevant. As part of the discussion on the use of multiple instruments, we address questions relating to the power and sample size requirements of Mendelian randomization studies. This part of the book is illustrated throughout using data from the CCGC, and a comprehensive analysis of the CCGC dataset for the causal effect of CRP on CHD risk is provided (Chapter 10). Although the details in this part require a greater depth of mathematical understanding, each chapter is introduced using non-technical language, and concludes with a set of key points to convey the main messages of the chapter.

Finally, we conclude with the final part, Chapter 11, by discussing possible future directions for research involving Mendelian randomization.

1.6 Summary

Distinguishing between a factor which is merely associated with an outcome and one which has a causal effect on the outcome is problematic outside of the context of a randomized controlled trial. Instrumental variables provide a way of assessing causal relationships in observational data, and Mendelian randomization is the use of genetic variants as instrumental variables.

In the next chapter, we provide more detail of what Mendelian randomization is, and when and why it may be useful.

2

What is Mendelian randomization?

In this chapter, we illustrate the conceptual framework and motivation for Mendelian randomization, explaining how Mendelian randomization offers opportunities to address some of the problems of conventional epidemiology. We describe the specific characteristics of genetic data which give rise to the Mendelian randomization approach.

2.1 What is Mendelian randomization?

Mendelian randomization is the use of genetic variants in non-experimental data to make causal inferences about the effect of an exposure on an outcome. We use the word "exposure" throughout this book to refer to the putative causal risk factor, sometimes called an intermediate phenotype, which can be a biomarker, an anthropometric measure, or any other risk factor that may affect the outcome. Usually the outcome is disease, although there is no methodological restriction as to what outcomes can be considered. Non-experimental data encompass all observational studies, including cross-sectional and longitudinal, cohort and case-control designs – any study where there is no intervention applied by the researcher.

2.1.1 Motivation

A foundational aim of epidemiological research is the estimation of the effect of changing an exposure on an outcome. This is known as the causal effect of the exposure on the outcome, and typically differs from the observational association between the exposure and outcome, for example due to confounding. Correlation between the exposure and the outcome cannot be reliably interpreted as evidence of a causal relationship. For example, those who drink red wine regularly have a lower incidence of heart disease. But socio-economic status is a common predictor of both wine consumption and better coronary health, and so it may be that socio-economic status rather than wine consumption underlies the risk of heart disease. Observational associations may also arise as a result of reverse causation. 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. Another example is vitamin D levels, which may decrease in individuals who are ill and therefore do not go outside, rather than vitamin D being a cause of illness.

The idea of Mendelian randomization is to find a genetic variant (or variants) associated with the exposure, but not associated with any other risk factor which affects the outcome, and not directly associated with the outcome. This means that any association of the genetic variant with the outcome must come via the variant's association with the exposure, and therefore implies a causal effect of the exposure on the outcome. Such a genetic variant would satisfy the assumptions of an instrumental variable (IV) [Greenland, 2000a; Sussman and Hayward, 2010]. As the theory of IVs was initially developed in the field of econometrics, a number of terms commonly used in the IV literature derive from this field and are not always well understood by medical statisticians or epidemiologists. Table 2.1 is a glossary of terms which are used in each field.

2.1.2 Instrumental variables

A technical definition of Mendelian randomization is "instrumental variable analysis using genetic instruments" [Wehby et al., 2008]. In Mendelian randomization, genetic variant(s) are used as IVs for assessing the causal effect of the exposure on the outcome [Thomas and Conti, 2004].

The fundamental conditions for a genetic variant to satisfy to be an IV are summarized as:

i. the variant is associated with the exposure,

ii. the variant is not associated with any confounder of the exposure–outcome association,

iii. the variant does not affect the outcome, except possibly via its association with the exposure.

Although Mendelian randomization analyses often involve a single genetic variant, multiple variants can be used either as separate IVs or combined into a single IV. More detail on the IV assumptions, which are key to the validity of Mendelian randomization investigations, is given in Chapter 3.

2.1.3 Confounding and endogeneity

One of the reasons why there may be a correlation between the exposure and outcome in an observational study is confounding, or the related concept, endogeneity of the exposure.

Confounding is defined as the presence of inherent differences between groups with different levels of the exposure [Greenland and Robins, 1986]. It

Econometrics term	Epidemiological term	Notes
Endogenous / endogeneity Exogenous / exogeneity	Confounded / confounding Unconfounded / no confounding	A variable is confounded / endogenous in a regression model if it is correlated with the error term, meaning that the regression coefficient is a biased estimate of the causal effect. A variable is unconfounded / exogenous if it is not correlated with the error term (see Section 2.1.3).
Outcome	Outcome	Denoted Y in this text.
Endogenous regressor	Exposure	Denoted X in this text; the causal effect of X on Y cannot be estimated by simple regression of Y on X if there is unmeasured confounding.
Instrumental variable / excluded instrument	Instrumental variable / instrument	Denoted G in this text; the instrument is called 'excluded' because 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 that is included in a model, such as a multivariable 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 an estimate of the causal effect.
Concentrate out	Profile out	To exclude a nuisance parameter from an equation by forming a profile likelihood with its maximum likelihood estimate given the other variables.
Panel data	Longitudinal data	Data on items at multiple timepoints.

TABLE 2.1
A summary of instrumental variable terms used in the fields of econometrics and epidemiology.

is often considered to result from the distribution of particular variables in the population, known as confounders. A confounder is a variable which is a common cause of both the exposure and the outcome. When confounders are recognized, measured and adjusted for, for example by multivariable regression, the remaining association between the exposure and outcome will often still be a biased estimate of the causal effect, due to the existence of unknown or unmeasured confounders or imprecision in measured confounders. Confounding not adjusted for in an analysis is termed 'residual confounding'.

Endogeneity means that there is a correlation between the regressor and the error term in a regression model. The words 'exogenous' and 'endogenous' are rarely used in epidemiology (see Table 2.1), but the terms have rigorous definitions that are useful in understanding confounding. Endogeneity literally means "coming from within". The opposite of endogenous is exogenous; an exogenous variable "comes from outside" of the regression model. The term endogeneity encompasses confounding, but also includes phenomena that are traditionally thought of as separate from confounding, such as measurement error and reverse causation. If the exposure in a model is endogenous in a regression model, then the regression coefficient for the exposure will be biased for the causal effect. An IV can be understood as 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 [Martens et al., 2006].

Mendelian randomization has also been named 'Mendelian deconfounding' [Tobin et al., 2004] as it aims to give estimates of the causal effect free from biases due to 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. While we can measure individual confounders and adjust for them in our analysis, we can never be certain that all confounders have been identified or measured precisely, leading to residual confounding. Additionally, if we adjust for a variable that lies on the true causal pathway between the exposure of interest and outcome (a mediator), this represents an over-adjustment and attenuates the estimate of the causal effect [Christenfeld et al., 2004]. By finding a genetic variant which satisfies the IV assumptions, we can estimate the unconfounded association between the exposure and the outcome.

2.1.4 Analogy with a randomized controlled trial

Mendelian randomization is analogous to a randomized controlled trial (RCT) [Nitsch et al., 2006]. An RCT, considered to provide the "gold standard" of medical evidence, involves dividing a set of individuals into two or more subgroups in a random way. These subgroups are each given different treatments. Randomization is preferred over any other assignment to subgroups as all

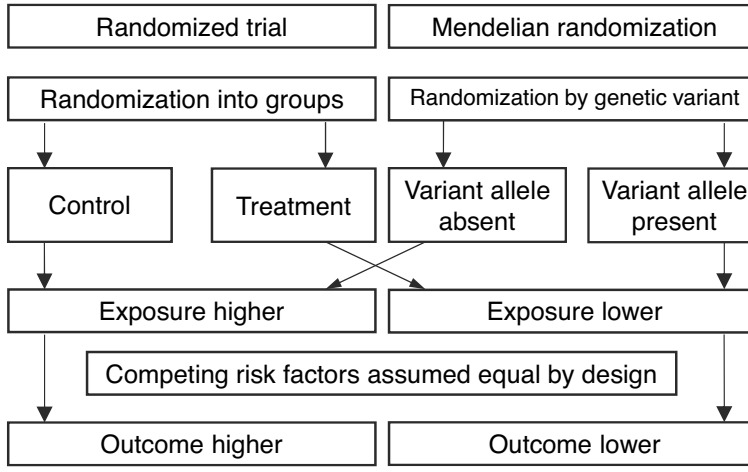


FIGURE 2.1
Comparison of a randomized controlled trial and Mendelian randomization.

possible confounders, known and unknown, are on average balanced between the subgroups.

In Mendelian randomization, we use a genetic variant to form subgroups analogous to those in an RCT, as shown in Figure 2.1. From the IV assumptions (Section 2.1.2), these subgroups differ systematically in the exposure, but not in any other factor except for those causally 'downstream' of the exposure. A difference in outcomes between these subgroups would therefore indicate a causal effect of the exposure on the outcome [Hernán and Robins, 2006]. Inferring a causal effect of the exposure on the outcome from an association between the genetic variant and the outcome is analogous to inferring an intention-to-treat effect from an association between randomization and the outcome in an RCT (that is, assignment to the treatment group affects the outcome).

Genetic variants for an individual are inherited from their parents, and so are not randomly assigned. For example, if neither of an individual's parents carry a particular genetic mutation, there is no way that the individual will carry that mutation. Nonetheless, under fairly realistic conditions the distribution of genetic variants in the population can be thought of as random with respect to environmental and social factors which may be important confounders. The necessary assumptions for a variant to be randomly distributed are random mating and lack of selection effects relating to the variant of interest. While there will be some departures from these assumptions, studies have shown that the distribution of most genetic variants is fairly uniform across the population, at least for example in a Western European context [Davey Smith, 2011]. Considerable departures from the random mating assumptions which

may invalidate the use of a genetic variant can be assessed by performing a test of Hardy–Weinberg equilibrium, to see if the frequency of heterozygotes and homozygotes (see Section 2.3) in the population is in line with what is expected. A variable which is distributed as if being randomly assigned despite the lack of true randomness in the assignment is known as quasi-randomized. Most natural experiments rely on quasi-randomization rather than the strict randomization of experimental units.

A recent observational study showed that linear regression gave a p-value less than 0.01 in 45% of 4560 associations between all pairs of 96 non-genetic variables [Davey Smith et al., 2007]. This suggests that many observed associations between environmental variables may not have a true causal interpretation. In contrast, the proportion of associations between genetic variants and these 96 variables with p-values less than 0.01 was not significantly higher than would be expected by chance. This gives plausibility to the assumption that genetic variants used as IVs will be distributed independently from many potential confounders, and so in many cases assignment to a genetic subgroup can be regarded as analogous to randomization in an RCT.

However, Mendelian randomization differs from a randomized trial in another respect. The aim of Mendelian randomization is not to estimate the size of a genetic effect, but the causal effect of the exposure on the outcome. The average change in the outcome associated with a genetic variant may differ in magnitude from that resulting from an intervention in the exposure (see Chapter 6). Additionally, even if the association of a genetic variant with the outcome is small in magnitude, the population attributable risk of the exposure is not necessarily low, as the exposure may vary to a considerably larger extent than that which can be explained by the variant. It may be possible to change the exposure by a greater amount than the difference in the mean exposure between genetic subgroups. For example, the effect of statin drug use on low-density lipoprotein cholesterol levels is several times larger than the association of low-density lipoprotein cholesterol levels with variants in the *HMGCR* gene, and consequently the effect on subsequent outcomes is greater.

2.2 Why use Mendelian randomization?

Although the main reason to use Mendelian randomization is to avoid the problem of residual confounding, there are additional reasons for using Mendelian randomization in specific contexts: with case-control data and with exposures that are difficult to measure.

2.2.1 Reverse causation and case-control studies

Reverse causation occurs when an association between the exposure and the outcome is not due to the exposure causing a change in the outcome, but the outcome causing a change in the exposure. This could happen if the exposure increased in response to pre-clinical disease, for example from cancer before it becomes clinically apparent or from atherosclerosis prior to clinical manifestations of coronary heart disease. As the genotype of an individual is determined at conception and cannot be changed, there is no possibility of reverse causation being responsible for an association between genotype and disease.

For this reason, Mendelian randomization has great strengths in a retrospective setting where genetic variants are measured after the disease outcome, such as in a case-control study. Many exposures of interest cannot be reliably measured in cases, that is in individuals who have already experienced an outcome event, as the event may distort the measurement. In this case, the genetic variant can be used as a proxy for the exposure, and the genetic association with the outcome can be assessed retrospectively. As the genotype of an individual can be measured in diseased individuals, causal inferences can be obtained using Mendelian randomization in a case-control setting.

2.2.2 Exposures that are expensive or difficult to measure

Mendelian randomization can be a useful technique when the exposure of interest is expensive or difficult to measure. For example, gold standard assays for biomarkers such as water-soluble vitamins may cost too much to be affordable for a large sample, or measurement of fasting blood glucose, which requires overnight fasting, may be impractical. If the genetic variant is associated with the exposure (this can be verified in a subsample or a separate dataset) and is a valid IV for the exposure, a causal relationship between the exposure and outcome can be inferred from an association between the genetic variant and the outcome even in the absence of measurement of the exposure.

Additionally, instrumental variable estimates do not attenuate due to classical measurement error (including within-individual variation) in the exposure [Pierce and VanderWeele, 2012]. This contrasts with observational studies, in which measurement error in the exposure usually leads to the attenuation of regression coefficients towards the null (known as regression dilution bias) [Frost and Thompson, 2000].

A further example is where the risk factor is not only difficult to measure, but also difficult to define. For example, a variant in the *IL6R* gene region that is associated with serum interleukin-6 concentrations (as well as levels of downstream inflammatory markers, including C-reactive protein and fibrinogen) was shown to be associated with coronary heart disease (CHD) risk [Swerdlow et al., 2012]. However, from knowledge about the functional role of the variant, the causal effect assessed is not thought to operate through

elevated serum interleukin-6 concentrations, but rather through changes in signalling in interleukin-6 receptor pathways. This is a cellular phenotype which varies over time, and so a representative measurement for an individual is not straightforward to define. However, as the genetic variant can be measured, the causal role of interleukin-6 receptor-related pathways on CHD risk can be assessed by Mendelian randomization [Sarwar et al., 2012].

2.3 A brief overview of genetics

In order to understand Mendelian randomization, it is necessary to have at least a cursory understanding of genetics. We here provide a brief overview of genetics, only covering the information necessary to understand Mendelian randomization. A glossary of genetic terminology, adapted from a Mendelian randomization review paper [Lawlor et al., 2008] is provided in Table 2.2. Further information on genetic terminology related to Mendelian randomization can be found in other papers [Davey Smith and Ebrahim, 2003; Sheehan et al., 2008].

2.3.1 Reading the genetic code

The genetic information (or genome) of many living organisms consists of long strings of genetic code in the form of DNA (deoxyribonucleic acid), the molecule that encodes life, packaged up into chromosomes. Humans have 23 pairs of chromosomes, with one chromosome in each of the pairs coming from the mother and one from the father. Chromosomes contain genes, which are locatable regions of the genetic code that encode a unit of heritable information. Not all of the genetic sequence falls into a gene region, and much of the chromosome consists of intermediate genetic material known as noncoding DNA.

A single chromosome has two strands, each consisting of a sequence of nucleotide bases which can represented by letters. There are four possible nucleotide bases (adenine, thymine, cytoside and guanine) represented by the letters A, T, C and G. These nucleotide bases pair up in such a way that the strands contain complementary sequences. Wherever the first strand has A, the other will have T – and vice versa. Wherever the first strand has C, the other will have G – and vice versa. In this way, each of the strands contains the same information, and so only one of the strands is considered. Suppose that a chromosome at a given locus (position) in the DNA sequence on one of its strands reads:

- *Alleles* are the variant forms of a single nucleotide polymorphism (SNP). For a *diallelic SNP* where there are two possible alleles, the more common allele is called the *major allele* or *wildtype allele*, and the less common allele is the *minor allele* or *variant allele*.
- *Canalization* (also known as *developmental compensation*) is the process by which potentially disruptive influences on normal development from genetic (and environmental) variation are damped or buffered by compensatory processes.
- A *chromosome* carries a collection of genes located on a long string of DNA. Humans have 22 pairs of autosomal (non-sex) chromosomes and 1 pair of sex chromosomes.
- A *copy number variant* (or *variation*) is a (possibly) repeating section of DNA where the number of copies of the section varies between individuals.
- *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 labelled A, T, C and G.
- A *gene* is a section of a chromosome comprising DNA which encodes information relevant to the function of an organism.
- The *genotype* of an individual at a particular locus refers to the two alleles at that locus. If the alleles are the same, the genotype is *homozygous*; if different, *heterozygous*.
- A *haplotype* describes a 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* (plural: *loci*) is the position in a DNA sequence and can be a SNP, a region of DNA sequence, or a whole gene.
- *Meiosis* is the process of cell division leading to gametes (sex cells) which contain half of the genetic material from the original cell.
- *Pleiotropy* is the potential for genes or genetic variants to have more than one independent phenotypic effect.
- *Polymorphism* is the existence of two or more variants at a locus. The term polymorphism is usually restricted to moderately common genetic variants, with at least two alleles having frequencies of greater than 1% in the population. A less common variant allele is called a mutation.
- *Single nucleotide polymorphisms* (SNPs) are genetic variations in which one base in the DNA is altered, for example a T instead of an A.

TABLE 2.2
A glossary of genetic terminology, adapted from Lawlor et al., 2008.

```
...ATTACGCTTCCGAGCTTCCGCAG...
```

and that same locus on the paired chromosome reads:

```
...ATTACGCCTCCGAGCTTCCGCAG...
```

The underlined nucleotide represents a nucleotide at a particular locus that is polymorphic: it exists in various forms. All individuals contain many genetic mutations, where the DNA code has changed from that generally seen in the population. A single nucleotide polymorphism (SNP) is a mutation where a single nucleotide base at a particular locus has been replaced with a different nucleotide. The different possible nucleotides which may appear at each locus are known as alleles. For example, at the highlighted locus above, one chromosome has the letter T, and the other has the letter C: so T and C are alleles of this particular SNP. If these are the only two possibilities, this is a diallelic SNP; triallelic and quadrallelic SNPs are far less common, but have also been observed.

For a diallelic SNP, it is conventional to denote the more common allele, known as the wildtype or major allele, by an upper case letter (for example, A) and the less common allele, the variant or minor allele, by a lower case letter (for example, a). The choice of letter is arbitrary; there is no connection between the letter A commonly used for the first variant considered, and the nucleotide base adenine represented by letter A. The proportion of minor alleles in a population for a given SNP is called the 'minor allele frequency'. Although some genetic mutations seem to be specific to particular individuals, others are more widespread, showing up in a substantial proportion of the population. SNPs occur on average about once in every 300 nucleotides along the genome, and extensive catalogues of SNPs have been compiled.

As people have two copies of each chromosome (one from each parent), individuals can be categorized for each diallelic SNP into three possible subgroups corresponding to their combination of alleles (their genotype). These subgroups are the 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 a more complicated genetic variant, 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.

When multiple SNPs on a single chromosome are considered, the combination of alleles on each of the chromosomes is known as a haplotype. For example, if an individual has one chromosome reading:

```
...GCACCTTAC...GTAGAATC...TCAACTGTCAT
```

and the other reading:

```
...GCACCGTAC...GTAAAATC...TCAACTGTCAT
```

then the individual is a heterozygote for the first two SNPs, and a homozygote for the final SNP. The haplotypes are `TGT` and `GAT`. One of these haplotypes

is inherited from each of the individual's parents. As a haplotype is a series of alleles on the same chromosome, haplotype patterns, especially for SNPs that are physically close together, are often inherited together. This means that genetic variants are not always independently distributed. Using patterns which have been observed in a large number of individuals, haplotypes can sometimes be inferred from SNP data using computer software, as generally not all possible combinations of alleles will be present on a chromosome in a population. In some cases, haplotypes can be determined uniquely from SNP data, whereas in other cases, there is uncertainty in this determination. If the SNPs satisfy the IV assumptions, then the haplotypes will also satisfy the IV assumptions.

Other patterns of genetic variation can also be used as IVs, such as copy number variations where a section of genetic material is repeated a variable number of times. Generally, throughout this book we shall assume that IVs are diallelic SNPs, although the majority of methods and findings discussed will apply similarly in other cases. SNPs are given numbers by which they can be uniquely referenced. Reference numbers begin "rs" (standing for "reference SNP"), such as rs1205.

2.3.2 Using a genetic variant as an instrumental variable

The use of any particular genetic variant as an IV requires caution as the IV assumptions cannot be fully tested and may be violated for various epidemiological and biological reasons (see Chapter 3). 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 IV to assess the causal relationship between alcohol consumption and oesophageal cancer. Here, alcohol consumption is the exposure and oesophageal cancer the outcome.

Assessment of the causal relationship using classical epidemiological studies is hindered by the strong association between alcohol and tobacco smoking, another risk factor for oesophageal cancer [Davey Smith and Ebrahim, 2004]. Individuals with two copies of the *ALDH2* polymorphism 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 [Lewis and Davey Smith, 2005]. 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 mutated allele 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. This is an example of a gene–environment interaction (here between the genotype and alcohol consumption). The conclusion is that alcohol consumption causes oesophageal cancer, since there is no association between this genetic variant and many other risk factors, and any single risk factor would have to have a massive

Genetic subgroup	Effect on alcohol metabolism	Genetic association with oesophageal cancer
Major homozygotes	No effect – can metabolize alcohol	(Reference group)
Heterozygotes	Mild effect – individuals can drink, but alcohol stays in bloodstream for longer	Increased disease risk
Minor homozygotes	Severe effect – cannot metabolize alcohol, individuals tend to abstain from alcohol	Decreased disease risk

TABLE 2.3
Example: alcohol intake and the *ALDH2* polymorphism in the Japanese population.

effect on oesophageal cancer risk as well as a strong association with the genetic variant to provide an alternative explanation for these results.

These associations are summarized in Table 2.3. 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 of the outcome. The evidence for a causal link between alcohol consumption, exposure to acetaldehyde and oesophageal cancer is compelling [Schatzkin et al., 2009]. However, in other cases, particularly if the genetic variant(s) do not explain much of the variation in the exposure, the power to detect a causal effect may be insufficient to provide such a convincing conclusion.

2.4 Summary

Mendelian randomization has the potential to be a useful tool in a range of scientific contexts to investigate claims of causal relationships. It must be applied with care, as its causal claims come at the price of assumptions which are not empirically testable. Its methods must be refined, as often data on multiple genetic variants or data taken from several study populations are required to achieve meaningful findings. But, when properly used, it gives an insight into the underlying causal relationships between variables which few other approaches can rival.

3

Assumptions for causal inference

In the previous chapters, we repeatedly used the word 'causal' to describe the inferences obtained by Mendelian randomization. In this chapter, we clarify what is meant by the causal effect of an exposure on an outcome. We give a more detailed explanation of the theory of instrumental variables, and explain in biological terms various situations that may lead to violations of the instrumental variable assumptions and thus misleading causal inferences. We conclude by discussing the difference between testing for the presence of a causal relationship and estimating a causal effect, and the additional assumptions necessary for causal effect estimation.

3.1 Observational and causal relationships

As the saying goes, "association is not causation" or in its more widely quoted form "correlation does not imply causation". Naive interpretation of an observed relationship between two variables as causal is a well-known logical fallacy. However, precise definitions of causality which correspond to our intuitive understanding have eluded philosophers for centuries [Pearl, 2000a]. Definitions are also complicated by the fact that, in many epidemiological contexts, causation is probabilistic rather than deterministic: for example, smoking does not always lead to lung cancer.

3.1.1 Causation as the result of manipulation

The fundamental concept in thinking about causal relationships is the idea of intervention on, or manipulation of, a variable. This is often cited as "no causation without manipulation", reflecting that direct experimentation is necessary to demonstrate a causal effect [Holland, 1986]. A causal effect is present if the outcome is different when the exposure is set to two different levels. This differs from an observational association, which represents the difference in the outcome when the exposure is observed at two different levels. If there are variables which are correlated with the exposure, the observational association reflects differences not only in the exposure of interest, but also in the

variables correlated with the exposure. With the causal effect, setting the value of the exposure only alters the exposure and variables on the causal pathway downstream of the exposure, not variables on alternative causal pathways.

The outcome variable Y for different observed values x of the exposure X is written as $Y|X = x$, read as Y conditional on X equalling x. Causal effects cannot be expressed in terms of probability distributions and so additional notation is required [Pearl, 2010]. The outcome variable Y when the exposure X is set to a given value x is written as $Y|do(X = x)$, where the *do* operator indicates that the variable is manipulated to be set to the given value.

3.1.2 Causation as a counterfactual contrast

One common definition of a causal effect is that of a counterfactual contrast [Maldonado and Greenland, 2002]. Counterfactual, literally meaning counter or contrary to fact, refers to a potential situation which could have happened, but did not [Greenland, 2000b]. For example, in the morning, Adam has a headache. He may or may not take an aspirin tablet. At the point of decision, we can conceive that there are two potential universes where Adam makes different choices about whether to take the aspirin or not. Associated with each universe is a potential outcome – does he still have a headache that afternoon? Once he has made this decision, one of these universes and outcomes becomes counterfactual; both outcomes cannot be observed. A causal effect is present if the two outcomes are different; if he still had a headache in the universe where he did not take the aspirin, but did not have a headache in the universe where he did take the aspirin, then the aspirin has caused the alleviation of the headache. With a probabilistic interpretation, assuming that the outcome is stochastic rather than deterministic, if the probability that he would still have a headache is lower in the aspirin universe than in the no-aspirin universe, then taking aspirin has a causal effect on alleviating the headache.

There are several conceptual difficulties with the counterfactual approach [Dawid, 2000]. The main difficulty is that the causal effect of an exposure for an individual can never be measured, as at least one of the two outcomes in the causal contrast is always unobserved. This is referred to as the 'fundamental problem of causal inference' [Holland, 1986]. It means that a counterfactual causal estimate is not the answer to any real experiment that could be conducted, but the answer to a hypothetical experiment requiring two parallel universes. However, the counterfactual approach has many appealing features. Chiefly, it gives a precise framework for defining causal effects, aiding both informal and mathematical thinking about causal relationships.

In terms of notation, the potential outcomes $Y|do(X = x)$ that the outcome variable can take are written as $Y(x)$. If the exposure is binary, the two potential outcomes for an individual are $Y(1)$ and $Y(0)$, and the causal effect of increasing X from $X = 0$ to $X = 1$ is $Y(1) - Y(0)$.

3.1.3 Causation using graphical models

Graphical models, and in particular directed acyclic graphs, can provide a helpful way of thinking about and expressing causal relationships. A graphical model comprises a set of nodes representing variables, and arrows representing causal effects. An arrow from variable A to variable B indicates that there is a causal effect of A on B. A graphical model need not contain all intermediate variables (such as C if $A \rightarrow C \rightarrow B$), but must contain all common causes of variables included in the graph (such as D if $A \leftarrow D \rightarrow B$). Relations between variables are expressed by directed arrows, indicating a (direct) causal effect (conditional dependence), or without an arrow, indicating no direct effect (conditional independence). A direct causal effect is only 'direct' with respect to the variables included in the graph and as such is not direct in an absolute sense, but could act via an intermediate variable. A directed acyclic graph (DAG) is a graph that does not contain any complete cycles, such as $A \leftrightarrow B$ or $A \rightarrow B \rightarrow C \rightarrow A$; a cycle would imply that a variable is its own cause.

As an example, Figure 3.1 shows the instrumental variable (IV) assumptions (Section 2.1.2) in the form of a graph. To simplify the graph, all confounding variables are subsumed into a single 'confounder', which has effects on both the exposure and outcome. We see that there are arrows from the IV to the exposure (assumption i.), from the exposure to the outcome, and from the confounder to the exposure and to the outcome. Just as importantly, there is no pathway between the IV and the confounder (assumption ii.) and no pathway from the IV to the outcome apart from that passing through the exposure (assumption iii.), indicating that a hypothetical intervention to change the value of the IV without varying the exposure or the confounder would not affect the outcome.

A pathway does not necessarily mean a route consisting only of directed arrows. For there to be no pathway from the IV G to the outcome Y (except via the exposure), there cannot be a sequence consisting of chains ($G \rightarrow C \rightarrow Y$) or forks ($G \leftarrow D \rightarrow Y$) of variables not including the exposure. There can be inverted forks ($G \rightarrow E \leftarrow Y$), provided neither E nor a descendent of E is adjusted for in the analysis (E may be referred to as a collider). In these examples, C may represent a competing risk factor to the exposure, and D may represent a selection variable, such as ethnicity, which must be accounted for in the analysis to prevent bias due to population stratification (Section 3.2.5). Formally, the genetic variants and outcome must be d-separated by the risk factor and confounders [Geiger et al., 1990].

3.1.4 Causation based on multivariable adjustment

Multivariable adjustment is often undertaken in the analysis of observational data in order to try to account for confounding. A set of covariates which, if known and conditioned on, would give an estimate of association equal to the causal effect, is referred to as 'sufficient'. Assuming that a set of covariates

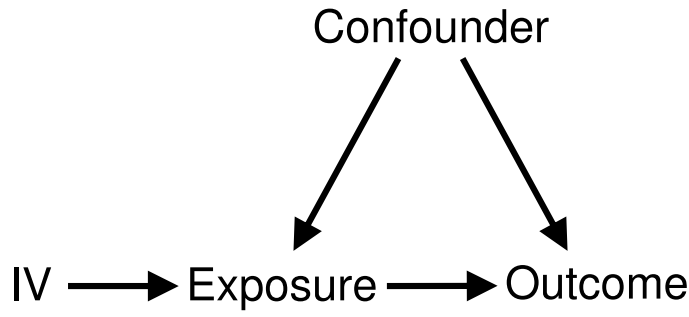


FIGURE 3.1
Directed acyclic graph illustrating instrumental variable (IV) assumptions.

is sufficient is necessary to interpret the result of a multivariable-adjusted regression analysis as a causal effect. On conditioning for a sufficient set of covariates, the counterfactual outcomes at different values of the exposure should be independent of the exposure, a property known as "conditional exchangeability" [Greenland and Robins, 1986].

If the causal relationships between all the variables in a model representing the generating mechanism for observational data were known, a set of covariates can be assessed as sufficient or otherwise using the "back-door criterion" [Pearl, 2000b]. For simple causal networks, a set of covariates is sufficient if it includes all common causes of the exposure and the outcome and does not include variables on the causal pathway from the exposure to the outcome, nor common effects of the exposure and outcome. In practice, neither the underlying network of associations between variables nor the sets of all common causes and all common effects of exposure and outcome are known, and so the use of multivariable-adjusted regression analyses to assess causal effects is unreliable. It is not possible to know if adjustment for a sufficient set of covariates has been made, or if there is residual confounding due to unmeasured covariates, or if the set of covariates includes variables on the causal pathway between the exposure and outcome, whose inclusion in a regression model also biases regression coefficients. This highlights the need to consider other methods for assessing causal relationships.

3.2 Finding a valid instrumental variable

Instrumental variable (IV) techniques represent one of the few ways available for estimating causal effects without complete knowledge of all confounders

of the exposure–outcome association. We continue by recalling and discussing the properties of an IV, and how the IV assumptions may be violated in practice.

3.2.1 Instrumental variable assumptions

In order for a genetic variant to be used to estimate a causal effect, it must satisfy the assumptions of an instrumental variable (Section 2.1.2), which we repeat here:

i. the variant is associated with the exposure,

ii. the variant is not associated with any confounder of the exposure–outcome association,

iii. the variant does not affect the outcome, except possibly via its association with the exposure.

These conditions can be understood intuitively. The first assumption guarantees that genetic subgroups defined by the variant will have different average levels of the exposure. This ensures that there is a systematic difference between the subgroups. If the genetic variant is not strongly associated with the exposure (in the sense of its statistical strength of association), then it is referred to as a weak instrument (see Chapter 7). A weak instrument differs from an invalid instrument in that a weak instrument can be made stronger by collecting more data. If a single genetic variant is a weak instrument, then it will still give a valid test of the null hypothesis of no causal effect, but the power to detect a true causal effect may be low. However, combining multiple weak instruments in an analysis model to obtain a single effect estimate can lead to misleading inferences.

The second assumption can be understood as ensuring that the comparison between the genetic subgroups is a fair test, that is, all other variables are distributed equally between the subgroups. The third assumption is often expressed using the concept of conditional independence as "the genetic variant is not associated with the outcome conditional on the value of the exposure and confounders of the exposure–outcome association". It ensures that the only causal pathway(s) from the genetic variant to the outcome are via the exposure. This means that the genetic variant is not directly associated with the outcome, nor is there any alternative pathway by which the variant is associated with the outcome other than that through the exposure.

3.2.2 Validity of the IV assumptions

The counterfactual framework for causation helps understanding of when and why a randomized controlled trial (RCT) can estimate a causal effect – thought of in a counterfactual sense as a contrast between parallel universes. The randomized subgroups in an RCT can be regarded as exchangeable. This means

that the same distribution of outcomes would be expected if each of the subgroups were exposed to the treatment or the control regime. Although an individual can only be exposed to one of the two treatment regimes (and so only observed in one universe), by exposing each subgroup to a different treatment regime, in effect we observe the population in each of the two counterfactual parallel universes, and the average outcomes in each of the universes (subgroups) can be compared [Greenland and Robins, 1986]. A causal effect can be consistently estimated which represents the average effect of being assigned to the treatment group as opposed to the control group. This means that an RCT can estimate an average causal effect for the population as the contrast between the average levels of the outcome in the randomized subgroups of the population (which will have the same characteristics on average as the overall population due to the random assignment into subgroups). An individual causal effect cannot be estimated, as an individual cannot in general be subjected to both the treatment and control regimes [Rubin, 1974].

For Mendelian randomization, the similar key property of an IV is that the division of the population into genetic subgroups is independent of competing risk factors, and so genetic subgroups defined by the IV are exchangeable. For a genetic variant to be an IV, it is necessary that same distribution of outcomes would be observed if individuals with no copies of the genetic variant instead had one copy of the genetic variant (and the exposure distributions were unchanged), and vice versa. However, empirical testing of the exchangeability criterion is not possible.

We return to the question of assessing the validity of genetic variants as IVs later in this section; firstly we consider reasons why a genetic variant may not be a valid IV. These include issues of biological mechanism, genetic coinheritance, and population effects. Invalid IVs lead to unreliable inferences for the causal effect of an exposure. The situations discussed here represent potential lack of internal validity of estimates; the question of the external validity of an IV estimate as an estimate of the effect of a clinical intervention is discussed in Chapter 6.

3.2.3 Violations of IV assumptions: biological mechanisms

The first category of ways that we consider by which the IV assumptions may be violated is because of an underlying biological mechanism.

Pleiotropy: Pleiotropy refers to a genetic variant being associated with multiple risk factors. If a genetic variant used as an IV is additionally associated with another risk factor for the outcome, then either the second or the third IV assumption is violated (depending on whether the risk factor is a confounder of the exposure–outcome association or not), and the variant is not a valid IV.

If the genetic variant is associated with an additional variable solely due to mediation of the genetic association via the exposure of interest (sometimes called vertical pleiotropy), that this is not regarded as pleiotropy for our

purposes. For example, the *FTO* gene is a determinant of satiety (how full of food a person feels) [Wardle et al., 2008]. If satiety affects body mass index (BMI), then a variant in the *FTO* gene can be used as an IV for BMI if the two variables are on the same causal pathway, and if there is no alternative causal pathway from the genetic variant to the outcome not via BMI. However, if the *FTO* gene was also associated with (say) blood pressure, and this association was not completely mediated by the association of the gene with BMI, then it would be misleading to use a variant in the *FTO* gene to make specific inferences about the causal effect of BMI on an outcome.

Concerns about pleiotropy can be alleviated by using genetic variants located in genes, the biological function of which are well-understood. For example, for C-reactive protein (CRP), we can use genetic variants in the *CRP* gene which are known to have functional relevance in the regulation of CRP levels. Associations of a variant with measured covariates can be assessed to investigate potential pleiotropy, although such associations may also reflect mediation, particularly if the associations are consistent across independent variants.

Canalization: Canalization, or developmental compensation, is the phenomenon by which an individual adapts in response to genetic change in such a way that the expected effect of the change is reduced or absent [Debat and David, 2001]. It is most evident in knockout studies, where a gene is rendered completely inactive in an organism, typically a mouse. Often the organism develops a compensatory mechanism to allow for the missing gene such that the functionality of the gene is expressed via a different biological pathway. This buffering of the genetic effect may have downstream effects on other variables. Canalization may be a problem in Mendelian randomization if groups with different levels of the genetic variants differ with respect not only to the exposure of interest, but also to other risk factors via a canalization mechanism.

In a sense, canalization is not a violation of the IV assumptions, but merely an (often unwanted) consequence. Canalization is the same process as that assessed by Mendelian randomization, as any change in other risk factors from canalization occurs as a causal effect of the genetic variant. However, the aim of Mendelian randomization is not simply to describe the effects of genetic change, but to assess the causal effect of the (non-genetic) exposure. If there is substantial canalization, Mendelian randomization estimates may be unrepresentative of clinical interventions on the exposure performed in a mature cohort.

3.2.4 Violations of IV assumptions: non-Mendelian inheritance

The second category of ways that we consider by which the IV assumptions may be violated is because of non-Mendelian inheritance. Although Mendelian principles state that separate characteristics are inherited separately, this is not always true in practice. Non-Mendelian inheritance refers to patterns of

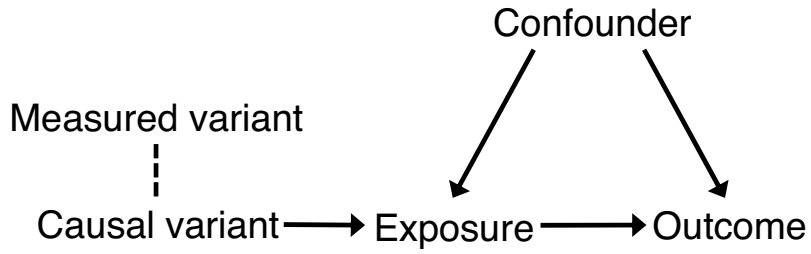


FIGURE 3.2
Graph of instrumental variable assumptions where a variant in linkage disequilibrium with the causal variant has been measured. Such a variant would still be a valid instrumental variable. The dashed line connecting the genetic variants indicates correlation without a causal interpretation.

inheritance which do not correspond to Mendel's laws, specifically the law of independent assortment.

Linkage disequilibrium: One particular reason for genetic variants to be inherited together is the physical proximity of the variants on the same chromosome. Variants whose distributions are correlated are said to be in linkage disequilibrium (LD). The opposite of LD is linkage equilibrium.

LD has both desirable and undesirable consequences. If genetic variants were truly independently distributed, then only the genetic variant which was causally responsible for variation in the exposure could be used as an IV, as all other genetic variants would not be associated with the exposure. In reality, it is not necessary for the genetic variant used as the IV to be the causal variant, merely to be correlated with the causal variant [Hernán and Robins, 2006]. This is because an IV must simply divide the population into subgroups which differ systematically only with respect to the exposure. This is illustrated in Figure 3.2.

An undesirable consequence of LD is that genetic variants correlated with the variant used in the analysis may have effects on competing risk factors. This would lead to the violation of the second or the third IV assumption (similar to violations due to pleiotropy). Concerns about invalid inferences due to LD can be alleviated by empirical testing of the association of known potential confounders with the measured variant.

Effect modification: Effect modification is a separate phenomenon from confounding, and relates to a statistical interaction between the effect of a variable (usually an effect of the exposure) and the value of a covariate, leading to the causal effect of the exposure varying across strata defined by the covariate. Factors that may lead to effect modification include (but are not limited to) issues of non-Mendelian inheritance, such epigenetic variation [Ogbuanu et al., 2009] and parent-of-origin effects [Bochud et al., 2008].

Effect modification alone is unlikely to represent a violation of the IV assumptions; however, it may lead to difficulties in interpreting Mendelian randomization investigations. Taking the example from Section 2.3.2 of the effect of alcohol intake on oesophageal cancer risk, in the Japanese population only men tend to drink alcohol. Hence, genetic associations with the outcome may be observed only in men and may not be present in women. If there are biological reasons for genetic associations to be stronger or weaker (or even absent) in some strata of the population, then associations measured in that stratum of the population would not be representative of the effect in the population as a whole. However, this may also provide an opportunity for verifying the IV assumptions; Japanese women are a natural control group for Japanese men. If the same genetic associations of alcohol-related variants with oesophageal cancer risk seen in Japanese men are not observed in Japanese women, this provides further evidence that the genetic associations with disease risk are driven by alcohol consumption, and not by violations of the IV assumptions.

3.2.5 Violations of IV assumptions: population effects

The final category of ways that we consider by which the IV assumptions may be violated is because of population effects.

Population stratification: Population stratification occurs when the population under investigation can be divided into distinct subpopulations. This may occur, for example, when the population is a mixture of individuals of different ethnic origins. If the frequency of the genetic variant and the distribution of the exposure are different in the different subpopulations, a spurious association between the variant and the exposure will be induced which is due to subpopulation differences, not the effect of the genetic variant. Violations of the IV assumptions may also occur if there is continuous variation in the structure of the population rather than distinct subpopulations.

Concerns about population stratification can be alleviated by restricting the study population to those with the same ethnic background (although there may still may be differences associated with ancestry in broadly-defined ethnic groups). In a genome-wide association study (GWAS), genomic control approaches, such as adjustment for genetic principal components, are possible. However, the use of Mendelian randomization in a population with a large amount of genetic heterogeneity is not advised.

Ascertainment effects: If the genetic variant is associated with recruitment into the study, then the relative proportions of individuals in each genetic subgroup are not the same as those in the population, and so a genetic association with the outcome in the sample may not be present in the original population. If the study cohort is taken from the general population, ascertainment effects are unlikely to be a major problem in practice. However, if, for example, the study cohort is pregnant mothers, and the genetic variant is associated with fertility, then the distributions of the covariates in the genetic

subgroups will differ and not be the same as those in the general population. This may introduce bias in the estimation of causal effects, as there is a pathway opened up from the genetic variant to the outcome by conditioning on a common cause of the variant and the outcome (sometimes called collider bias).

This would also be a problem in studies looking at genetic associations in populations of diseased individuals, such as clinical trials of secondary disease prevention. Individuals with greater genetically determined disease risk are less likely to survive to study recruitment, and so the randomization of individuals into genetic subgroups at conception would not hold in the study population, leading to biased genetic associations.

3.2.6 Statistical assessment of the IV assumptions

Although it is not possible to demonstrate conclusively the validity of the IV assumptions, several tests and assessments are possible to increase or decrease confidence in the use of genetic variants as IVs.

The simplest assessment of instrument validity is to test the association between the genetic variant and known confounders. Association of the variant with a covariate associated with the outcome which is not on the causal pathway between the exposure and outcome would violate the second IV assumption. However, there is no definitive way to tell whether the association with the covariate is due to violation of the IV assumptions (such as by pleiotropy or linkage disequilibrium), or due to mediation through the exposure of interest. Additionally, there is no way of testing whether or not the variant is associated with an unmeasured confounder. If there are multiple covariates and/or genetic variants, then any hypothesis testing approach needs to account for the multiple comparisons of each covariate, leading to a lack of power to detect any specific association. Additionally, as several covariates may be correlated, a simple Bonferroni correction may be an over-correction. A sensible way to proceed is to combine a hypothesis testing approach with a quantitative and qualitative assessment of the imbalance of the covariates between genetic subgroups and the degree to which this may bias the IV estimate.

A further approach for testing instrument validity is to see whether the association of a genetic variant with the outcome attenuates on adjustment for the risk factor [Glymour et al., 2012]. Although attenuation may not be complete even when the instrumental variable assumptions are satisfied for the risk factor due to confounding and measurement error [Didelez and Sheehan, 2007], if the attenuation is not substantial then the risk factor is unlikely to be on the causal pathway from the variant to the outcome.

If multiple genetic variants are available, each of which is a valid IV, then a separate IV estimate can be calculated using each of the instruments in turn. Assuming that each variant affects the exposure in a similar way, even if the genetic associations with the exposure are of different magnitude, the separate IV estimates should be similar, as they are targeting the same

quantity. This can be assessed graphically by plotting the genetic associations for an additional variant allele with the exposure and outcome for multiple variants: a straight line through the origin is expected, as in Figure 6.1. Formally, heterogeneity between variants can be tested using an overidentification test (Section 4.5.3). Failure of an overidentification test may be due to one or more of the IVs being invalid. However, the power of such tests may be limited in practice, and so testing should not be relied on for justification of the IV assumptions.

Other mathematical results for testing IV validity are available [Glymour et al., 2012], but these are only likely to detect gross violations of the IV assumptions. Biological knowledge rather than statistical testing should form the backbone of any justification of the use of a particular genetic variant as an IV in Mendelian randomization. The Bradford Hill criteria form a systematic summary of common-sense principles for assessing causality in epidemiological investigations [Hill, 1965]. In Table 3.1, we apply the relevant Bradford Hill criteria for causation to Mendelian randomization as a checklist to judge whether the validity of genetic variant(s) as an IV is plausible.

3.2.7 Summary of issues relating to IV validity

The validity of IVs is of vital importance to Mendelian randomization. It is our view that the choice of genetic variants as IVs should be justified mainly by basic biological knowledge but can be verified by empirical statistical testing. Appropriate caution should be attached to the interpretation of Mendelian randomization findings depending on the plausibility of the IV assumptions, and particularly to those where the justification of the IV assumptions is mainly empirical. This suggests that variants from candidate gene investigations, where the function of the genetic variant(s) is well-understood, will have more credibility for use in Mendelian randomization studies than variants outside of gene coding regions, such as those discovered in genome-wide association studies. However, it should be remembered that all statistical methods for assessing causal effects rely on untestable assumptions, and as such, Mendelian randomization has an important role in building up the case for the causal nature of a given exposure even if the validity of the IV assumptions can be challenged.

On a more positive note, a British study into the distribution of genetic variants 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 [Ebrahim and Davey Smith, 2008], indicating that genetic factors seem to be distributed independently of possible confounders in the population of the United Kingdom [Davey Smith, 2011]. 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 variants in question.

- **Strength:** If a genetic association with the outcome is slight, then the association could be explained by only a small imbalance in a covariate associated with the genetic variant. A small violation of the instrumental variable assumptions is less likely to be detected by testing the association of the variant with known covariates.

- **Consistency:** A causal relationship is more plausible if multiple genetic variants associated with the same exposure are all concordantly associated with the outcome, especially if the variants are located in different gene regions and/or have different mechanisms of association with the outcome.

- **Biological gradient:** Further, a causal relationship is more plausible if the genetic associations with the outcome and with the exposure for each variant are proportional (for example, as in Figure 6.1).

- **Specificity:** A causal relationship is more plausible if the genetic variant(s) are associated with a specific risk factor and outcome, and do not have associations with a wide range of covariates and outcomes. A specific association is most likely if the genetic variant(s) are biologically proximal to the exposure, and not biologically distant. This is most likely for risk factors that are biomarkers (such as C-reactive protein and low-density lipoprotein cholesterol), rather than generic risk factors (such as body mass index and blood pressure).

- **Plausibility:** If the function of the genetic variant(s) is known, a causal relationship is more plausible if the mechanism by which the variant acts is credibly and specifically related to the exposure.

- **Coherence:** If an intervention on the exposure has been performed (for example, if a drug has been developed that acts on the exposure), associations with intermediate outcomes (covariates) observed in the experimental context should also be present in the genetic context; directionally concordant genetic associations should be observed with the same covariates. For example, associations of genetic variants in the *IL6R* gene region with C-reactive protein and fibrinogen should be similar to those observed for tocilizumab, an interleukin-6 receptor inhibitor [Swerdlow et al., 2012].

TABLE 3.1
Bradford Hill criteria applied to Mendelian randomization for judging the biological plausibility of a genetic variant as an instrumental variable.

3.2.8* Definition of an IV as a random variable

For the more mathematically inclined, we give a further characterization of an IV in terms of random variables. We assume that we have an outcome Y that is a function of a measured exposure X and an unmeasured confounder U; that the confounding factors can be summarized by a single random variable U [Palmer et al., 2008], which satisfies the requirements of a sufficient covariate (Section 3.1.4); and that the exposure X can be expressed as a function of the confounder U and the genetic variant G. G may be a single genetic variant or a matrix corresponding to several genetic variants. The IV assumptions of Section 3.2.1 are 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 implies that the joint distribution of Y, X, U, G factorizes as

$$p(y, x, u, g) = p(y|u, x)p(x|u, g)p(u)p(g) \tag{3.1}$$

which corresponds to the directed acyclic graph (DAG) in Figure 3.3 [Dawid, 2002; Didelez and Sheehan, 2007].

It is a common mistake to think that the third IV assumption should read not $G \perp\!\!\!\perp Y|X, U$, but $G \perp\!\!\!\perp Y|X$, that is conditioning on U is not necessary. As X is a common descendent of G and U, conditioning on X induces an association between G and U, and therefore between G and Y. For example, if X and U are positively correlated and both have positive causal effects on Y, then conditional on X taking a value around the middle of its distribution, a large value of Y is associated with a low value of G. This is because the large value of Y is associated with a large value of U, and so G is more likely to be

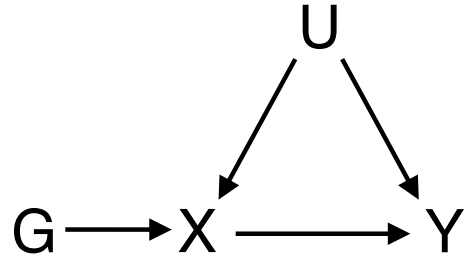


FIGURE 3.3
Directed acyclic graph of Mendelian randomization assumptions as random variables.

low so that the value of X is moderate and not large. The lack of independence ($G \not\perp\!\!\!\perp Y|X$) means that, in the regression of Y on X and G, the coefficient for G will generally be close to, but not equal to zero in a large sample (and especially if X is measured with error).

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

$$p(y, u, g, x|do(X = x_0)) = p(y|u, x_0)1(X = x_0)p(u)p(g) \tag{3.2}$$

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 [Didelez et al., 2010].

3.2.9* Definition of an IV in potential outcomes

In the "potential outcomes" or counterfactual causal framework (Section 3.1.2), a set of outcomes $Y(x), x \in \mathfrak{X}$ are considered to exist, where $Y(x)$ is the outcome which would be observed if the exposure were set to $X = x$ and $\mathfrak{X}$ is the set of possible values of the exposure. At most one of these outcomes is ever observed. The three assumptions of Section 3.2.1 necessary for the assessment of a causal relationship can be expressed in the language of potential outcomes as follows [Angrist et al., 1996]:

i'. Causal effect of IV on exposure: $p(x|g)$ is a non-trivial function of g

ii'. Independence of the potential exposures and outcomes from the IV: $X(g), Y(x, g) \perp\!\!\!\perp G$.

iii'. Exclusion restriction: $Y(x, g) = Y(x)$

where $p(x|g)$ is the probability distribution function of X conditional on $G = g$, $Y(x, g)$ is the potential outcome that would be observed if X were set to x and G were set to g, $Y(x)$ is the potential outcome observed when $X = x$, and $X(g)$ is the potential value of the exposure when $G = g$. Assumption ii'. states that the potential values of the exposure and outcome for each value of the IV do not depend on the actual value of the IV. This would not be true if, for example, the IV were associated with a confounder. Assumption iii'. is named 'exclusion restriction' and states that the observed outcome for each value of the exposure 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 exposure [Clarke and Windmeijer, 2010].

3.3 Testing for a causal relationship

Mendelian randomization studies are able to address two related questions: whether there is a causal effect of the exposure on the outcome, and what is the size of the causal effect [Tobin et al., 2004].

Under the assumption that the genetic variant is a valid IV, the hypothesis of a causal effect of the exposure on the outcome can be assessed by testing for independence of the variant and the outcome. A non-zero association is indicative of a causal relationship [Hernán and Robins, 2006]. The presence and direction of effect can be tested statistically by straightforward regression of the outcome on the genetic variant to see whether the estimated association is compatible with no causal effect based on a chosen threshold for statistical significance.

3.3.1 Converse of the test

The converse statement to the test for a causal relationship is that if the correlation between the outcome and variant is zero, then there is no causal effect of the exposure on the outcome. Although this converse statement is not always true, as there may be zero linear correlation between the variant and outcome without independence [Spirtes et al., 2000], it is true for most biologically plausible models of the exposure–outcome association.

3.3.2 Does Mendelian randomization really assess a causal relationship?

In a natural experiment such as Mendelian randomization, as there is no intervention or manipulation of the exposure, use of the label 'causal' relies on the assumption that the observational relationships between the genetic variant(s), exposure, and outcome are informative about the structural relationship between the exposure and the outcome (structural meaning relating to the distribution of the variables under intervention). Put simply, this assumption states that the effect on the outcome of the unconfounded observed difference in the exposure due to the genetic variant would be similar (same direction of effect) if the genetic variant (or equivalently, the exposure) were manipulated to take different values, rather than being observed at different values. Hence although Mendelian randomization is an observational rather than an experimental technique, under this assumption it does assess a causal relationship.

3.3.3 Interpreting a null result

A difficulty faced by practitioners of Mendelian randomization is how to interpret a 'null' (for example $p > 0.05$) finding. In such cases, above all, caution must be exercised against the overinterpretation of a null finding which may simply be due to low power.

One common approach is to compare the observed and 'expected' association between the exposure and the outcome; the latter is based on triangulating the associations between the genetic variant and the exposure and between the variant and the outcome (Figure 3.4). This 'expected' association is calculated as the coefficient from the regression of the outcome on the variant divided by the coefficient from the regression of the exposure on the variant. This is a ratio estimate (Section 4.1), and is the change in the outcome expected for a unit change in the exposure if there were no confounding in the observational association between the exposure and the outcome. While there is some merit in comparing the 'expected' and observed association estimates of the exposure with the outcome, this comparison should be seen as a guide rather than a conclusive statistical test. (The formal test of comparison of these estimates is known as an endogeneity test; reasons why we discourage reliance on an endogeneity test are given in Section 4.5.4.) If the expected and observed association estimates are similar, then a null finding may give little evidence as to the causal nature of the exposure. Even if the estimates are different, there may be good biological reasons for a difference other than residual confounding (Chapter 6). A better approach is to consider an estimate of the causal effect and of its precision using an IV method.

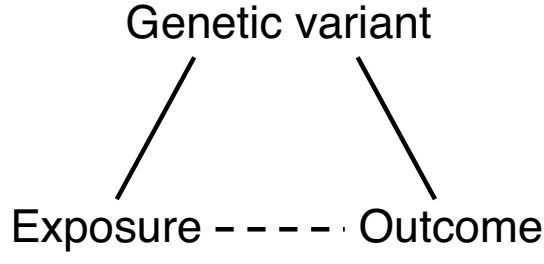


FIGURE 3.4
Triangle of associations: an 'expected' association estimate between the exposure and the outcome can be calculated by dividing the coefficient for the association between the genetic variant and the outcome by the coefficient for the association between the exposure and the variant. The dashed line is the association which is estimated.

3.4 Estimating a causal effect

Although testing for a causal relationship is useful and may be sufficient in some cases, there are several reasons why it is desirable to go beyond this and to estimate the size of a causal effect. First, this is usually the parameter representing the answer to the question of interest. 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 at a given level even if none of the estimates from the individual IVs achieve significance. Thirdly, often a null association is expected. By estimating a confidence interval for the causal effect, we obtain bounds on its plausible size. Although it is not statistically possible to prove the null hypothesis, it may be possible to obtain a sample size large enough such that the confidence interval bounds for the causal effect are narrow enough that the range of plausible causal effect values excludes a minimally clinically relevant causal effect.

In this section, we consider technical issues associated with parameter estimation: the assumptions necessary to estimate a causal effect, and definitions of the causal parameters to be estimated. Having discussed these points, we proceed in the next chapter to consider methods for constructing different IV estimators.

3.4.1* Additional IV assumptions for estimating a causal effect

In order to estimate a causal effect, it is necessary to make further assumptions to the ones listed in Section 3.2.1 [Angrist et al., 1996]:

1. The stable unit treatment value assumption (SUTVA), which states that the potential outcomes for each individual should be unaffected by how the exposure was assigned, and unaffected by variables in the model relating to other individuals [Cox, 1958];
2. Strong monotonicity, which means that varying the IV should alter the exposure for at least one individual in the population, and that any change in the exposure from varying the IV should be in the same direction (an increase or a decrease) for all individuals.

The monotonicity assumption is credible for most biologically plausible situations in which Mendelian randomization investigations for estimating a causal effect are undertaken. It would not be plausible in the example from Section 2.3.2 of the effect of alcohol intake on oesophageal cancer risk, as the average levels of alcohol intake and the associated disease risk are not monotone in the number of variant alleles. If the monotonicity assumption is not plausible (for example, if the IV is an unweighted allele score, Section 8.2),

then a causal effect can be identified under a homogeneity assumption that the causal effect has the same magnitude in all individuals [Swanson and Hernán, 2013]. If the monotonicity assumption is plausible, then an IV analysis typically estimates an average causal effect, known as the local average treatment effect or complier-average causal effect. This is the average causal effect (see below) amongst individuals whose exposure value is influenced by the IV. This may be the whole population; an example where it is a subset of the population is for the exposure of alcohol intake, where individuals who abstain from alcohol for cultural or religious reasons would do so regardless of their IV value.

If there is a single IV, then this could be used to calculate the average causal effect of changing the value of the IV from one value to another. However, it is usually desired to express a causal effect in terms of the exposure. For this, it is necessary to assume a parametric relationship between the exposure and outcome. For a continuous outcome, this is usually a linear model; the expected value of the outcome is a linear function of the exposure. For a binary outcome, this may be a linear model for the probability of an outcome, but is more often a log-linear model or a logistic-linear model; the log-transformed (or logit-transformed) probability of an outcome is a linear function of the exposure. Non-linear parametric models have been considered for IV analysis; however, inference from such models has been shown to be highly dependent on the parametric form considered [Mogstad and Wiswall, 2010; Horowitz, 2011]. Non-parametric models are discussed in Section 11.1.2.

The SUTVA is generally not plausible in Mendelian randomization, as the effect on an outcome associated with a genetic variant is likely to be different to the effect from intervention on the exposure in a number of qualitative and quantitative ways (Chapter 6) – for example, due to the duration of the intervention (life-long or short-term), the timing of the intervention (on long-term levels of the exposure or on acute levels), the magnitude of the intervention (genetic effects are usually small, clinical interventions are typically larger), and the mechanism of the intervention (genetic effects and clinical interventions may operate via different pathways). Estimates from Mendelian randomization should therefore not be interpreted naively as the expected outcome of an intervention in the risk factor of interest (Section 6.3.3).

3.4.2* 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) [Greenland, 1987].

The average causal effect (ACE) [Didelez and Sheehan, 2007] under intervention on the exposure is the expected difference in the outcome when the

exposure is set to two different values:

$$ACE(x_0, x_1) = \begin{array}{l}\text{Expected outcome when the exposure is set at } x_1 \\ \quad - \text{ expected outcome when the exposure is set at } x_0.\end{array} \quad (3.3)$$

This can be written as:

$$ACE(x_0, x_1) = \mathbb{E}(Y|do(X = x_1)) - \mathbb{E}(Y|do(X = x_0)). \quad (3.4)$$

The ACE is zero when there is conditional independence between Y and X given U, but the converse is not generally true [Didelez and Sheehan, 2007].

With a binary outcome ($Y = 0$ or 1), 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):

$$CRR(x_0, x_1) = \frac{\text{Probability of outcome when the exposure is set at } x_1}{\text{Probability of outcome when the exposure is set at } x_0}, \quad (3.5)$$

$$COR(x_0, x_1) = \frac{\text{Odds of outcome when the exposure is set at } x_1}{\text{Odds of outcome when the exposure is set at } x_0}. \quad (3.6)$$

These can be written as:

$$CRR(x_0, x_1) = \frac{\mathbb{P}(Y = 1|do(X = x_1))}{\mathbb{P}(Y = 1|do(X = x_0))}, \quad (3.7)$$

$$COR(x_0, x_1) = \frac{\mathbb{P}(Y = 1|do(X = x_1))\mathbb{P}(Y = 0|do(X = x_0))}{\mathbb{P}(Y = 1|do(X = x_0))\mathbb{P}(Y = 0|do(X = x_1))}. \quad (3.8)$$

3.5 Summary

The instrumental variable assumptions make assessment of causation in an observational setting possible without complete knowledge of all the confounders of the exposure–outcome association. Genetic variants have good theoretical and empirical plausibility for use as instrumental variables in general, but the instrumental variable assumptions may be violated for a number of reasons.

We continue in the next chapter to consider methods for estimating the magnitude of a causal effect using instrumental variables.

4

Methods for instrumental variable analysis

In this chapter, we discuss methods for the estimation of causal effects using instrumental variables (IVs) with both continuous and binary outcomes. We focus attention on the case of a single continuous exposure variable, as this is the usual situation in Mendelian randomization studies; although the same methods could be used in the case of a single binary exposure. We explain for each method how to estimate a causal effect, and describe specific properties of the estimator. In turn, we consider the ratio of coefficients method, two-stage methods, likelihood-based methods, and semi-parametric methods. This order corresponds roughly to the complexity of the methods, with the simplest ones first. These methods are contrasted in terms of bias, coverage, efficiency, power, robustness to misspecification, and existence of finite moments. We have included a simple explanation of each method at first, and then further details for more technical readers. Also discussed are implementations of the methods using standard statistical software packages.

4.1 Ratio of coefficients method

The ratio of coefficients method, or the Wald method [Wald, 1940], is the simplest way of estimating the causal effect of the exposure (X) on the outcome (Y). The ratio method uses a single IV. If more than one variant is available which is an IV then the causal estimates from the ratio method using each variant can be calculated separately, or the variants can be combined into a single IV in an allele score approach (Section 8.2). Otherwise, other estimation methods in this chapter can be used.

4.1.1 Continuous outcome, dichotomous IV

We initially assume that we have an IV G which takes the values 0 or 1, dividing the population into two genetic subgroups. The IV can be thought of as a single nucleotide polymorphism (SNP) where two of the three subgroups are merged together, for example reflecting a dominant or recessive genetic model, or because there are very few individuals in the least common genetic

subgroup (the minor homozygotes). In a recessive model, a single copy of the major (wildtype) allele A is sufficient to mask a minor (variant) allele; the genetic subgroups are AA/Aa (major homozygote/heterozygote) and aa (minor homozygote). A dominant model is similar, except that the heterozygotes are combined with the minor homozygotes; the two genetic subgroups are AA and Aa/aa.

From the IV assumptions, the distribution of the exposure differs in the two genetic subgroups. If the distribution of the outcome also differs, then there is a causal effect of the exposure on the outcome. We define $\bar{y}_j$ for $j = 0, 1$ as the average value of outcome for all individuals with genotype $G = j$, and define $\bar{x}_j$ similarly for the exposure. Figure 4.1 displays the mean exposure and outcome in the two genetic subgroups in a fictitious example with a positive causal effect of X on Y.

IV estimates are usually expressed as the change in the outcome resulting from a unit change in the exposure, although changes in the outcome

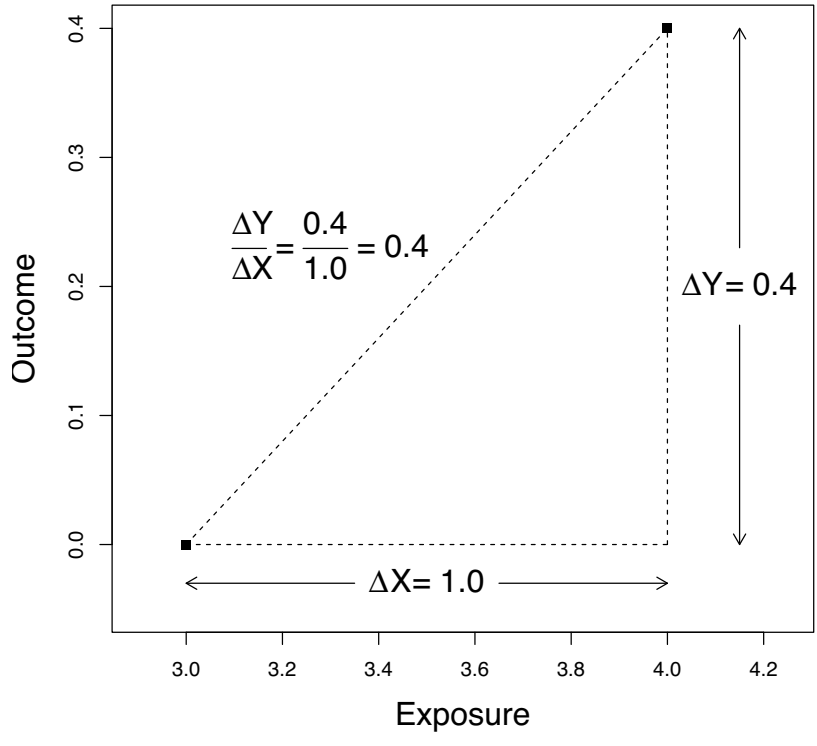


FIGURE 4.1
Points representing mean exposure and outcome in two genetic subgroups with IV ratio estimate.

corresponding to different magnitudes of change in the exposure could be quoted instead. If the exposure has been (natural) log-transformed, a unit increase in the log-transformed exposure corresponds to a $\exp(1) = 2.72$-fold multiplicative in the untransformed exposure. The effect of a (say) 20% increase in the exposure can be considered by multiplying the causal estimate by $\log(1.2) = 0.182$, or of a 30% decrease by multiplying by $\log(0.7) = -0.357$ (where log is the natural logarithm or ln). If an IV estimate is expressed for a change in the exposure much greater than that associated with the genetic variant, this extrapolation may not be justified and the IV estimate may not be realistic. However, some extrapolation is often necessary to convert the genetic association to a clinically relevant causal effect of the exposure.

We see that an average difference in the exposure between the two subgroups of $\Delta X = \bar{x}_1 - \bar{x}_0$ results in an average difference in the outcome of $\Delta Y = \bar{y}_1 - \bar{y}_0$. Assuming that the effect of the exposure on the outcome is linear, the ratio estimate for the change in outcome due to a unit increase in the exposure is:

$$\text{Ratio method estimate (dichotomous IV)} = \frac{\Delta Y}{\Delta X} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}. \qquad (4.1)$$

In the example shown (Figure 4.1), $\Delta Y = 0.4$ and $\Delta X = 1.0$, giving a ratio estimate of $\frac{0.4}{1.0} = 0.4$.

The numerator and denominator in the ratio estimate are the average causal effects on the outcome and exposure respectively of being in genetic subgroup 1 versus being in genetic subgroup 0. If we assume that the effect of the exposure on the outcome is linear, then the ratio estimate is the average causal effect on the outcome of an exposure of $x + 1$ units versus an exposure of x units. (Under the linearity assumption, the causal effect of a unit increase in the exposure is equal for all values of x.) If the effect is not linear, then a ratio estimate approximates the average causal effect of a population intervention in the exposure [Burgess et al., 2014b]. (Non-linear exposure–outcome relationships are discussed further in Section 11.1.2.)

4.1.2 Continuous outcome, polytomous or continuous IV

Alternatively, the IV may not be dichotomous, but polytomous (takes more than two distinct values). This is the usual case for a diallelic SNP; the three levels AA (major homozygote), Aa (heterozygote), and aa (minor homozygote) will be referred to as 0, 1, and 2, corresponding to the number of minor alleles. In a linear 'per allele' model, we assume that the association of the genetic variant with the exposure is proportional to the number of variant alleles. The IV could also be a continuous allele score (Section 8.2), under the assumption that the association of the score with the exposure is also linear.

The coefficient of G in the regression of X on G is written as $\hat{\beta}_{X|G}$, and represents the change in X for a unit change in G. Similarly, the coefficient of G in the regression of Y on G is written as $\hat{\beta}_{Y|G}$. The ratio estimate of the

causal effect is:

$$\text{Ratio method estimate (polytomous/continuous IV)} = \frac{\hat{\beta}_{Y|G}}{\hat{\beta}_{X|G}}. \tag{4.2}$$

Intuitively, we can think of the ratio method as saying that the change in Y for a unit increase in X is equal to the change in Y for a unit increase in G, scaled by the change in X for a unit increase in G.

Illustrative data are shown in Figure 4.2. Each of the graphs is plotted on the same scale. The top-left panel shows that the exposure and outcome are negatively correlated, with the line showing the observational association from linear regression. However, as shown in the top-right panel, where individuals in different genetic subgroups are marked with different plotting symbols, individuals in the subgroup marked with circles tend to congregate towards the south-west of the graph and individuals in the subgroup marked with squares tend towards the north-east of the graph. The bottom-left panel shows the mean values of the exposure and outcome in each genetic subgroup with lines representing 95% confidence intervals intervals for the means. The bottom-right panel includes the individual data points, the subgroup means and the causal estimate from the ratio method. We see that the causal estimate is positive. The 95% confidence intervals for the lines passing through the points show that the uncertainty in the ratio IV estimate is greater than that of the observational estimate.

From a technical point of view, the ratio estimator is valid under the assumption of monotonicity of the genetic effect on the exposure and linearity of the causal X–Y association [Angrist et al., 2000]. Because of this, the ratio estimate has been named the linear IV average effect (LIVAE) [Didelez et al., 2010]. Monotonicity means that the exposure for each individual would be increased (or alternatively for each individual would be decreased) or unchanged if that person had $G = g_1$ compared to if they had $G = g_0$ for all $g_1 > g_0$. We note that it is not necessary for the genetic effect on the exposure to be constant in magnitude for all individuals, merely consistent in direction (that is, there may be effect modification), or for the exposure effect on the outcome to be constant in magnitude. If the monotonicity assumption is not satisfied, then the causal effect of the exposure on the outcome can only be estimated consistently if it is constant for all individuals across the population.

The linearity assumption is that the expected value of the outcome Y conditional on the exposure X and confounders U is:

$$\mathbb{E}(Y|X = x, U = u) = \beta_0 + \beta_1 x + h(u) \tag{4.3}$$

where $h(u)$ is a function of U. Hence, there is no interaction term between X and U in the conditional expectation of Y. It is also required that the structural model:

$$\mathbb{E}(Y|do(X = x)) = \beta_0' + \beta_1 x \tag{4.4}$$

holds, where the causal effect β_1 is the same as in the equation above. This is

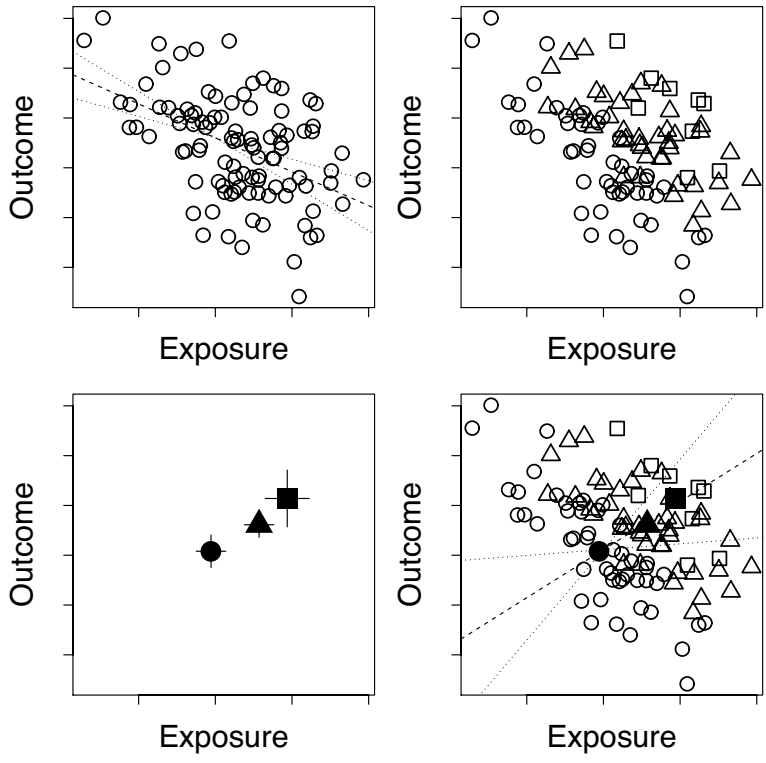


FIGURE 4.2
Illustration of ratio method for polytomous IV taking three values with a continuous outcome in a fictitious dataset: (top-left) exposure and outcome for all individuals, observational estimate with 95% confidence interval; (top-right) individuals divided into genetic subgroups by plot symbol; (bottom-left) mean exposure and outcome in each genetic subgroup (lines represent 95% confidence intervals); (bottom-right) ratio IV estimate with 95% confidence interval.

similar to a consistency assumption, which states that the outcome for an individual would be the same if the value of the exposure were observed naturally or set due to an intervention [VanderWeele, 2009]. Although confounding is represented by a single variable U, this is simply for presentation; U represents the combined effect of all confounding variables.

We note that the ratio estimate can be calculated simply from the coefficients $\hat{\beta}_{X|G}$ and $\hat{\beta}_{Y|G}$, and as such only requires the availability of summarized data, not individual-level data. Methods for obtaining IV estimates using summarized data are discussed further in Section 9.4. The two coefficients can also be estimated in different groups of individuals. Common examples include where the IV–outcome association is measured on the whole sample and the IV–exposure association on a subsample (subsample Mendelian randomization, see Section 8.5.2), or the associations are estimated on non-overlapping datasets (two-sample Mendelian randomization, see Section 9.8.2).

4.1.3 Binary outcome

Generally in epidemiological applications, disease is the outcome of interest. Disease outcomes are often dichotomous. We use the epidemiological terminology of referring to an individual with an outcome event as a case ($Y = 1$), and an individual with no event as a control ($Y = 0$).

With a binary outcome and a dichotomous IV, the ratio estimate is defined similarly as with a continuous outcome:

$$\begin{aligned}\text{Ratio method log risk ratio estimate (dichotomous IV)} &= \frac{\Delta Y}{\Delta X} \\ &= \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}\end{aligned} \tag{4.5}$$

where $\bar{y}_j$ is commonly the log of the probability of an event, or the log odds of an event, in genetic subgroup j. The term "risk ratio" is used as a generic term meaning relative risk (for the log of the probability) or odds ratio (for the log odds) as appropriate.

With a polytomous or continuous IV, the coefficient $\hat{\beta}_{Y|G}$ in the ratio estimate (equation 4.2) is taken from regression of Y on G. The regression model used could in principle be linear, where the IV estimate represents the change in the probability of an event for a unit change in the exposure. However, with a dichotomous outcome, log-linear or logistic regression models are generally preferred, where the IV estimate represents the log relative risk or log odds ratio, respectively, for a unit change in the exposure. With logistic models, the odds ratio being estimated depends on the choice of covariates included in the model (Section 4.2.3*).

The ratio estimate is also commonly quoted in its exponentiated form:

$$\text{Ratio method risk ratio estimate (dichotomous IV)} = R^{1/\Delta X} \tag{4.6}$$

where R is the estimated risk ratio between the two genetic subgroups.

As in the continuous case, this estimator is valid under the assumption of monotonicity of X on G and a log-linear or logistic-linear model [Didelez et al., 2010]. In the log-linear case, the association model is:

$$\log(\mathbb{E}(Y|X = x, U = u)) = \beta_0 + \beta_1 x + h(u) \tag{4.7}$$

and the structural model is:

$$\log(\mathbb{E}(Y|do(X = x))) = \beta_0' + \beta_1 x \tag{4.8}$$

for some β_0, β_0', β_1, $h(u)$ as above.

4.1.4 Retrospective and case-control data

In Mendelian randomization, when retrospective data are available, it is usual to make inferences on the gene–exposure association using only non-diseased individuals, such as the control population in a case-control study [Minelli et al., 2004]. This makes the assumption that the distribution of the exposure in the controls is similar to that of the general population, which is true for a rare disease [Bowden and Vansteelandt, 2011]. This is necessary to prevent bias of the causal estimate for two reasons. The first reason is reverse causation, whereby post-event measurements of the exposure may be distorted by the outcome event. Secondly, in a case-control setting, over-recruitment of cases into the study means that the distribution of confounders in the ascertained population is different to that in the general population. An association is then induced between the IV and the confounders, leading to possible bias in the IV estimate [Didelez and Sheehan, 2007]. This affects not only the ratio method, but all IV methods.

If the outcome is common and its prevalence in the population from which the case-control sample was taken is known, such as in a nested case-control study, then inferences on the gene–exposure association can be obtained using both cases and controls, provided that measurements of the exposure in cases were taken prior to the outcome event. This analysis can be performed by weighting the sample so that the proportions of cases and controls in the reweighted sample match those in the underlying population [Bowden and Vansteelandt, 2011].

4.1.5 Confidence intervals

Confidence intervals for the ratio estimate can be calculated in several ways.

Normal approximation: The simplest way is to use a normal approximation. With a continuous outcome, standard errors (SEs) and confidence intervals from the two-stage least squares method, introduced below, are given in standard software commands (Section 4.6). Alternatively, the following approximation can be used, based on the first two terms of the delta method

expansion for the variance of a ratio:

$$\text{Standard error of ratio estimate} \simeq \sqrt{\frac{\text{se}(\hat{\beta}_{Y|G})^2}{\hat{\beta}_{X|G}{}^2} + \frac{\hat{\beta}_{Y|G}{}^2 \, \text{se}(\hat{\beta}_{X|G})^2}{\hat{\beta}_{X|G}{}^4}} \quad (4.9)$$

This approximation assumes that the numerator and denominator of the ratio estimator are uncorrelated; such correlation could be accounted for by including a third term of the delta expansion [Thomas et al., 2007], but is unlikely to have a considerable impact on the estimate of the standard error.

However, asymptotic (large sample) normal approximations may result in overly narrow confidence intervals, especially if the sample size is not large or the IV is 'weak'. This is because IV estimates are not normally distributed.

Fieller's theorem: If the regression coefficients in the ratio method $\hat{\beta}_{Y|G}$ and $\hat{\beta}_{X|G}$ are assumed to be normally distributed, critical values and confidence intervals for the ratio estimator may be calculated using Fieller's theorem [Fieller, 1954; Lawlor et al., 2008]. We assume that the correlation between $\hat{\beta}_{Y|G}$ and $\hat{\beta}_{X|G}$ is zero; other values can be used, but the impact on the confidence interval is usually small [Minelli et al., 2004]. If the standard errors are $\text{se}(\hat{\beta}_{Y|G})$ and $\text{se}(\hat{\beta}_{X|G})$ and the sample size is N, then we define:

$$\begin{aligned} f_0 &= \hat{\beta}_{Y|G}{}^2 - t_N(0.975)^2 \, \text{se}(\hat{\beta}_{Y|G})^2 \\ f_1 &= \hat{\beta}_{X|G}{}^2 - t_N(0.975)^2 \, \text{se}(\hat{\beta}_{X|G})^2 \\ f_2 &= \hat{\beta}_{Y|G} \, \hat{\beta}_{X|G} \\ D &= f_2{}^2 - f_0 \, f_1 \end{aligned} \quad (4.10)$$

where $t_N(0.975)$ is the 97.5th percentile point of a t-distribution with N degrees of freedom (for $N > 100$, $t_N(0.975) \approx 1.96$).

If $D > 0$ and $f_1 > 0$, then the 95% confidence interval is from $(f_2 - \sqrt{D})/f_1$ to $(f_2 + \sqrt{D})/f_1$. The confidence interval is more likely to be a closed interval like this if we have a 'strong' instrument, that is, an instrument which explains a large proportion of the variation of the exposure in the population. Confidence intervals of size α can be similarly constructed by using the $(1 - \alpha/2)$ point of the t-distribution.

If $D < 0$, then there is no interval which covers the true parameter with 95% confidence. This occurs when there is little differentiation in both the exposure and outcome distributions between the genetic subgroups (due to a weak instrument), and so a gradient corresponding to any size of causal effect is plausible. The only valid 95% confidence interval is the unbounded interval from minus infinity to plus infinity. An example where Fieller's theorem would give an unbounded confidence interval is displayed in Figure 4.3. This situation is likely to occur when the IV explains little of the variation in the exposure; it is a weak instrument.

If $D > 0$ and $f_1 < 0$, then the 95% confidence interval is the union of two intervals from minus infinity to $(f_2 + \sqrt{D})/f_1$ and from $(f_2 - \sqrt{D})/f_1$

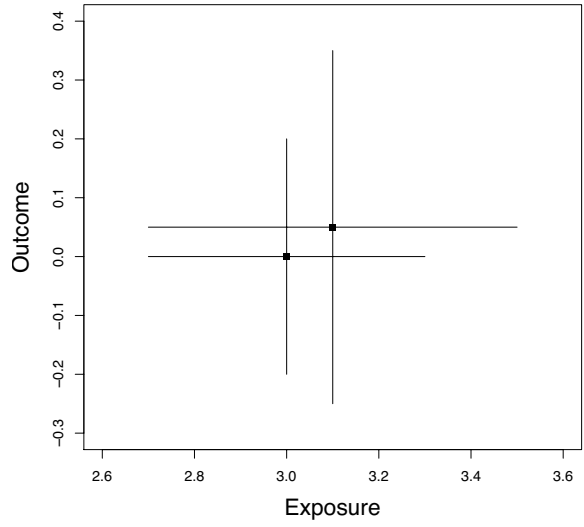


FIGURE 4.3
Points representing mean exposure and outcome (lines are 95% confidence intervals) in two genetic subgroups where the confidence interval from Fieller's theorem for the IV ratio estimate is unbounded.

to plus infinity. All possible values are included in the interval except those between $(f_2+\sqrt{D})/f_1$ and $(f_2-\sqrt{D})/f_1$. An example where Fieller's theorem would give such a confidence interval including infinity but excluding zero, is displayed in Figure 4.4. This suggests that the differences in the outcome are not caused solely by differences in the exposure, and so the IV assumptions are violated.

To summarize, Fieller's theorem gives confidence intervals that have one of three possible forms [Buonaccorsi, 2005]:

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

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

iii. The interval may be unbounded.

where $a = (f_2-\sqrt{D})/f_1$, $b = (f_2+\sqrt{D})/f_1$. Confidence intervals from Fieller's theorem are preferred to those from an asymptotic normal approximation when the IV is weak. A tool to calculate confidence intervals from Fieller's theorem based on the gene–exposure and gene–outcome associations is available online (http://spark.rstudio.com/sb452/fieller/).

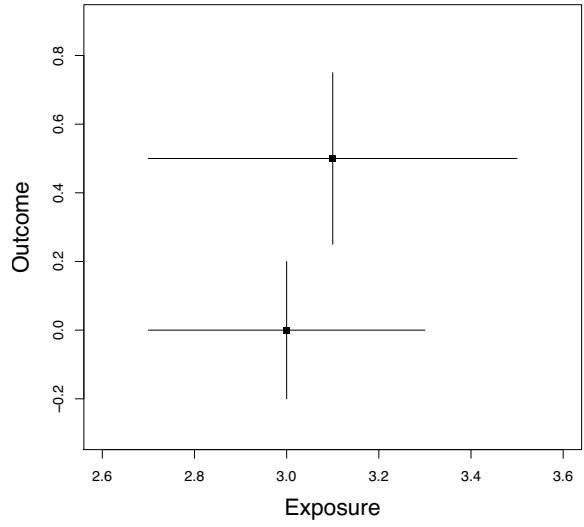


FIGURE 4.4
Points representing mean exposure and outcome (lines are 95% confidence intervals) in two genetic subgroups where the confidence interval from Fieller's theorem for the IV ratio estimate is compatible with an infinite (vertical) association, but not a null (horizontal) association.

Bootstrapping: As an alternative approach, also applicable to any of the following methods, confidence intervals can be calculated by bootstrapping [Efron and Tibshirani, 1993]. The simplest way of constructing a bootstrapped confidence interval is by taking several random samples with replacement from the data of the same sample size. The empirical distribution of the IV estimator in the bootstrapped samples approximates the true distribution of the IV estimator [Imbens and Rosenbaum, 2005]. However, there are some concerns about the behaviour of bootstrapped confidence intervals with weak instruments [Moreira et al., 2009].

Other approaches: Alternative approaches for inference with weak instruments not discussed further here are confidence based on inverting a test statistic, such as the Anderson–Rubin test statistic [Anderson and Rubin, 1949] or the conditional likelihood ratio test statistic [Moreira, 2003]. These intervals give appropriate confidence levels under the null hypothesis with weak instruments, but may be underpowered with stronger instruments. They have been discussed in detail elsewhere [Mikusheva, 2010; Davidson and MacKinnon, 2014] and implemented in Stata [Mikusheva and Poi, 2006] and R [Small, 2014].

4.1.6 Absence of finite moments

One peculiar property of the ratio estimator is that its mean (also known as its first moment) is not finite. This implies that, if you generated data on the exposure and outcome from a model with a valid IV and calculated the ratio IV estimate a large number of times, the mean value of these IV estimates could become arbitrarily high (or low). This is due to the fact that there is a finite probability that the denominator in the ratio estimate (ΔX or $\hat{\beta}_{X|G}$) is very close to zero, leading to a large IV estimate. In practice, this is unlikely to be a serious issue since, if ΔX were close to zero, the IV would be considered invalid as assumption i. (Section 3.2.1) would appear to be violated. Theoretically, the absence of a finite mean makes comparison of IV methods more difficult, as the (mean) bias of the ratio estimate, defined as the difference between the mean IV estimate (the expected value of the IV estimate) and the true value of the causal effect, cannot be calculated for any finite sample size. We therefore additionally consider the median bias, the difference between the median of the estimator over its distribution and the true causal effect, when comparing different methods for IV estimation.

Central moments (often simply called the moments) are the expectations of the powers of a random variable with its mean subtracted. The kth moment of the random variable Z with mean μ is $\mathbb{E}((Z-\mu)^k)$, for $k = 1, 2, \ldots$. All of the central moments of the ratio IV estimator are infinite. In particular, its mean and variance are undefined.

4.1.7 Coverage and efficiency

The coverage of a confidence interval is the probability that the confidence interval contains the true parameter value. By definition, a 95% confidence interval should contain the true parameter value 95% of the time. However, in practice, this may not be true, due to approximations and distributional approximations made in constructing the interval. By simulating data where the true parameter values are known, the coverage properties of differently estimated confidence intervals can be investigated.

Efficiency is a property of an estimator relating to its variance. An efficient estimator has low variance and therefore a narrow confidence interval. A desirable estimator has a narrow confidence interval, but maintains the correct coverage. The coverage and efficiency of various IV estimators are discussed in this chapter, but addressed in more detail in Chapter 7 onwards.

4.1.8 Reduced power of IV analyses

Figure 4.2 illustrates the wider confidence interval of an IV estimate compared with that of an observational estimate. As in many areas of applied statistics, there is a trade-off in choice of estimation procedure between bias and variance. The observational estimate is precisely estimated, but typically

biased for the causal effect, whereas the IV estimate is unbiased, but typically imprecisely estimated. The loss of precision in the IV estimate is the cost of unbiased estimation. When making causal assessments, we would argue that no appreciable amount of bias should be introduced in order to reduce the variance of the estimate [Zohoori and Savitz, 1997].

However, the sample size required to obtain precise enough causal estimates to be clinically relevant can be very large [Ebrahim and Davey Smith, 2008]. A rule of thumb for power is that the sample size for a conventional analysis should be divided by the coefficient of determination (R^2) of the IV on the exposure (Section 8.3) [Wooldridge, 2009]. For example, if the sample size for an observational regression analysis of the outcome on the exposure to detect a given effect size requires a sample size of 400, and the IV explains 2% of the variation in the exposure, then the sample size required for an IV analysis is approximately $400/0.02 = 20000$. For this reason, while for some researchers the ratio method may be sufficient for the analysis in question, we are motivated to consider methods which can incorporate data on more than one IV, and hence give more precise estimates of causal effects.

4.2 Two-stage methods

A two-stage method comprises two regression stages: the first-stage regression of the exposure on the genetic IVs, and the second-stage regression of the outcome on the fitted values of the exposure from the first stage.

4.2.1 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). It can be used with multiple IVs. In the first-stage (G–X) regression, the exposure is regressed on the IV(s) to give fitted values of the exposure ($\hat{X}|G$). In the second-stage (X–Y) regression, the outcome is regressed on the fitted values for the exposure from the first stage regression. The causal estimate is this second-stage regression coefficient for the change in outcome caused by a unit change in the exposure.

With a single IV, the 2SLS estimate is the same as the ratio estimate (with a continuous and with a binary outcome). With multiple IVs, the 2SLS estimator may be viewed as a weighted average of the ratio estimates calculated using the instruments one at the time, where the weights are determined by the relative strengths of the instruments in the first-stage regression [Angrist et al., 2000; Angrist and Pischke, 2009].

Suppose we have K instrumental variables available. With data on individuals indexed by $i = 1, \ldots, N$ who have exposure x_i, outcome y_i and assuming an additive per allele model for the IVs g_{ik} indexed by $k = 1, \ldots, K$, the

first-stage regression model is:

$$x_i = \alpha_0 + \sum_k \alpha_k g_{ik} + \varepsilon_{Xi}. \tag{4.11}$$

The fitted values $\hat{x}_i = \hat{\alpha}_0 + \sum_k \hat{\alpha}_k g_{ik}$ are then used in the second-stage regression model:

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \varepsilon_{Yi} \tag{4.12}$$

where ε_{Xi} and ε_{Yi} are independent error terms. The causal parameter of interest is β_1. If both models are estimated by standard least-squares regression, both the error terms are implicitly assumed to be normally distributed.

Although estimation of the causal effect in two stages (a sequential regression method) gives the correct point estimate, the standard error from the second-stage regression (equation 4.12) is not correct. This is because it does not take into account the uncertainty in the first-stage regression. Under homoscedasticity of the error term in the equation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon'_{Yi} \tag{4.13}$$

the asymptotic variance of the 2SLS estimator is:

$$\hat{\sigma}^2 (X^T G (G^T G)^{-1} G^T X)^{-1} = \hat{\sigma}^2 (\hat{X}^T \hat{X})^{-1} \tag{4.14}$$

where $\hat{\sigma}^2$ is an estimate of the variance of the residuals from equation (4.13), and the matrices G of IVs and X for the exposure contain constant terms. The use of 2SLS software is recommended for estimation (Section 4.6) [Angrist and Pischke, 2009]. Robust standard errors are often used in practice, as estimates are sensitive to heteroscedasticity and misspecification of the equations in the model.

When all the associations are linear and the error terms normally distributed, the 2SLS estimator has a finite kth moment when there are at least $(k + 1)$ IVs [Kinal, 1980]. Therefore the mean of a 2SLS estimator is only defined when there are at least 2 IVs, and the variance is only defined when there are at least 3 IVs.

4.2.2 Binary outcome

The analogue of 2SLS with binary outcomes is a two-stage estimator where the second-stage (X–Y) regression uses a log-linear or logistic regression model. This can be implemented using a sequential regression method by performing the two regression stages in turn (also known as two-stage predictor substitution). Estimates from such an approach will be overly precise, as uncertainty in the first-stage regression is not accounted for; however, this over-precision may be slight if the standard error in the first-stage coefficients is low. This can be resolved by the use of a likelihood-based method (Section 4.3.4*) or a bootstrap method, such as that implemented in Stata using the `qvf` command (Section 4.6.2).

As with the ratio IV estimator, in a case-control study it is important to undertake the first-stage regression only in the controls, not the cases (Section 4.1.4). Fitted exposure values for the cases are be obtained by substituting their genetic variants into the first-stage regression model.

Two-stage regression methods with non-linear second-stage regression models (such as with binary outcomes) have been criticized and called "forbidden regressions" [Angrist and Pischke, 2009, page 190]. This is because the non-linear model does not guarantee that the residuals from the second-stage regression are uncorrelated with the instruments [Foster, 1997]. There is current debate about the interpretation and validity of such estimates, especially when the measure of association is non-collapsible.

4.2.3* Non-collapsibility

Several measures of association, including odds ratios, differ depending on whether they are considered conditional or marginal on a covariate. For example in the left half of Table 4.1, the odds ratio of an outcome for exposed versus unexposed individuals is equal to 2 for men and 2 for women. Even under the assumption of no confounding (that the proportion of exposed and non-exposed individuals is the same in both men and women), the odds ratio for a population with equal numbers of men and women is not 2. In contrast, as the example in the right half of Table 4.1 shows, a relative risk is the same whether considered conditional or marginal on sex.

A measure of association, such as an odds ratio or relative risk, would be termed collapsible 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 [Greenland et al., 1999]. The relative risk and absolute risk difference are collapsible measures of association. Odds ratios are generally non-collapsible [Ducharme and LePage, 1986]. This means that the conditional model:

$$\text{logit}(\mathbb{E}(Y|X = x, U = u)) = \beta_0 + \beta_1 x + h(u) \tag{4.15}$$

	Probability of event		Odds	Probability of event		Relative
	Unexposed	Exposed	ratio	Unexposed	Exposed	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	0.139	0.233	1.88	0.175	0.35	2

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

and the structural model:

$$\log(\mathbb{E}(Y|do(X = x))) = \beta_0' + \beta_1 x \tag{4.16}$$

for a logistic model of association cannot in general both be true simultaneously for the same value of β_1.

An odds ratio estimated in an observational study by conventional multivariable logistic regression is conditional on those covariates adjusted for in the analysis. Unless adjustment is made in the instrumental variable analysis, an odds ratio estimated in a Mendelian randomization study is marginal on these covariates. The odds ratio from a ratio or two-stage analysis method is conditional on the IV, but marginal in all other variables, including the exposure itself if it is continuous [Burgess and CCGC, 2013].

This has several consequences for Mendelian randomization. First, the parameter estimated by a two-stage analysis is best interpreted as a population-averaged causal effect. This approximates the effect estimated by a RCT, without adjustment for covariates, where the intervention is to change the distribution of the exposure by increasing the exposure uniformly for all individuals in the population [Stukel et al., 2007]. Generally, a population-averaged causal effect marginal across all covariates is the estimate of interest for a policy-maker as it represents the effect of intervention on the exposure at a population level [Vansteelandt et al., 2011].

Secondly, naive comparison of odds ratio estimates from multivariable regression and from two-stage IV analysis is not strictly valid, as the two odds ratios represent different quantities. The degree of attenuation of the IV estimate depends on the prevalence of the outcome (greater attenuation for common outcomes), the magnitude of the causal effect (greater proportional attenuation for larger effects) and the heterogeneity in individual risks (greater attenuation for more heterogeneous populations). Epidemiological data for coronary heart disease risk has shown attenuation towards unity of 5–14% for odds ratios of around 1.2 to 1.4, although this is likely to be an underestimate of the true attenuation as not all predictors in the risk model are known in practice [Burgess, 2012a]. Attenuation is more substantial when the odds ratio estimate is further from the null.

Thirdly, the apparent inconsistency of estimates from two-stage methods with a non-collapsible measure of association can be explained as manifestation of non-collapsibility. For example, the estimate from a two-stage method with a logistic second-stage model is not in general consistent for the parameter β_1 in equation (4.15) or equation (4.16). This is discussed further below.

Despite the consequences of non-collapsibility, the two-stage estimator with a logistic second-stage model still provides a valid test of the null hypothesis [Vansteelandt et al., 2011].

4.2.4* Adjusted two-stage method

An adjusted two-stage method has been proposed, where the residuals from the first-stage regression of the exposure on the IV are included in the second-stage regression of the outcome on the fitted values of the exposure. This has been referred to as a control function approach [Nagelkerke et al., 2000], or two-stage residual inclusion (2SRI) method [Terza et al., 2008]. 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 adjusted two-stage estimate comes from a second-stage regression additively on $\hat{X}|G$ and $\hat{R}|G$ (or equivalently on X and $\hat{R}|G$). The residuals from the first-stage regression incorporate information on confounders.

If the second-stage regression is linear, as with a continuous outcome, then inclusion of these residuals in the second-stage regression model does not change the estimate, as the residuals are orthogonal to the fitted values. If the outcome is binary, inclusion of these residuals in a second-stage logistic regression model means that the IV estimate will be conditional on these residuals. Numerically, it brings the IV estimate closer to the conditional log odds ratio, the parameter β_1 in the logistic-linear model (equation 4.15) [Palmer et al., 2008]. Some investigators have therefore recommended the adjusted two-stage method when the second-stage regression is logistic on the premise that it is less biased than the unadjusted two-stage method.

Under a particular choice of mathematical model, the adjusted two-stage estimate is consistent for the parameter β_1 [Terza et al., 2008]. However, this mathematical model is unrealistic, and in general the adjusted two-stage estimate is biased for this parameter [Cai et al., 2011]. Further, when the confounders are unknown, as is usual in an IV analysis, it is not clear what variable is represented by the first-stage residuals, and so which covariates the adjusted two-stage estimate is conditional on and which it is marginal across. It is uncertain what odds ratio is being estimated by an adjusted two-stage approach, that is, to what question is the adjusted two-stage estimate the answer. This is in contrast to the unadjusted two-stage method, which consistently estimates an odds ratio which is marginal across all covariates except for the IV itself [Burgess and Thompson, 2012]. We therefore do not recommend adjustment for the first-stage residuals in a two-stage method.

4.3 Likelihood-based methods

The above methods are not likelihood-based and do not provide maximum likelihood estimates, which have the desirable properties of asymptotic unbiasedness, normality and efficiency. So we next consider likelihood-based methods.

4.3.1 Full information maximum likelihood

If we have the same situation as for the two-stage model equations (4.11) and (4.12), such that each individual $i = 1, \ldots, N$ has exposure x_i, continuous outcome y_i and IVs g_{ik} indexed by $k = 1, \ldots, K$, we can assume the following model:

$$\begin{aligned} x_i &= \alpha_0 + \sum_k \alpha_k g_{ik} + \varepsilon_{Xi} \\ y_i &= \beta_0 + \beta_1 x_i + \varepsilon_{Yi} \end{aligned} \tag{4.17}$$

where the error terms $\boldsymbol{\varepsilon} = (\varepsilon_X, \varepsilon_Y)^T$ have a bivariate normal distribution $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. (These error terms differ from those defined in equations 4.11 and 4.12.) The causal parameter of interest is β_1. Correlation between ε_X and ε_Y is due to confounding. We can simultaneously calculate the maximum likelihood estimates of β_1 and each of the other parameters in the model. This is known as full information maximum likelihood (FIML) [Davidson and MacKinnon, 1993].

Confidence intervals can be obtained by the assumption of asymptotic normality of the parameter estimates.

4.3.2 Limited information maximum likelihood

A disadvantage of FIML is that all the parameters in each of the equations are estimated. This means that each of the regression equations has to be correctly specified to give a consistent estimate of β_1. In practice, we are only interested in β_1, and not in the other parameters. In limited information maximum likelihood (LIML), we maximize the likelihood substituting for and profiling out (referred to by economists as 'concentrating out') each of the parameters except β_1.

LIML has been called the 'maximum likelihood counterpart of 2SLS' [Hayashi, 2000, page 227] and gives the same causal estimate as the 2SLS and ratio methods with a single IV. As with 2SLS, estimates are sensitive to heteroscedasticity and misspecification of the equations in the model. Use of LIML has been strongly discouraged by some, as LIML estimates do not have defined moments for any number of instruments [Hahn et al., 2004]. However, use has also been encouraged by others, especially with weak instruments (Section 4.5.2), as the median of the distribution of the estimator is close to unbiased even with weak instruments [Angrist and Pischke, 2009]. With large numbers of IVs (10 or more), standard confidence intervals from the LIML method with weak instruments are too narrow and a correction is needed (known as Bekker standard errors) [Bekker, 1994]. Although this correction is required to maintain nominal coverage levels, the efficiency of the LIML estimator is reduced, and it may be outperformed by a simple allele score approach [Davies et al., 2014].

The LIML estimate can be intuitively understood as the effect β_1 that minimizes the residual sum of squares from the regression of the component of Y not caused by X, $(y_i - \beta_1 x_i)$, on G. Informally, the LIML estimator is the causal parameter for which the component of Y due to confounding is as badly predicted by G as possible.

4.3.3 Bayesian methods

Inference from a similar likelihood model can be undertaken in a Bayesian framework. For each individual i, we model the measured exposure x_i and outcome y_i as coming from a bivariate normal distribution for $(X_i, Y_i)^T$ with mean $(\xi_i, \eta_i)^T$ and variance-covariance matrix Σ. The mean of the exposure distribution ξ_i is assumed to be a linear function of the instruments $g_{ik}, k = 1, \ldots, K$, and the mean of the outcome distribution η_i is assumed to be a linear function of the mean exposure [Jones et al., 2012].

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix}, \Sigma \right) \tag{4.18}$$
$$\xi_i = \alpha_0 + \sum_k \alpha_k g_{ik}$$
$$\eta_i = \beta_0 + \beta_1 \xi_i$$

This model is similar to that in the FIML and LIML methods, except that the causal parameter β_1 represents the causal effect between the true means ξ_i and η_i rather than the measured values of outcome and exposure. The model can be estimated in a Markov chain Monte Carlo (MCMC) framework, such as that implemented in WinBUGS [Spiegelhalter et al., 2003]. The output is a posterior distribution, from which the posterior mean or median can be interpreted as a point estimate, and the 2.5th and 97.5th percentiles as a '95% confidence interval'. Under certain choices of prior, estimates based on the Bayesian posterior distribution are similar to those from the 2SLS or LIML methods [Kleibergen and Zivot, 2003]. With vague priors, the joint posterior distribution is similar to the frequentist likelihood function [Burgess and Thompson, 2012].

An advantage of the Bayesian approach is that no distributional assumption is made for the posterior distribution of the causal parameter. Inference is therefore more robust using weak instruments [Burgess and Thompson, 2012].

4.3.4* Likelihood-based methods with binary outcomes

Maximum likelihood and Bayesian estimates can be estimated with binary outcomes. If we assume a linear model of association between the logit-transformed probability of an event (π_i) and the exposure (a logistic-linear model), and a Bernoulli distribution for the outcome event, as in the

following model:

$$\begin{aligned} x_i &\sim \mathcal{N}(\xi_i, \sigma_X^2) \\ y_i &\sim \text{Bernoulli}(\pi_i) \\ \xi_i &= \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 x_i \end{aligned} \tag{4.19}$$

then the joint likelihood L is given by:

$$L = \prod_{i=1,\ldots,N} \left(\pi_i^{y_i} (1-\pi_i)^{1-y_i} \frac{1}{\sqrt{2\pi}\sigma_X} \{\exp(-\frac{1}{\sigma_X^2}(x_i - \xi_i)^2)\} \right). \tag{4.20}$$

Estimates can be obtained by maximization of the joint likelihood. As all coefficients are simultaneously estimated, this is a full information maximum likelihood (FIML) approach. Alternatively, model parameters can be estimated in a Bayesian framework, obtaining posterior distributions from the model by MCMC methods.

Log-linear models can in principle be estimated in the same way, although care is needed to ensure that the probabilities π_i do not exceed 1 at any point in the estimation.

4.3.5 Comparison of two-stage and likelihood-based methods

In the two-stage methods, the two stages are performed sequentially. The output from the first-stage regression is fed into the second-stage regression with no acknowledgement of uncertainty. In the likelihood-based methods, the two stages are performed simultaneously: the α and β parameters are estimated at the same time. Uncertainty in the first-stage parameters is acknowledged and feedback between the regression stages is possible. The uncertainty in the estimate of the causal parameter β_1 is therefore better represented in the likelihood-based approaches if there is non-negligible uncertainty in the first-stage regression model.

4.4* Semi-parametric methods

A semi-parametric model has both parametric and non-parametric components. Typically semi-parametric estimators with IVs assume a parametric form for the equation relating the outcome and exposure, 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 [Clarke and Windmeijer, 2010].

4.4.1* 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 heteroscedasticity of error distributions and non-linearity in the two-stage structural equations [Foster, 1997; Johnston et al., 2008]. 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

$$\mathbb{E}(Y) = f(X; \boldsymbol{\beta}) \tag{4.21}$$

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

$$\begin{aligned} \sum_i \left(y_i - f(x_i; \boldsymbol{\beta})\right) &= 0 \\ \text{and } \sum_i g_i \left(y_i - f(x_i; \boldsymbol{\beta})\right) &= 0 \end{aligned} \tag{4.22}$$

where the summation is across i, which indexes study participants. In the linear (or additive) case, $f(x_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_i$; in the log-linear (or multiplicative) case, $f(x_i; \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_i)$; and in the logistic case, $f(x_i; \boldsymbol{\beta}) = \text{expit}(\beta_0 + \beta_1 x_i)$; where β_1 is our causal parameter of interest and $\text{expit}(z) = (1 - \exp(-z))^{-1}$, the inverse of the logit function. These two equations can be solved numerically [Palmer et al., 2011b].

GMM estimates are sensitive to the parametrization of the model used. For example, estimates from the estimating equations (4.22) and from:

$$\begin{aligned} \sum_i \left(y_i f(x_i; \boldsymbol{\beta})^{-1} - 1\right) &= 0 \\ \text{and } \sum_i g_i \left(y_i f(x_i; \boldsymbol{\beta})^{-1} - 1\right) &= 0 \end{aligned} \tag{4.23}$$

may be different in finite samples, although they each assume the same structural model between Y and X.

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

$$(\mathbf{y} - f(\mathbf{x}; \boldsymbol{\beta}))^T \mathrm{G} (\mathrm{G}^T \Omega \mathrm{G})^{-1} \mathrm{G}^T (\mathbf{y} - f(\mathbf{x}; \boldsymbol{\beta})) \tag{4.24}$$

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 matrix Ω, efficient estimation is achieved when $\Omega_{ij} = \text{cov}(\varepsilon_i, \varepsilon_j)$ $(i, j = 1, \ldots, N)$, where ε_i is the residual $y_i - f(x_i; \boldsymbol{\beta})$ [Hansen, 1982].

As the estimation of Ω requires knowledge of the unknown $\boldsymbol{\beta}$, a two-step approach is suggested. 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 $\mathrm{G}^T\Omega\mathrm{G} = \sum_i \mathbf{g_i}\mathbf{g_i}^T\varepsilon_i^2$ as $\sum_i \mathbf{g_i}\mathbf{g_i}^T e_i^2$ in a second-stage estimation [Johnston et al., 2008].

4.4.2* Structural mean models

The structural mean model (SMM) approach is another semi-parametric estimation method designed in the context of randomized trials with incomplete compliance [Robins, 1994; Fischer-Lapp and Goetghebeur, 1999]. (Technically, g-estimation is the method by which a structural mean model is fitted, but we refer to the approach as SMM [Robins, 1986; Greenland et al., 2008].) We recall that the potential outcome $Y(x)$ is the outcome which would have been observed if the exposure X were set to x. In particular, the exposure-free outcome $Y(0)|X = x$ is the outcome which would have been observed if we had set X to 0 rather than it taking its observed value of x [Clarke and Windmeijer, 2010]. Conditioning is performed on $X = x$ so that no other variable is changed from the value it would take if $X = x$. 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:

$$\mathbb{E}(Y(x)) - \mathbb{E}(Y(0)|X = x) = \beta_1 x \tag{4.25}$$

and β_1 is taken as the causal parameter of interest. In the context of non-compliance in randomized trials, this is referred to as the 'effect of treatment on the treated' [Dunn et al., 2005].

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. The estimating equations are:

$$\sum_i (g_{ik} - \bar{g}_k)(y_i - \beta_1 x_i) = 0 \qquad k = 1, \ldots, K \tag{4.26}$$

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. With a binary outcome, it is natural to use a log-linear (or multiplicative) SMM:

$$\log \mathbb{E}(Y(x)) - \log \mathbb{E}(Y(0)|X = x) = \beta_1 x \tag{4.27}$$

Due to non-collapsibility of the odds ratio, the logistic SMM cannot be

estimated in the same way, as the expectation $\text{logit}\,\mathbb{E}(Y(x))$ depends on the distribution of the IV [Robins, 1999]. This problem can be addressed by estimating $Y(x)$ assuming an observational model [Vansteelandt and Goetghebeur, 2003]:

$$\text{logit}\,\mathbb{E}(Y(x)) = \beta_{0a} + \beta_{1a}x \tag{4.28}$$

where the subscripts a indicate associational, as well as a structural model:

$$\text{logit}\,\mathbb{E}(Y(x)) - \text{logit}\,\mathbb{E}(Y(0)|X = x) = \beta_{1c}x \tag{4.29}$$

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

$$\sum_i (g_{ik} - \bar{g}_k)\ \text{expit}(\hat{Y}(x) - \beta_{1c}x_i) = 0 \qquad k = 1, \ldots, K \tag{4.30}$$

where $\text{logit}\,\hat{Y}(x) = \hat{\beta}_{0a} + \hat{\beta}_{1a}x$ [Vansteelandt et al., 2011].

We note that the choice of estimating equations presented here is not the most efficient, but leads to consistent estimates [Vansteelandt and Goetghebeur, 2003]. In the general case, the linear (additive) and log-linear (multiplicative) GMM and SMM approaches give rise to the same estimates. This is not true in the logistic case [Clarke and Windmeijer, 2010].

4.4.3* Lack of identification with binary outcomes

An issue with semi-parametric IV estimation in practice is lack of identification of the causal parameter. A parameter in a statistical model is identified if an estimate of its value can be uniquely determined on the basis of the data. For a semi-parametric instrumental variable analysis with a binary outcome, the causal parameter of interest may not be identified; there may be multiple or no parameter values which satisfy the estimating equations [Burgess et al., 2014c]. This is especially likely if the IV is weak (Section 4.5.2). Consequently, estimates and standard errors reported by automated commands for GMM or SMM estimation can be misleading.

It is recommended that investigators wanting to use a GMM or SMM approach should plot the relevant estimating equations for a large range of values of the parameter of interest to check if there is a unique solution. If there is not, this should be reported as an indication that there is a lack of information on the parameter in the data. An alternative estimation technique can be used, such as a two-stage method, but identification will be rely on stronger assumptions.

4.5 Efficiency and validity of instruments

Having discussed the methods for IV estimation, we present statistical approaches to improve the efficiency of estimates, and to test the validity of IVs.

4.5.1 Use of measured covariates

If we can find measured covariates which explain variation in the exposure or a continuous outcome, and which are not correlated with the IV nor on the causal pathway between exposure and outcome, then we can incorporate such covariates into our analysis. In econometrics, such a variable is called an exogenous regressor or included instrument, as opposed to an IV, which is called an excluded instrument [Baum et al., 2003]. This is because the covariate is included in the second-stage regression model for the outcome. Incorporation of covariates generally increases efficiency and hence the precision of the causal estimate. However, it may lead to bias in the causal estimate if the covariate is on the causal pathway between exposure and outcome, or if the analysis model including the covariate is misspecified. In a two-stage estimation, any covariate adjusted for in the first-stage regression should also be adjusted for in the second-stage regression [Wooldridge, 2009]; failure to do so can cause associations between the IV and confounders leading to bias [Angrist and Pischke, 2009, page 189].

4.5.2 Weak instruments

A 'weak instrument' is defined as an IV for which the statistical evidence of association with the exposure is not strong [Lawlor et al., 2008]. An IV is weak if it explains only a small amount of the variation of the exposure, where the amount defined as 'small' depends on the sample size. The F statistic in the regression of the exposure on the IV (also known as the Cragg–Donald F statistic [Baum et al., 2007]) is usually quoted as a measure of the strength of an instrument [Stock et al., 2002]. Although IV methods are asymptotically unbiased, they typically demonstrate systematic finite sample bias, typically in the direction of the observational (confounded) association between the exposure and outcome. The bias of the IV estimate from the two-stage method with a continuous outcome is approximately $1/\mathbb{E}(F)$ of the bias of the observational association, where $\mathbb{E}(F)$ is the expected F statistic from the first-stage regression.

IVs with an F statistic less than 10 are often labelled as 'weak instruments' [Staiger and Stock, 1997]. The value 10 was chosen as this limits the bias of the two-stage IV estimate to 10% of the bias of the observational association. Such characterization of IVs is misleading for several reasons. First, it gives a binary

classification of IVs as either weak or strong based on an arbitrarily chosen threshold F statistic, whereas the true magnitude of bias relates to instrument strength in a continuous way. Secondly, the F statistic is not simply a measure of the intrinsic strength of the IV (unlike the coefficient of determination R^2) as it depends on the sample size. Labelling an IV as a weak instrument leads researchers to think that 'weak instrument bias' is due to an intrinsic property of the instrument, whereas any instrument can be made stronger by increasing the sample size. Thirdly, the measured F statistic in a given dataset is an unreliable guide to the true strength of an instrument, due to the large sampling variability of the F statistic. Fourthly, the use of rules for the *post hoc* selection of data based on measured F statistics can lead to more bias than it prevents [Burgess et al., 2011b]. Fifthly, the threshold was determined based on the 2SLS method, and is not necessarily relevant to other IV methods. Indeed, the F statistic may not even be a reliable measure of instrument strength for obtaining identification in a semi-parametric model [Burgess et al., 2014c].

It is the authors' view that weakness in instruments is best combatted through *a priori* specification of the variable(s) used as IVs in the analysis and careful choice of analysis method. Further advice on weak instruments is given in Chapter 7.

4.5.3 Overidentification tests

When more than one instrument is used, an overidentification test, such as the Basmann test [Basmann, 1960] or Sargan test [Sargan, 1958], can be carried out to test whether the instruments have additional effects on the outcome beyond that mediated by the exposure. Overidentification means that the number of instruments used is greater than the number of exposures measured. The latter is almost always one in Mendelian randomization, so when there is more than one IV, separate causal estimates can be calculated using each IV in turn. The overidentification test assesses whether these IV estimates are compatible, or equivalently whether the IVs have residual associations with the outcome once the main effect of the exposure has been removed [Wehby et al., 2008]. Such a residual association may indicate that at least one of the IVs has a pathway of association with the outcome not via the exposure (such as via another risk factor), meaning that the IV assumptions may be violated.

Overidentification tests are omnibus tests, where the alternative hypothesis includes failure of the IV assumptions for one IV, failure for all IVs, a non-linear relationship between the exposure and outcome, and that different variants identify different magnitudes of causal effect (treatment effect heterogeneity, Section 8.5.1) [Baum et al., 2003]. They generally have low power and so have limited practical use in detecting violations of the IV assumptions [Glymour et al., 2012].

4.5.4 Endogeneity tests

Some applied Mendelian randomization analyses have reported on whether there is a difference between the observational and IV estimates as the primary outcome of interest [Hingorani and Humphries, 2005]. This can be formally tested using the Durbin–Wu–Hausman test [Baum et al., 2003]. This is a test of equality of the observational and IV estimates, where a significant result indicates disagreement between the two estimates. Such a test is known as an endogeneity test (see Table 2.1).

While an informal comparison of the observational and causal estimates may be reasonable, there are several reasons why reliance on an endogeneity test as a primary analysis result is not recommended in practice. If a non-significant result is achieved, it would be fallacious to assume that the exposure was exogenous, that is to assume that the observational association is unconfounded. A non-significant result may simply reflect the limited power of the test. If a significant result is achieved, this does not imply that there is no causal effect. There may be a causal effect, but this may be different in magnitude to the observational association. The conclusion from a significant endogeneity test is that the exposure is endogenous, and so there is confounding. If a researcher believed that there was no (residual) confounding, then they would be content with interpreting the observational association as causal, and IV analysis would be unnecessary. For this reason, it is more appropriate to consider the presence or absence of a causal effect as the subject of investigation [Thomas et al., 2007]. The confidence interval of the causal estimate gives the researcher bounds on the plausible size of any possible causal effect.

4.6 Computer implementation

Several commands are available in statistical software packages for IV estimation, such as Stata [StataCorp, 2009], SAS [SAS, 2004], and R [R Development Core Team, 2011]. We assume that the reader is familiar enough with the software to calculate the ratio method 'by hand' (that is without the use of pre-written software commands). Code is also given below for the estimation of Bayesian models in WinBUGS [Spiegelhalter et al., 2003].

4.6.1 IV analysis of continuous outcomes in Stata

The commands in Stata `ivreg`, `ivreg2`, `ivhettest`, `overid`, and `ivendog`, have been written to implement the 2SLS, LIML and GMM methods, with estimators and tests, including the Cragg–Donald F statistic (weak instruments) and the Sargan statistic (overidentification) [Baum et al., 2003]. The

main command in Stata for IV analysis is `ivreg2`. If the exposure is `x`, the outcome is `y` and the IV is `g`, the syntax for a 2SLS analysis is:

```
ivreg2 y (x=g)
```

The syntax for a LIML analysis is:

```
ivreg2 y (x=g), liml
```

The syntax for a (linear) GMM analysis is:

```
ivreg2 y (x=g), gmm
```

In the output of the `ivreg2` command, several additional results are displayed, as follows:

The underidentification test assesses whether the IV is sufficiently associated with the exposure to give reliable identification of the causal parameter. A parameter is formally identified if the data-generating model corresponds to a unique set of parameter values. Poor identification means that there are multiple parameter values which fit the data well. Underidentification tests are rarely performed in practice, and are unlikely to be useful in Mendelian randomization, as there is typically only one parameter of interest, and underidentification would be reflected in a wide confidence interval for this parameter.

Critical values for the F statistic in determining the potential impact of the weakness of the IV are provided. The values cited are from a simulation study [Stock and Yogo, 2002], where the authors sought to improve on the general arbitrary threshold of 10 for a weak instrument to give more accurate bias and coverage thresholds for different numbers of IVs. With a single IV, there is no threshold limiting relative bias in the 2SLS method. The coverage thresholds cited relate to the coverage of the IV estimate. For example, with a single IV, an F statistic of 16.38 or greater is needed to guarantee that the 95% confidence interval will exclude the true parameter value no more than 10% of the time, compared to the nominal 5%. The coverage thresholds were calculated, however, assuming an unrealistically large correlation between the exposure and the outcome. This means that the coverage levels should be more conservative than the upper bounds cited from Stock and Yogo, although there may well be some undercoverage of confidence intervals from the 2SLS method with weak instruments (Section 4.1.5). The large sampling variability in the F statistic, and the selection bias induced by data-driven procedures based on the measured value of the F statistic (see Chapter 7), mean that the apparent precision of these cited threshold values should not be relied on to protect against bias.

Overidentification tests are described above (Section 4.5.3). With a single IV, an overidentification test is not possible. In the 2SLS analysis, the Sargan test alone is given. In the LIML and GMM analyses, different overidentification tests are computed.

The commands `ivregress 2sls y (x=g)` and `ivreg y (x=g)` give the same estimates as `ivreg2 y (x=g)`, but a more limited output. The command `ivhettest` performs a test of heteroscedasticity of the errors in the second-stage regression. If heteroscedasticity is present, a GMM analysis with robust standard errors is preferred to a 2SLS analysis. The command `overid` gives more information about overidentification tests. The command `ivendog` gives more information about endogeneity tests. The command `qvf` has been written to implement a fast bootstrap estimation of standard errors for IV analysis [Hardin et al., 2003]. Each of these commands can be used with multiple instruments, for example `ivreg2 y (x=g1 g2 g3)`.

4.6.2 IV analyses of binary outcomes in Stata

With a binary outcome and a logistic-linear model, the two-stage estimates can be obtained by the commands:

```
reg x g
predict xhat
logit y xhat, robust
```

where robust standard errors are calculated in the second-stage regression.

Generic estimating equations for GMM or SMM analyses can be solved in Stata using the `gmm` command [Drukker, 2009]. For example, a linear GMM estimate can be obtained using:

```
gmm (y - {beta0} - x*{beta1}), instruments(g)
```

A logistic GMM estimate can be obtained using:

```
gmm (y - invlogit({beta0} + x*{beta1})), instruments(g)
```

A log-linear (multiplicative) GMM estimate can be obtained using the command `ivpois`:

```
ivpois y, endog(x) exog(g)
```

The same log-linear GMM estimate can be obtained using the `gmm` command:

```
gmm (y*exp(-x*{beta1}) - {beta0}), instruments(g)
```

An alternative log-linear GMM estimate can be estimated using:

```
gmm (y - exp({beta0}+x*{beta1})), instruments(g)
```

These GMM models assume the same structural relationship between the exposure and outcome, but give different answers in finite samples (Section 4.4.1*). The first formulation of the log-linear GMM model is equivalent to a log-linear SMM [Palmer et al., 2011b].

Useful notes are available for the estimation of SMMs with a binary outcome [Clarke et al., 2011]. Each of these commands can be used with multiple instruments: for example `gmm (y - beta0 - x*beta1), instruments(g1 g2)`. The command `qvf` can also be used in non-linear cases [Hardin et al., 2003], such as a two-stage analysis with a non-linear second-stage model, to prevent the over-precision of estimates resulting from a sequential regression method. A probit IV model (not considered in this book) can be estimated using the command `ivprobit`.

4.6.3 IV analysis in SAS

The command `proc syslin` in SAS has been written to implement the 2SLS, FIML, and LIML methods:

```
proc syslin data=in 2sls;
 endogenous x;
 instruments g;
 model y=x;
      run;
```

where `2sls` can be replaced by `liml` or `fiml` as appropriate.

4.6.4 IV analysis in R

The R command `tsls` in the library *sem* carries out a 2SLS procedure [Fox, 2006]. Care must be taken as the constant term usually used in regression equations is not included by default. If the exposure is `x`, the outcome is `y` and the IV is `g`, the syntax for a two-stage (2SLS) analysis is:

```
tsls(y, cbind(x, rep(1, length(x))), cbind(g, rep(1, length(g))),
 w=rep(1, length(x)))
```

where `w` are the weights, here set to 1 for each individual. Also available are the function `ivreg` in the *aer* package [Kleiber and Zeileis, 2014], and the *ivpack* package with some additional functions, such as implementation of the Anderson–Rubin confidence intervals [Small, 2014].

In a sequential regression two-stage analysis of a binary outcome in a case-control setting, inference on the controls only (where $Y = 0$) can be made using the `predict` function:

```
g0=g[y==0]
glm(y~predict(lm(x[y==0]~g0), newdata=list(g0=g)), family=binomial)
```

Generic estimating equations for GMM or SMM can be solved in R using the *gmm* package [Chaussé, 2010]; details and sample code are available [Clarke et al., 2011].

4.6.5 IV analysis in WinBUGS

Bayesian analyses can be performed in WinBUGS. The following annotated code can be used with a continuous outcome. We here assume vague priors for all the parameters: Normal(0,10^6) for the regression parameters, Uniform(0,20) for standard deviations, and Uniform(-1,1) for correlations, to mimic a likelihood-based analysis. These could be changed for particular datasets, to be better represent 'non-informative' priors or to alternatively to represent prior information.

```
model {
   beta1 ~ dnorm(0, 1E-6)
   beta0 ~ dnorm(0, 1E-6)
  alpha0 ~ dnorm(0, 1E-6)
for (k in 1:K) {  alpha1[k] ~ dnorm(0, 1E-6) }
     xsd ~ dunif(0, 20)
     ysd ~ dunif(0, 20)
     rho ~ dunif(-1, 1)
 # priors for the parameters
   xtau <- pow(xsd, -2)
   ytau <- pow(ysd, -2)
   tauy <- ytau/(1-pow(rho,2))
 # tauy is the precision of y conditional on x
for (i in 1:N) {
 ksi[i] <- alpha0 + inprod(alpha1[1:K], g[i,1:K])
    x[i] ~ dnorm(ksi[i], xtau)
 eta[i] <- beta0 + beta1 * ksi[i]
 muy[i] <- eta[i] + sqrt(xtau/ytau)*rho*(x[i]-ksi[i])
 # muy[i] is the mean of y[i] conditional on x[i]
    y[i] ~ dnorm(muy[i], tauy)
  } }
```

In the above, the bivariate normal distribution of $(X,Y)^T$ from equation (4.18) has been equivalently replaced by the marginal distribution of X and the conditional distribution of $Y|X=x$ [Burgess and Thompson, 2012].

In a case-control study with a binary outcome and a logistic model of association, the following code can be used (the first P individuals in the dataset are the controls):

```
model {
   beta1 ~ dnorm(0, 1E-6)
   beta0 ~ dnorm(0, 1E-6)
  alpha0 ~ dnorm(0, 1E-6)
for (k in 1:K) {  alpha1[k] ~ dnorm(0, 1E-6)  }
```

```
   xtau <- pow(xsd, -2)
   xsd ~ dunif(0, 20)
for (i in 1:P) {  x[i] ~ dnorm(ksi[i], xtau)  }
 # where P is the number of controls
for (i in 1:N) {
 ksi[i] <- alpha0 + inprod(alpha1[1:K], g[i,1:K])
 logit(pi[i]) <- beta0 + beta1 * ksi[i]
    y[i] ~ dbern(pi[i])  } }
```

4.7 Summary

Methods for IV analysis range from the very simple (calculate the difference between two pairs of numbers and divide one by the other) to the more complicated. The development of complex methods has been driven by the desire to produce efficient estimates, for example by integrating data on multiple IVs, to allow for more flexible modelling assumptions, or to provide robustness against misspecification of modelling assumptions. Each method has its own advantages and disadvantages. The properties of many of these estimators will be discussed in the chapters to come in the specific contexts of weak instruments, binary outcomes, and evidence synthesis.

In the next chapter, we consider examples of the use of Mendelian randomization, focusing particularly on practical aspects of the analyses, such as study design, and their impact on methods.

5

Examples of Mendelian randomization analysis

Having discussed several of the statistical issues regarding Mendelian randomization analyses, in this chapter we present four published examples of the use of Mendelian randomization from the literature, commenting on interesting features of the analysis which help to clarify the methodology and aid readers performing similar investigations.

5.1 Fibrinogen and coronary heart disease

The paper entitled "Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization' " [Keavney et al., 2006] assesses the causal effect of fibrinogen on risk of coronary heart disease (CHD). Fibrinogen is observationally associated with CHD risk, although the magnitude of association attenuates on adjustment for age and sex, and further on adjustment for other covariates such as smoking and body mass index. When the plasma apolipoprotein B/A_1 ratio is additionally adjusted for, the observational association is compatible with the null. However, it may be that some of the variables adjusted for are on the causal pathway between fibrinogen and CHD, and so this may represent an over-adjustment (Section 3.1.4).

5.1.1 Study design

Two approaches are proposed for the assessment of causality using Mendelian randomization. First, the authors analyse individual participant data from a case-control study, the International Studies of Infarct Survival (ISIS). ISIS contains 4685 cases with confirmed myocardial infarction (MI) and 3460 disease-free control participants with measurements of fibrinogen levels. Secondly, they conduct a meta-analysis for the association between a particular genetic variant and the risk of CHD, following a literature-based search for relevant summary genetic estimates. The meta-analysis contains 20 studies

measuring beta-fibrinogen genotypes, including the original study, comprising a total of 12 220 CHD cases and 18 716 controls.

In the context of a disease outcome, a case-control design may be necessary if the outcome of interest is not common, as the power of the analysis depends on the precision of the estimate of the gene–outcome association, which in turn depends on the number of cases. Although measurement of the exposure in the cases is unreliable due to possible reverse causation, the genetic instrumental variable is not affected by the outcome, and so the gene–outcome association can be reliably estimated in a case-control study (Section 2.2.1).

5.1.2 Genetic instruments

A single genetic variant is used as an instrumental variable (IV). This variant is a single nucleotide polymorphism (SNP) in the beta-fibrinogen gene promoter which regulates fibrinogen production, giving some biological credibility to its specific association with fibrinogen, and therefore its validity as an IV. Tests of association between the variant and a range of confounders show no strong associations, except for that with plasma apolipoprotein B/A_1 ratio. Although the p-value of 0.01 is not particularly extreme in view of the multiple comparisons, and would not be judged conventionally significant using a threshold of $p = 0.05$ and a standard Bonferroni correction procedure, the result may indicate a pleiotropic association of the variant with fibrinogen and the plasma apolipoprotein B/A_1 ratio. This would be problematic if the Mendelian randomization estimate indicated a causal relationship, as it would not be possible empirically to distinguish between causal effects of fibrinogen and of the plasma apolipoprotein B/A_1 ratio on CHD risk. Alternatively, it may be that changes in the plasma apolipoprotein B/A_1 ratio associated with the genetic variant are not directly associated with the genetic variant, but occur as a result of the increase in fibrinogen levels. This would mean that the Mendelian randomization analysis was valid, as a clinical intervention on fibrinogen levels would also increase the plasma apolipoprotein B/A_1 ratio. If the plasma apolipoprotein B/A_1 ratio is a mediator on the causal pathway from fibrinogen to CHD risk, it should not be adjusted for in the observational analysis.

It is not possible to differentiate between the association of the variant with the plasma apolipoprotein B/A_1 ratio being a chance finding, evidence of pleiotropy of the genetic variant, or evidence of a causal pathway from fibrinogen.

5.1.3 Statistical methodology

In both the single study and meta-analysis, the causal effect of fibrinogen on CHD risk is assessed, but no causal parameter is estimated. In the single study, the association of fibrinogen levels with the genetic variant is estimated in control participants using linear regression, and the association of CHD

risk with the variant is estimated in the whole study population using logistic regression. In the meta-analysis, summary-level data from each study on the number of cases and controls in each genetic subgroup are used to estimate the association of the variant with the risk of CHD in each study. The study-specific estimates are then combined using a fixed-effect inverse-variance weighted meta-analysis (Chapter 9). A per allele genetic model is used, as this is best supported by the data on fibrinogen levels. In both cases, the result is cited as the risk ratio of CHD per additional variant allele.

5.1.4 Results

The analyses show a null association of the variant with CHD risk, with a narrow confidence interval (CI) for the genetic association with disease risk: risk ratios of 1.06 (95% CI 0.96 to 1.16) per fibrinogen-increasing allele in ISIS alone and of 1.00 (95% CI 0.95 to 1.04) in the meta-analysis. On the basis of this, the authors conclude that "these genetic results provide strong evidence that long-term differences in fibrinogen concentrations are not a major determinant of coronary disease risk".

5.1.5 Commentary

A weakness of the presentation of the results is that a causal estimate of the effect of fibrinogen on CHD risk is not presented. Although the risk ratio estimate of 1.00 (95% CI, 0.95 to 1.04) per additional allele appears to be a small effect, each additional allele is only associated with a small increase in fibrinogen levels (0.14 (standard error 0.024) g/l), meaning that a standard deviation increase in fibrinogen (0.81 g/l, estimated in control participants) could still lead to an approximate 25% increase in CHD risk based on the upper bound of the 95% CI (assuming a log-linear relationship between fibrinogen and the risk of CHD).

5.2 Adiposity and blood pressure

The paper "Does greater adiposity increase blood pressure and hypertension risk? Mendelian randomization using the *FTO/MC4R* genotype" [Timpson et al., 2009] considers the causal effect of adiposity on blood pressure. Adiposity is observationally associated with blood pressure, although there are many potential confounders that may bias the observational estimate. Randomized trials of weight reduction have shown related decreases in blood pressure, but such interventions may additionally affect other variables, such as physical activity and diet. Although the prevalence of obesity has increased over time, the secular trend in blood pressure and hypertension has been in the opposite

direction, leading some to question whether the observational association is in fact causal. Hypertension (severe hypertension) was defined as a systolic blood pressure of over 140 mmHg (over 160 mmHg for severe hypertension), a diastolic blood pressure of over 90 mmHg (over 100 mmHg), or (in both cases) the taking of antihypertensive drugs.

5.2.1 Study design

The authors analyse cross-sectional data on 37 027 unrelated individuals from a population-based study, the Copenhagen General Population Study. All participants are of the same ethnic background (Danish), and were selected to reflect the composition of the general population of Copenhagen.

For an outcome that is a continuous trait rather than a disease outcome, a cross-sectional study is able to provide all the information necessary for a Mendelian randomization experiment without necessitating the expense of following up participants over a period of time. A further advantage of a well-designed population study is increased external validity, whereby an estimate from a Mendelian randomization study represents an effect estimate for a cohort similar to the population on whom an intervention could be performed.

5.2.2 Genetic instruments

Two genetic variants are used as IVs. The SNPs are located in the *FTO* and *MC4R* loci, which have been shown to be associated with body mass index (BMI) in a number of previous studies. The precise functions of the two genetic regions are unknown, although variation in the *FTO* gene is known to be linked with food intake [Wardle et al., 2008].

Although knowledge of the function of genetic variants is not necessary for Mendelian randomization, instrumental variable analysis with variants of unknown function can be problematic to interpret. As the instrumental variable assumptions are scientifically more uncertain, a conclusion that the specific risk factor of interest is in fact the causal agent is less reliable. This is especially true for a risk factor such as BMI, in the same way that a single causal agent in a randomized trial for weight loss is difficult to isolate. Unlike a biomarker such as fibrinogen, there is no single regulatory gene for "BMI production" or for a "BMI receptor". There are additional difficulties in comparing the Mendelian randomization estimate to the effect of a potential clinical intervention, as the intervention and genetic effect on BMI reduction may have different pathways of action. It may be that there is heterogeneity in the proportional effect of changes in BMI on blood pressure and hypertension as instrumented by different variants resulting from differences between pathways (treatment effect heterogeneity). For example, if there were several genetic variants used in the analysis, it is possible for some of the variants to be associated with changes in BMI that do affect blood pressure, and some to be associated with changes that do not.

5.2.3 Statistical methodology

The causal effect of adiposity on blood pressure is estimated using the generalized method of moments (GMM). Adiposity is represented by 'relative BMI', calculated as the ratio of an individual's observed BMI to predicted BMI from a linear regression model on age, sex and height. Results are also calculated using the two-stage least squares (2SLS) and limited information maximum likelihood (LIML) methods; similar results are obtained from each method. The observational and IV estimates of association are compared using a Durbin–Wu–Hausmann test of the equality of the observational and IV estimates. As we discuss in Section 4.5.4, we do not advocate such tests, as there is a multitude of reasons for the estimates to be different unrelated to the question of causality, and neither a significant nor a non-significant finding is directly interpretable as evidence for or against a causal effect.

5.2.4 Results

The IV analysis shows a positive causal effect of BMI on blood pressure and hypertension of similar magnitude to the observational association. For example, the estimate for the increase in systolic blood pressure associated with a 10% increase in BMI is 2.75 mmHg (95% CI 2.62 to 2.88) from the observational analysis with adjustment for age, sex and height, and 2.54 mmHg (95% CI 2.39 to 2.69) with further adjustment for socio-behavioural factors. The corresponding estimate of the increase in systolic blood pressure caused by a 10% increase in BMI from the IV analysis is 3.85 mmHg (95% CI 1.88 to 5.83). The *FTO* SNP has statistically robust associations with BMI (1.18% [95% CI 0.96 to 1.41] increase in BMI on a multiplicative scale per additional allele) and with blood pressure (0.63 mmHg [95% CI 0.33 to 0.93] increase in systolic blood pressure per additional allele), whereas the *MC4R* SNP has a smaller magnitude of association with BMI (0.78%, 95% CI 0.53 to 1.04), and an association with blood pressure compatible with the null (0.20 mmHg, 95% CI -0.14 to 0.54). This may be due to the *MC4R* SNP's reduced association with BMI and the statistical uncertainty in the association estimates, but it may reflect heterogeneity of the causal effects identified by the two variants.

The association of the *FTO* SNP with severe hypertension does not fully attenuate on adjustment for BMI: attenuation from an odds ratio of 1.07 (95% 1.04 to 1.11) on adjustment for age and sex, to 1.07 (95% 1.03 to 1.11) on additional adjustment for socio-behavioural factors, and to 1.04 (95% 1.01 to 1.08) on additional adjustment for log(BMI). Although a complete attenuation is not expected, the limited attenuation suggests that the causal effect of adiposity on hypertension may not simply be explained as a function of BMI.

The Durbin–Wu–Hausmann tests for each variant are not significant, indicating no difference between the observational and IV estimates beyond that compatible with chance.

5.2.5 Commentary

Although the Mendelian randomization analysis suggests that adiposity is causally associated with blood pressure, the unknown function of the genetic variants limits the certainty of the conclusions that can be drawn.

5.3 Lipoprotein(a) and myocardial infarction

The paper "Genetically elevated lipoprotein(a) and increased risk of myocardial infarction" [Kamstrup et al., 2009] examines the causal effect of lipoprotein(a) [denoted lp(a)] on the risk of myocardial infarction (MI). Lipoprotein(a) is an assembly of a lipid, essentially a low-density lipoprotein (LDL) particle, and a protein, known as apolipoprotein(a). Concentrations of lp(a) vary widely between individuals and are highly heritable.

5.3.1 Study design

The authors analyse data from three related studies of Danish participants: a prospective study with 16 years of follow-up, the Copenhagen City Heart Study, comprising 9867 participants with genetic data of whom 4514 have a lp(a) plasma level measurement and 599 suffered a MI event during the follow-up period; a cross-sectional study, the Copenhagen General Population Study, comprising 29 388 participants with genetic data of whom 5543 have a lp(a) plasma level measurement and 994 suffered a MI event in a defined period prior to study entry; and a case-control study, the Copenhagen Ischemic Heart Disease Study, comprising 1231 participants with genetic data and a MI event, and 1230 matched controls taken from the Copenhagen City Heart Study (which reduces the effective size of that study to 8637 participants).

By combining evidence from prospective, cross-sectional and case-control designs, the advantages of each approach are exploited. The prospective study measured lp(a) levels at a range of timepoints, enabling assessment of the long-term associations of genetic variation. The cross-sectional study is the simplest study design, enabling assessment of the genetic association with the exposure in a large population. The case-control study design has known potential weaknesses, including selection bias, but enables more precise estimation of the genetic association with the outcome in a sample enriched for cases. Although lp(a) levels were not measured in all participants, this does not invalidate findings of the Mendelian randomization experiment, and may even be a worthwhile design strategy if the exposure is difficult or expensive to measure (see Section 8.5.2).

5.3.2 Genetic instruments

In this study, the genetic variant is not a SNP, but a copy number variant in the *LPA* gene, the kringle IV type 2 (KIV-2) size polymorphism. Individuals have a variable number of repeating sections of DNA known as kringle repeats, and this number correlates inversely with lp(a) concentration. There is good biological plausibility for the use of the polymorphism as an IV. (The IV in kringle IV type 2 is the Roman numeral 4, rather than the abbreviation for instrumental variable.) While the two variants in the previous example explained less than 1% of the variation in BMI, the KIV-2 polymorphism here explains more than 20% of the variation in lp(a).

5.3.3 Statistical methodology

Two approaches are taken to assess and estimate the causal effect of lp(a) on MI risk. First, the association between the IV and MI risk is assessed in each of the datasets. To address potential non-linearity, the IV is defined by dividing the population into quartiles based on the number of kringle repeats. In the prospective study, the association is assessed using Cox proportional hazards regression with adjustment for a range of covariates. In the cross-sectional and case-control studies, logistic and matched logistic regression are used. Adjustment is made for a limited set of covariates which would not be thought to be affected by potential reverse causation, such as age, sex and diabetes status. Secondly, a formal IV method is conducted in the prospective study only, using the average level of lp(a) and the risk of MI in the top and bottom quartiles of the IV to construct a ratio estimate. Confidence intervals are evaluated using Fieller's theorem (Section 4.1.5).

5.3.4 Results

The analyses show a positive causal effect of lp(a) on MI risk. The odds ratios of MI in the quartiles of the IV (fourth quartile is reference group) were 1.3 (95% CI, 1.1 to 1.5) in the first quartile, 1.1 (95% CI, 0.9 to 1.3) in the second quartile, and 0.9 (95% CI, 0.8 to 1.1) in the third quartile in the Copenhagen General Population Study ($p = 0.005$ for trend), and 1.4 (95% CI, 1.1 to 1.7), 1.2 (95% CI, 1.0 to 1.6), and 1.3 (95% CI, 1.0 to 1.6) in the Copenhagen Ischemic Heart Disease Study ($p = .01$ for trend). In the Copenhagen City Heart Study, the IV estimate for the hazard ratio (HR) of MI per doubling of lp(a) (HR 1.22, 95% CI 1.09 to 1.37) is considerably larger than the observational estimate (HR 1.08, 95% CI 1.03 to 1.12). This finding, which was replicated in a similar study [Clarke et al., 2009], may reflect the increased effect of lifelong differences in lp(a) levels, similar to that observed for low-density lipoprotein cholesterol (LDL-C) (Section 6.2.1). It also may result from the association of the KIV-2 polymorphism with both the concentration of lp(a) and the lp(a) particle size, which is also implicated

as a potential risk factor for MI. In the absence of further evidence, it is difficult to disentangle these two variables.

5.3.5 Commentary

A limitation of the interpretation of the IV estimate is the non-linear association of the number of kringle repeats with lp(a) levels. The IV estimate should be interpreted as a population-averaged effect, comparing genetic subgroups which have different average levels of the exposure. With non-linear relationships, the IV estimate does not necessary represent the effect of intervening on lp(a) for an individual (Section 11.1.2).

5.4 High-density lipoprotein cholesterol and myocardial infarction

The paper "Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study" [Voight et al., 2012] examines the causal effect of high-density lipoprotein cholesterol (HDL-C) on risk of MI. As a proof of concept, the causal effect of LDL-C on risk of MI is also assessed.

5.4.1 Study design

The authors analyse individual participant data from six prospective studies and 14 cross-sectional studies, comprising 20 913 MI cases and 95 407 controls, although assessment of the assumptions for IV analysis is performed in a larger set of studies.

5.4.2 Genetic instruments

Two approaches are proposed for the assessment and estimation of the causal effect of HDL-C on MI risk. First, a single SNP is used as an IV. This SNP is a loss-of-function coding variant at the endothelial lipase gene which has known functional association with HDL-C concentration, and does not show any association with LDL-C or triglycerides in the dataset ($p > 0.05$). Secondly, an allele score (or gene score, Section 8.2) is used, comprising 14 variants associated with HDL-C ($p < 5 \times 10^{-8}$), but not associated with LDL-C or triglycerides ($p > 0.01$). For comparison, an allele score comprising 13 variants associated with LDL-C, but not HDL-C or triglycerides, is also constructed. The reason for the two approaches is that the first is more scientifically rigorous, as the function of the variant used as an IV is known, whereas the second gives more statistical power, as the allele score explains more of the

variation in the exposure than any of the constituent variants individually (bias–variance trade-off). Another practical reason for including both analyses is that the second analyses is performed in a smaller subset of participants, comprising 12 482 MI cases and 41 331 controls, due to missing genetic data on one or more variants (Section 8.4).

The dilemma between only including genetic variants where there is strong evidence of their validity as instrumental variables, risking an underpowered estimate, and including all variants even if their function is not fully known, risking a biased estimate, is an example of a bias–variance trade-off. A sensible compromise in practice is to present the estimate using fewer "safer" variants as the primary analysis result, acknowledging the statistical imprecision in the estimate, and to present the estimate using more variants as a secondary analysis result, acknowledging both the statistical imprecision and the scientific uncertainty in the assumptions necessary to interpret the estimate as a causal effect.

5.4.3 Statistical methodology

In the first approach using a single variant, causal estimates from each of the prospective studies are calculated using the `qvf` command in Stata to fit two-stage logistic models with robust standard errors. In two of the studies, a two-stage method is employed with sequential regression using generalized estimating equations in the first-stage of the analysis, to account for related individuals. These study-level causal estimates are combined in a fixed-effect inverse-variance weighted meta-analysis. In the second approach, a weighted allele score is constructed for both HDL-C and LDL-C using coefficients as weights for the association of each variant with the exposure of interest taken from a large meta-analysis. The association of the allele score with MI case status is assessed using logistic regression in the cross-sectional studies. The data source for the weights is not entirely independent from the data under analysis, as some studies are included in both analyses (Section 8.2.1).

5.4.4 Results

From observational epidemiology, the expected odds ratio (OR) for each variant allele in the endothelial lipase gene is 0.87 (95% CI 0.84 to 0.91). This is obtained by triangulation of the observed estimate of the association of HDL-C on MI risk from multivariable adjusted logistic regression with the observed genetic association of the variant with HDL-C (Section 3.3.3). However, the variant is not associated with risk of myocardial infarction (OR 0.99, 95% CI 0.88 to 1.11). With the allele score, the expected OR for a 1 standard deviation increase in HDL-C from observational epidemiology (OR 0.62, 95% CI 0.58 to 0.66) is not compatible with the estimated OR for a 1 SD increase in HDL cholesterol from Mendelian randomization (OR 0.93, 95% CI 0.68-1.26). For a 1 standard deviation increase in LDL-C, the observational epidemiology

(OR 1.54, 95% CI 1.45 to 1.63) and Mendelian randomization (OR 2.13, 95% CI 1.69 to 2.69) estimates are directionally concordant.

The authors conclude that "some genetic mechanisms that raise plasma HDL-C do not seem to lower risk of myocardial infarction". This tentative conclusion reflects the limited power of the single SNP analysis, where the CIs for the causal effect and the observational estimate substantially overlap, and the limited knowledge of the specific function of the allele score, which may contain variants not exclusively or not directly associated with HDL-C.

5.4.5 Commentary

Aside from the limitations of the conclusions stated above, this paper demonstrates the statistical difficulty of applied Mendelian randomization analysis where the studies under analysis are heterogeneous. Although Mendelian randomization investigations can be undertaken in a number of study designs, differences between studies and specific features of each study may make integrated analysis of the entirety of the data available challenging.

The authors choose a pragmatic approach, combining a more conservative analysis using a single genetic variant with a more speculative analysis using an allele score. This is contrasted with a parallel analysis of LDL-C, which provides plausibility of the allele score approach, as a positive causal effect of LDL-C on MI risk is estimated.

5.5 Discussion

A question of interpretation relating to each of these analyses, and to Mendelian randomization more widely, is how much weight of evidence to attach to the result of a Mendelian randomization investigation. In a hierarchy of evidence, Mendelian randomization has been advocated as providing "critical evidence" on exposure–outcome relationships [Gidding et al., 2012]. However, the true weight of evidence in each case depends strongly on the plausibility of the instrumental variable assumptions for the genetic variants. If the function of the genetic variants is poorly understood, then a causal conclusion is in doubt, particularly if there are multiple genetic variants and there is little consistency in the causal effect estimates using each of the variants. Equally, if the genetic variants explain a small proportion of the variance in the exposure, then a Mendelian randomization investigation using those variants will be inconclusive unless the sample size is very large.

For exposures which are biomarkers, variants can be employed as IVs which are located in the gene coding the biomarker (such as for fibrinogen in the example above). For exposures which are complex multifactorial traits, such as body mass index and blood pressure, the association between the genetic

variants and exposure is less proximal, giving more opportunities for violations of the IV assumptions. A non-null Mendelian randomization estimate is indicative that genetic predictors of the exposure are also associated with the outcome, but there may be an alternative causal pathway other than that through the exposure of interest. An analogous situation is inferring a causal effect of a specific biomarker based on a pharmaceutical intervention with multiple effects, such as those of statins on lipid fractions and inflammation markers. The existence of pleiotropic associations of variants is particularly likely if a large number of variants is included in the analysis.

In conclusion, the reliability of the findings from a Mendelian randomization study depends heavily on both the validity of the variant(s) used as an IV, and the power of the analysis to detect a clinically relevant causal effect.

Before further considering the statistical properties of IV estimators, in the next chapter we consider a more fundamental question: what does a Mendelian randomization estimate represent?

6

Generalizability of estimates from Mendelian randomization

In the previous chapters, we have discussed the meaning of causation and presented methods and examples of estimating causal effects using instrumental variables (IVs). In this chapter, we consider the interpretation of causal effects assessed and estimated in Mendelian randomization, and address the question of under what circumstances a Mendelian randomization estimate may be a reliable guide to the effect of an intervention on the exposure of interest in practice.

6.1 Internal and external validity

From the first discussions of Mendelian randomization, researchers have emphasized that the assumptions leading to the assertion of a causal relationship may be invalid for many genetic variants. Violations in the assumptions of no direct effect of the genetic variant on the outcome or of no association with a confounding risk factor may occur for several reasons, as discussed in Chapter 3. Such violations of internal validity can potentially lead to misleading conclusions. An aspect of Mendelian randomization which is less well appreciated is the issue of external validity. If the IV assumptions about the genetic variant are true and a valid estimate is made which corresponds to a causal effect, what questions are raised in generalizing this estimate to an experimental context? For example, is the estimate of lowered risk derived from considering genetically reduced levels of cholesterol the same as the lowered risk conferred by an intervention that reduces levels of cholesterol?

Mendelian randomization is different from a randomized trial in a fundamental way which impacts on questions of external validity [Rothwell, 2010]. In a randomized trial, the intervention applied to the treatment group is usually identical or similar to the intervention which is proposed to be applied in clinical practice. In Mendelian randomization, the "intervention" leading to differences between genetically-defined subgroups within the study is the presence of a genetic variant. The question of external validity is whether the

causal effect due to the change in the exposure as a result of the presence of the genetic variant is similar to the causal effect due to the proposed intervention on the exposure. There are several reasons why these effects may be unequal, as we now discuss.

6.1.1 Time-scale and developmental compensation

First, the presence or absence of the genetic variant in an individual is determined at conception. This means that the Mendelian randomization estimate represents the result of a life-long difference in the exposure between the genetic subgroups [Davey Smith, 2006]. In contrast, most clinical interventions are performed on mature individuals. For some exposures, an individual may develop compensatory mechanisms in response to long-term elevated (or lowered) levels of the exposure, known as canalization (Section 3.2.3).

Secondly, it may be that a stage of disease progression is irreversible. There may be no intervention on the exposure in a mature cohort which can imitate the genetic effect. This may be especially relevant if the genetic change in the exposure affects intra-uterine or early-stage development.

6.1.2 Usual versus pathological levels

Secondly, the genetic variant would be expected to affect average or "usual" levels of the exposure. This is often the target of interest for epidemiologists interested in disease prevention. Mendelian randomization has a particular role to play here, as typically life-long randomized trials affecting usual levels of exposures cannot be undertaken. However, Mendelian randomization studies are unlikely to be informative about the acute response behaviour of an individual to a stimulus, such as a sudden large increase in an inflammation biomarker. It is plausible that long-term elevated average levels of an exposure for an individual do not affect the outcome, but acute response of the exposure does. The efficacy of short-term targeted interventions on pathological levels of an exposure cannot be validly assessed by a Mendelian randomization approach.

An example is that of C-reactive protein (CRP). Genetic variants which are associated with usual levels of CRP have been used to assess the causal effect of long-term elevated average levels of CRP on cardiovascular risk [Elliott et al., 2009; CCGC, 2011]. Although the causal effect of CRP on cardiovascular risk appears to be null, this does not preclude the efficacy of a therapeutic intervention on acute levels of CRP.

6.1.3 Extrapolation of small differences

Thirdly, the change in an exposure due to genetic variants is generally small. For evolutionary reasons, genetic variants associated with substantial changes

in clinically relevant exposures are uncommon. Most genetic variants which have been used in Mendelian randomization studies have explained in the region of 1 to 4% of the variation in the exposure [Schatzkin et al., 2009; Davey Smith, 2011]. If the target of interest for the epidemiologist is an intervention lowering (or raising) the exposure uniformly by a small amount for everyone in the population, then a Mendelian randomization study may provide a relevant estimate of the effect of the intervention. However, if the proposed intervention effect is more substantial, then the Mendelian randomization estimate relies on extrapolation beyond the genetic change in the exposure observed. Estimates relying on a linear assumption for the effect of the exposure on the outcome may not be valid; moreover this assumption may not be testable from empirical data.

6.1.4 Different pathways of genetic and intervention effects

Fourthly, the genetic variant and the proposed intervention will not, in general, have the same specific mechanism of effect on the exposure. The genetic change in the exposure may be associated with another variable, as in the case of a variant in the *FTO* gene which has been used to study body mass index (BMI) [Brennan et al., 2009]. The effect of variation in the *FTO* gene on BMI is not direct; rather the genetic variant affects satiety, which in turn affects BMI [Wardle et al., 2008]. An intervention on BMI which is not based on reducing food intake may have a different effect on the outcome to the estimate from a Mendelian randomization study using a variant in the *FTO* gene as an IV.

Equivalently, the effect of the intervention may not be limited to the exposure of interest. For example, bariatric surgery aimed at reducing BMI may also result in dietary and lifestyle changes. It is difficult to assess which changes in covariates are direct results of a decrease in BMI and so are on the causal pathway from BMI to disease, and which are separate consequences of the intervention. Even when both the genetic change and proposed intervention specifically target the exposure, it may be that they are on different biological, biochemical or physiological pathways, and so the genetic and clinical changes in exposure may affect the outcome in different ways and to different extents.

6.1.5 Differences in populations

Fifthly, the genetic variant potentially affects all members of a population. If the proposed intervention is to be made across the whole population, then Mendelian randomization using a population-based cohort may give a valid estimate of its potential effect. However, if the intervention is intended to be made in a particular subpopulation, it may not be possible to choose a cohort for a Mendelian randomization study which would give a relevant estimate. For example, an intervention on blood pressure may only be applied to those with clinically-determined hypertension, whereas a genetic variant associated with blood pressure would potentially affect the whole population.

6.2 Comparison of estimates

We give some examples to illustrate the differences between Mendelian randomization estimates and those from other epidemiological approaches, such as effect estimates from randomized controlled trials (RCTs), and observational associations from multivariable adjusted regression models.

6.2.1 Cholesterol and coronary heart disease

Coronary heart disease (CHD) is the result of a build-up of atheromatous plaques in the coronary arteries. A major component of such plaques is cholesterol, and low-density lipoprotein cholesterol (LDL-C) is an established causal risk factor for CHD. We here use the available literature to assess the magnitude of the effect of LDL-C on CHD risk as estimated from Mendelian randomization, and from RCTs where statin drugs have been used as a clinical intervention to lower LDL-C.

A recent meta-analysis of genome-wide association studies reported five SNPs associated with LDL-C, but not with high-density lipoprotein cholesterol (HDL-C) nor triglycerides [Waterworth et al., 2010]. Table 6.1 gives the SNPs and relevant genes, the estimates of association of each SNP with log-transformed LDL-C and risk of CHD, and estimates using each SNP of the causal odds ratio of CHD per 30% decrease in LDL-C using the ratio method (Section 4.1). We note that this relies on a log-linear assumption of the effect of log(LDL-C) on CHD risk, and between eight- and twenty-fold extrapolation of the genetic effects on log(LDL-C). Although further SNPs associated with LDL-C are known, these five were chosen as they represent variants with known strong associations with LDL-C, where there is some biological knowledge to justify the assumption of the specific effect of the SNP on LDL-C. The dose–response relationship between the genetic associations with LDL-C and with CHD risk is noted; this gives plausibility that LDL-C is a causal risk factor for CHD risk (Table 3.1).

Odds ratio estimates for each SNP individually range from 0.27 to 0.45. If we assume that the five estimates of causal effect in Table 6.1 are independent, then a fixed-effect inverse-variance weighted meta-analysis method (see Section 9.4.1) gives a combined odds ratio of 0.33 (95% CI 0.24 to 0.46) [Thompson et al., 2005]. The estimates will not be strictly independent, as they are derived from the same data and affect related pathways, but as the SNPs are on different chromosomes and therefore independently distributed, the correlation between estimates can reasonably be assumed to be small [Burgess et al., 2013].

In comparison, RCTs of statins have given lesser estimates of the benefits of reducing LDL-C levels. A meta-analysis of 9 trials of the effect of statin use on CHD comprising 69 139 participants with 6406 CHD events gave a

SNP (relevant gene)	Per allele change in log(LDL-C) (SE)	Per allele odds ratio of CHD (95% CI)	Odds ratio of CHD per 30% decrease in LDL-C (95% CI) [1]
rs11206510 (*PCSK9*)	0.026 (0.004)	1.07 (1.01–1.13)	0.40 (0.15–0.85)
rs660240 (*SORT1*)	−0.044 (0.004)	0.85 (0.80–0.90)	0.27 (0.15–0.44)
rs515135 (*APOB*)	−0.038 (0.004)	0.90 (0.85–0.96)	0.37 (0.19–0.66)
rs12916 (*HMGCR*)	−0.023 (0.003)	0.94 (0.90–0.99)	0.38 (0.16–0.80)
rs2738459 (*LDLR*)	−0.018 (0.004)	0.96 (0.89–1.03)	0.45 (0.07–1.95)

TABLE 6.1

Association of five SNPs with log-transformed low-density lipoprotein cholesterol (LDL-C) and coronary heart disease (CHD) risk. Causal estimates of odds ratio for 30% reduction in LDL-C on coronary heart disease from Mendelian randomization using each SNP in turn.

[1] A 30% decrease in LDL-C is equivalent to a change in log(LDL-C) of -0.357.

relative risk of 0.73 (95% CI 0.70 to 0.77) based on a reduction of around 30% in LDL-C over an average follow-up time of at least 3 years [Cheung et al., 2004]. A more focused meta-analysis examining the effect of statin use for primary disease prevention, comprising around 27 969 individuals without a history of coronary heart disease with 1677 events, gave a similar relative risk of 0.72 (95% CI 0.65 to 0.79) over 1.5 to three years' follow-up [Taylor et al., 2013]. The data on the genetic variants, together with the combined Mendelian randomization estimate and the proportional estimate of the effect of statins (assuming the relative risk of 0.73 approximates an odds ratio of the same magnitude) are displayed in Figure 6.1.

The Mendelian randomization estimate of the effect of LDL-C reduction is greater (further from the null) than the estimate from RCTs of LDL-C reduction using statins. It is known that the effect of statins in reducing CHD increases over time [Law et al., 2003]. As atherosclerosis is a chronic condition which develops progressively, it is not surprising that the estimates of the effect of the life-long lowering of LDL-C associated with the SNPs considered corresponds to a greater proportional change in cardiovascular risk than the effect on LDL-C due to statin usage. Further possible reasons for differences between the estimates include the non-specific action of statins, which also reduce inflammatory response [Davignon and Laaksonen, 1999]. However, any effects of statins on inflammatory response may further lessen the causal role of low-density lipoprotein cholesterol, and make the contrast with the genetic effects more extreme.

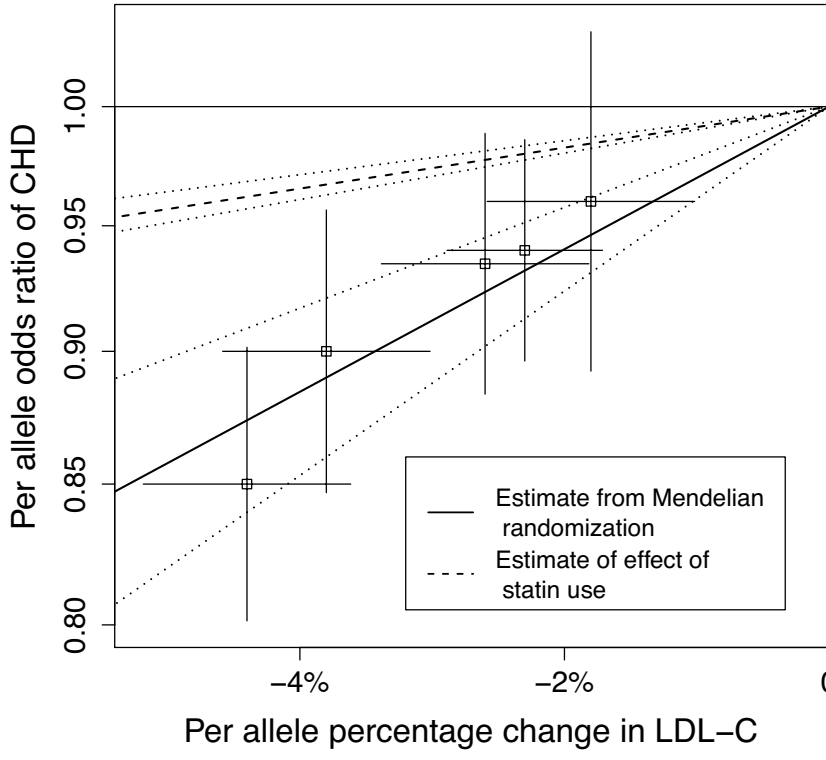


FIGURE 6.1
Estimates of percentage change in low-density lipoprotein cholesterol (LDL-C) and odds ratio of coronary heart disease (CHD) per LDL-C decreasing allele for five SNPs (point estimates with 95% confidence intervals), plus estimate of causal effect of LDL-C on CHD risk from Mendelian randomization using all 5 SNPs (solid line) with 95% confidence interval (dotted lines). Proportionate effect from meta-analysis of statin use on CHD risk in RCTs (dashed line with dotted lines for 95% confidence interval) is displayed for comparison.

6.2.2 Blood pressure and coronary heart disease

A similar example can be observed in the association between blood pressure and CHD. An allele score associated with a 1.6mmHg decrease in systolic blood pressure is associated with a odds ratio for CHD of 0.91 (95% CI 0.89 to 0.92) [Ehret et al., 2011]. Assuming a log-linear association, this corresponds to an odds ratio of 0.55 (95% CI 0.47 to 0.61) for a 10mmHg decrease in systolic blood pressure, compared to the relative risks from a meta-analysis of 0.78 (95% CI 0.73 to 0.83) in clinical trials and 0.75 (95% CI 0.73 to 0.77) in cohort studies [Law et al., 2009].

The Mendelian randomization estimate of the effect of blood pressure lowering is greater (further from the null) than the estimates from both RCTs and observational studies. However, unlike in the lipids example above, the mechanisms of the effects of all the 29 SNPs included in the allele score are not well-known. Hence there is a possibility of pleiotropic effects (or equivalent, such as linkage with other variants) leading to violation of the Mendelian randomization assumptions and lack of internal validity of the causal estimate (Section 3.2.3). However, the pleiotropic associations with alternative risk factors would have to be reasonably strong to explain a sizeable proportion of the reduction in CHD risk [Martens et al., 2006]. As none of the variants in the allele score are known to be strongly associated with other known CHD risk factors, it would seem that the reduction in CHD risk is most plausibly due to the effect of the blood pressure reduction and not due to other factors.

6.3 Discussion

External validity in epidemiology is often thought of in terms of generalizability to a population other than the one considered in the original study [Dekkers et al., 2010]. Although variation in populations may cause some difficulties, the differences between the change in exposure levels associated with natural genetic variation and with any proposed clinical intervention on the exposure lead to inescapable problems in generalizing Mendelian randomization estimates to clinical questions of interest.

Mendelian randomization is a useful tool for exploring causal relationships between modifiable exposures and outcomes of interest. It is one of the few methodologies that can aid the selection of targets for therapeutic intervention. However, it would be misleading to assume that the estimate from a Mendelian randomization study gave the definitive answer to every question of causal relevance of an exposure. Mendelian randomization estimates are especially relevant when the effect of interest is that of a long-term population-based intervention; otherwise, although a Mendelian randomization approach

may be qualitatively informative, the quantity estimated may not correspond to the clinical effect of interest.

6.3.1 Using Mendelian randomization in drug assessment

Questions of generalizability of results are important when using Mendelian randomization to prioritize or de-prioritize targets for drug development, especially in the context of the primary prevention of disease. A considerable proportion of large-scale and expensive clinical trials of drugs targeting suspected novel mechanisms of action fail to demonstrate efficacy. A prudent approach to drug development would be to only go forward with research on targets where there is evidence on the causal nature of the exposure and/or mechanism from human genetics [Plenge et al., 2013]. Where suitable genetic variants for the application of Mendelian randomization on a given exposure are known and available in a large enough sample, assessment of the association between the variants and the disease outcome, and consequently of the causal effect of the exposure on the outcome, is simple, quick, and relatively inexpensive to perform.

Association between a relevant genetic variant affecting the exposure and the outcome may be taken as evidence for the potential efficacy of a drug affecting the exposure pathway. However, absence of evidence for such an association does not necessarily imply lack of efficacy. A drug which blocks a particular biological pathway may have a profound effect on downstream markers, which may lead to a substantially different effect on outcome compared to the slight changes associated with a genetic variant. Although we may expect Mendelian randomization in many circumstances to provide a good qualitative indication of the efficacy of clinical intervention, the magnitude of the Mendelian randomization estimate will not necessarily be a reliable guide to the potential benefit of a drug. Additionally, in many cases the drug will be aimed at secondary disease prevention or targeted at a particular population group (such as individuals who have high or low levels of the exposure) rather than at the general population. As the randomization of genetic variants is valid only in the population, testing the genetic association with the disease in a sample population chosen according to their disease status or exposure value may lead to misleading inference (ascertainment bias, Section 3.2.5).

The Mendelian randomization paradigm can also aid in target reassessment (drug repositioning). If an existing drug has a genetic variant that mimics its effect, then an association of the variant with another outcome may indicate that the drug is also an effective treatment for that outcome. For example, the effects of anakinra, an interleukin-1 receptor antagonist and licensed treatment for rheumatoid arthritis, can be assessed for a range of further auto-immune diseases by considering the association of variants in the *IL1RN* with those disease outcomes. Equally, genetic associations can inform the assessment of mechanism-associated safety. For example, a variant in the *GCKR* gene is associated with lower plasma glucose levels, but also with

higher triglyceride levels [Beer et al., 2009]. This observation may suggest that additional monitoring of triglyceride levels would be advisable in clinical trials of glucokinase activators.

6.3.2 Using Mendelian randomization in drug discovery

'Reverse Mendelian randomization' can be used when a SNP is found in genome-wide association data to be associated with a disease outcome, but the mechanism for the association is not known. An exposure is sought which is associated with the variant and could explain the gene–disease association. This concept is closely related to that of functional genomics. For example, associations between variants in the *HMGCR* gene and risk of coronary heart disease are indicative of the causal role of low-density lipoprotein cholesterol in cardiovascular pathogenesis, and also point towards the potential efficacy of drugs to inhibit HMG-CoA reductase (statins). This approach should provide a fruitful source of targets for ongoing pharmacological research, and has already been used successfully in the discovery of the PCSK9 enzyme for cholesterol lowering (a variant in the *PCSK9* gene having been previously shown to be associated with CHD risk [Cohen et al., 2006]). PCSK9 inhibitors have already been demonstrated to be effective in lowering LDL-C [Robinson et al., 2014] and lipoprotein(a) [Raal et al., 2014], and phase III trials for the secondary prevention of cardiovascular endpoints were underway at the time of writing [Farnier, 2013].

6.3.3 Relevance of causal estimation in Mendelian randomization

As has been emphasized throughout this book, Mendelian randomization investigations can assess a causal relationship, or estimate a causal effect. The arguments in this chapter suggest that the magnitude of causal effect estimates using Mendelian randomization should not be taken too literally. While they provide some indication of the potential relevance of an exposure, the direction of the causal effect and whether it is compatible or not with the null are more important.

One reason for this is that the true causal risk factor may be difficult to define, and so the measured exposure may only be a surrogate (proxy) measure of the underlying risk factor. For example, in a Mendelian randomization analysis of the causal effect of BMI, genetic variants associated with BMI are likely to be associated with the outcome by causal pathways via other adiposity-related variables. In this case, formally the IV assumptions are violated [Glymour et al., 2012]. However, if the investigation is interpreted not narrowly as estimating the causal effect of BMI on the outcome, but more broadly as estimating the causal effect of adiposity (for which BMI is used as a proxy measure) on the outcome, then the estimate may still a valid test

of the causal null hypothesis if there is no causal pathway from the genetic variant(s) to the outcome not via adiposity, in spite of the IV assumptions for BMI being violated.

For these reasons, some authors have questioned whether causal effect estimates should ever be considered as part of a Mendelian randomization analysis [VanderWeele et al., 2014]. Although there is a danger of estimates being overinterpreted, there are several reasons why causal estimates are useful. First, in epidemiology generally, estimates with confidence intervals are preferred to hypothesis tests with p-values, as they are more informative [Sterne and Davey Smith, 2001]. For instance, if a p-value does not achieve conventional levels of statistical significance, a point estimate with a confidence interval allows the reader to judge in a quantitative way whether the null result reflects a lack of evidence or a genuine negative finding in comparison with either the observational association, or with a minimal clinically relevant effect. Secondly, if several genetic variants are valid instrumental variables for the same exposure, greater power to detect a clinically relevant causal effect can be obtained using information on all of the variants simultaneously rather than that using the variants individually. Causal estimates from multiple variants also enable the quantitative comparison of the consistency of genetic associations, using a heterogeneity or overidentification test, as a statistical assessment of pleiotropy (Section 4.5.3). Finally, although the causal estimate in a Mendelian randomization analysis may not be equal to the effect of an intervention in the exposure, it does have a well-defined interpretation as the effect of an intervention in the genetic code at conception. Hence, although assessment of causation should be the primary outcome of a Mendelian randomization investigation, the estimate of a causal effect also has considerable utility.

6.4 Summary

In Mendelian randomization, differences in the exposure distribution due to genetic variation are materially distinct from the change due to any proposed therapeutic intervention on the exposure, and so may affect the outcome differently. Consequently, it may be misleading to generalize the magnitude of a Mendelian randomization association to the effect of a potential intervention on the exposure in practice. Awareness of this is important for the use of Mendelian randomization in target-based drug development.

In this chapter, we have considered qualitative and quantitative issues relating to the interpretation of causal effects using Mendelian randomization and their relationship to the effects of interventions. In the next part of this book, we consider statistical aspects of Mendelian randomization analyses relating to the topics discussed in this and previous chapters.

Part II

Statistical issues in instrumental variable analysis and Mendelian randomization

7

Weak instruments and finite-sample bias

In this chapter, we consider the effect of weak instruments on instrumental variable (IV) analyses. Weak instruments, which were introduced in Section 4.5.2, are those that do not explain a large proportion of the variation in the exposure, and so the statistical association between the IV and the exposure is not strong. This is of particular relevance in Mendelian randomization studies since the associations of genetic variants with exposures of interest are often weak. This chapter focuses on the impact of weak instruments on the bias and coverage of IV estimates.

7.1 Introduction

Although IV techniques can be used to give asymptotically unbiased estimates of causal effects in the presence of confounding, these estimates suffer from bias when evaluated in finite samples [Nelson and Startz, 1990]. A weak instrument (or a weak IV) is still a valid IV, in that it satisfies the IV assumptions, and estimates using the IV with an infinite sample size will be unbiased; but for any finite sample size, the average value of the IV estimator will be biased. This bias, known as weak instrument bias, is towards the observational confounded estimate. Its magnitude depends on the strength of association between the IV and the exposure, which is measured by the F statistic in the regression of the exposure on the IV [Bound et al., 1995]. In this chapter, we assume the context of 'one-sample' Mendelian randomization, in which evidence on the genetic variant, exposure, and outcome are taken on the same set of individuals, rather than subsample (Section 8.5.2) or two-sample (Section 9.8.2) Mendelian randomization, in which genetic associations with the exposure and outcome are estimated in different sets of individuals (overlapping sets in subsample, non-overlapping sets in two-sample Mendelian randomization).

We illustrate this chapter using data from the CRP CHD Genetics Collaboration (CCGC) to estimate the causal effect of blood concentrations of C-reactive protein (CRP) on plasma fibrinogen concentrations (Section 1.3). As the distribution of CRP is positively skewed, we take its logarithm and assume a linear relationship between log(CRP) and fibrinogen. Although log(CRP)

and fibrinogen are highly positively correlated ($r = 0.45$ to 0.55 in the examples below), it is thought that long-term elevated levels of CRP are not causally associated with an increase in fibrinogen.

We first demonstrate the direction and magnitude of weak instrument bias for IV estimates from real and simulated data (Section 7.2). 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 7.3). We discuss simulated results that quantify the size of this bias for different strengths of instruments and different analysis methods (Section 7.4). When multiple IVs are available, we show how the choice of IV affects the variance and bias of IV estimators (Section 7.5). We propose ways of designing and analysing Mendelian randomization studies to minimize bias (Section 7.6). We conclude with a discussion of this bias from both theoretical and practical viewpoints, ending with a summary of recommendations aimed at applied researchers on how to design and analyse a Mendelian randomization study to minimize bias from weak instruments (Section 7.7).

7.2 Demonstrating the bias of IV estimates

First, we demonstrate the existence and nature of weak instrument bias in IV estimation using both real and simulated data.

7.2.1 Bias of IV estimates in small studies

As a motivating example, we consider the Copenhagen General Population Study [Zacho et al., 2008], a cohort study from the CCGC with complete cross-sectional baseline data for 35 679 participants on CRP, fibrinogen, and three SNPs from the *CRP* gene region: rs1205, rs1130864, and rs3093077. We calculate the observational estimate by regressing fibrinogen on log(CRP), and the IV estimate by the two-stage least squares (2SLS) method using all three SNPs as IVs in a per allele additive model (Section 4.2.1). We then analyse the same data as if it came from multiple studies by dividing the data randomly into substudies of equal size, calculating estimates of association in each substudy, and combining the results using inverse-variance weighted fixed-effect meta-analysis. We divide the whole study into, in turn, 5, 10, 16, 40, 100, and 250 substudies. We recall that the F statistic from the regression of the exposure on the IV is used as a measure of instrument strength (Section 4.5.2).

We see from Table 7.1 that the observational estimate stays almost unchanged whether the data are analysed as one study or as several studies. However, as the number of substudies increases, the pooled IV estimate increases from near zero until it approaches the observational estimate. At the

Substudies	Observational estimate	2SLS IV estimate	Mean F statistic
1	1.68 (0.01)	−0.05 (0.15)	152.0
5	1.68 (0.01)	−0.01 (0.15)	31.4
10	1.68 (0.01)	0.09 (0.14)	16.4
16	1.68 (0.01)	0.23 (0.14)	10.8
40	1.68 (0.01)	0.46 (0.13)	4.8
100	1.67 (0.01)	0.83 (0.11)	2.5
250	1.67 (0.01)	1.27 (0.08)	1.6

TABLE 7.1
Estimates of effect (standard error) of log(CRP) on fibrinogen (μmol/l) from the Copenhagen General Population Study ($N = 35\,679$) divided randomly into substudies of equal size and combined using fixed-effect meta-analysis: observational estimates using unadjusted linear regression, IV estimates using 2SLS. Mean F statistics averaged across substudies from linear regression of log(CRP) on three genetic variants.

same time, the standard error of the pooled IV estimates 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. The causal estimate with 16 substudies is positive ($p = 0.09$) despite the causal estimate with the data analysed as one study being near to zero.

7.2.2 Distribution of the ratio IV estimate

In order to investigate the distribution of IV estimates with weak instruments, we use a simulation exercise, taking a simple example of a confounded association with a single dichotomous IV [Burgess and Thompson, 2011]. Parameters are chosen such that the causal effect is null, but simply regressing the outcome on the exposure yields a strong positive confounded observational association of close to 0.5. We took 6 different values of the strength of the IV–exposure association, corresponding to mean F statistic values between 1.1 and 8.7.

Causal estimates are calculated using the ratio method, although with a single IV the estimates from the ratio, 2SLS and limited information maximum likelihood (LIML) methods are the same (Section 4.3.2). The resulting distributions for the estimate of the causal parameter are shown in Figure 7.1. For weaker IVs, there is a marked bias in the median of the distribution in the positive direction and the distribution of the IV estimate has long tails. For the weakest IV considered, the mean F statistic is barely above its null expectation of 1 and the median IV estimate is close to the confounded observational estimate of 0.5. For stronger IVs, the median of the distribution of IV estimates is close to zero. The distribution is skew with more extreme causal estimates tending to take negative values.

The analyses in Table 7.1 and simulations in Figure 7.1 show that IV estimates can be biased. This bias has two notable features: it is larger when the F statistic for the IV–exposure relationship is smaller, and it is in the direction of the confounded observational estimate.

7.3 Explaining the bias of IV estimates

We now try to provide some more intuitive understanding of why weak instrument bias occurs. We give three separate explanations for its existence, in terms of the definition of the ratio estimator, finite-sample violation of the IV assumptions, and sampling variation of IV estimators.

7.3.1 Correlation of IV associations

First, there is a correlation between the numerator (estimate of the G–Y association) and denominator (estimate of the G–X association) in the ratio estimator. To understand this, we consider a simple model of confounded association with causal effect β_1 of X on Y, with a dichotomous IV $G = 0$ or 1, and further correlation between X and Y due to association with a confounder U:

$$\begin{aligned} X &= \alpha_1 G + \alpha_2 U + \varepsilon_X \\ Y &= \beta_1 X + \beta_2 U + \varepsilon_Y \\ U &\sim \mathcal{N}(0, \sigma_U^2); \quad \varepsilon_X \sim \mathcal{N}(0, \sigma_X^2); \quad \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2) \text{ independently.} \end{aligned} \tag{7.1}$$

We initially assume that $\sigma_X^2 = \sigma_Y^2 = 0$ for ease of explanation.

If $\bar{u}_j$ is the average confounder level for the subgroup with $G = j$ (where $j = 0, 1$), an expression for the causal effect from the ratio method is:

$$\beta_1^R = \frac{\Delta Y}{\Delta X} = \frac{\beta_1 \Delta X + \beta_2 \Delta U}{\Delta X} = \beta_1 + \frac{\beta_2 \Delta U}{\alpha_1 + \alpha_2 \Delta U} \tag{7.2}$$

where $\Delta U = \bar{u}_1 - \bar{u}_0$ is normally distributed with expectation zero; ΔX and ΔY are defined similarly. When the instrument is strong, α_1 is large compared to $\alpha_2 \Delta U$. Then the expression β_1^R will be close to β_1. When the instrument is weak, α_1 may be small compared to $\beta_2 \Delta U$ and $\alpha_2 \Delta U$. Then the bias $\beta_1^R - \beta_1$ is close to $\frac{\beta_2}{\alpha_2}$, which is approximately the bias of the confounded observational association (it is exactly this if α_1 is zero). This is true whether ΔU is positive or negative. Figure 7.2 (top panel) 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 ΔU tends towards zero, a bias in the direction of the confounded association will be present in finite samples. From Figure 7.2 (top panel), the median bias will be positive, as the

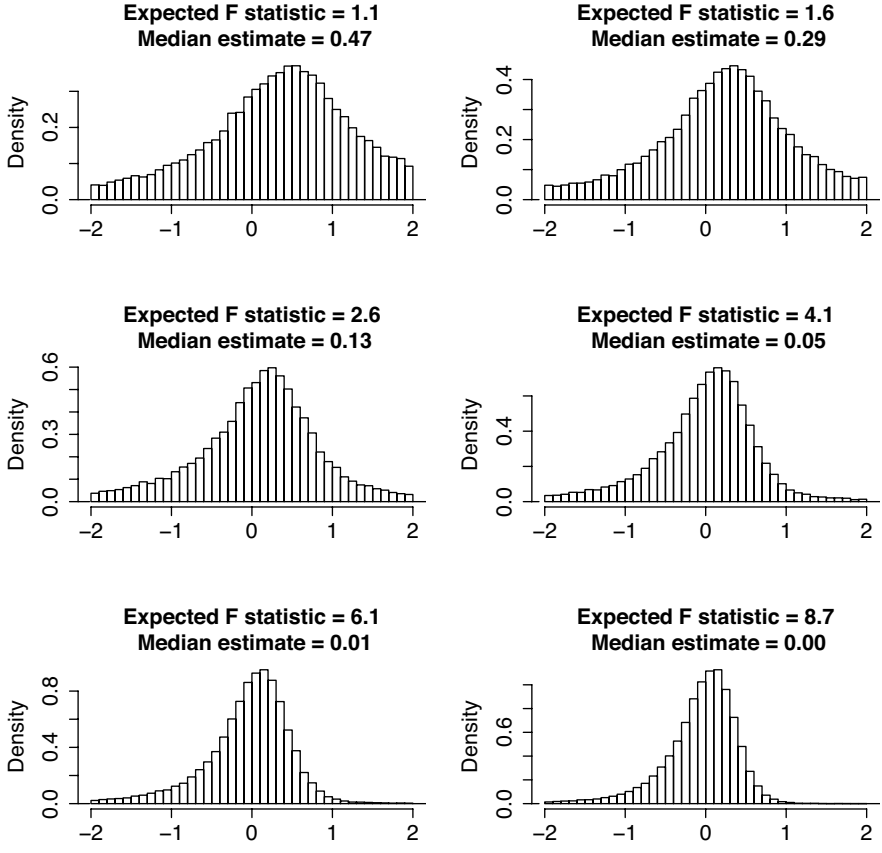


FIGURE 7.1
Histograms of IV estimates of a null causal effect using weak instruments from simulated data for six strengths of the IV–exposure association. Average F statistics and median IV estimates for each scenario are shown.

estimate is greater than β_1 when $\Delta U > 0$ or $\Delta U < -\frac{\alpha_1}{\alpha_2}$, which happens with probability greater than 0.5.

This also explains the heavier negative tail in the histograms in Figure 7.1. 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 7.2.2, this is associated with a negative value of ΔU, whence the numerator $\beta_2 \Delta U$ 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_1^R than positive ones.

If there is independent error in X and Y (that is, σ_X^2 and σ_Y^2 in equation (7.1) are non-zero), then the picture is similar, but more noisy, as seen in Figure 7.2 (bottom panel). The expression for the IV estimator is:

$$\beta_1^R = \beta_1 + \frac{\beta_2 \Delta U + \Delta\varepsilon_Y}{\alpha_1 + \alpha_2 \Delta U + \Delta\varepsilon_X}$$

where $\Delta\varepsilon_X = \bar{\varepsilon}_{X1} - \bar{\varepsilon}_{X0}$ and $\Delta\varepsilon_Y = \bar{\varepsilon}_{Y1} - \bar{\varepsilon}_{Y0}$ are defined analogously to ΔU above.

7.3.2 Finite-sample violation of IV assumptions

An alternative explanation of weak instrument bias is in terms of violation of the second IV assumption in a finite sample. Although a valid instrument will be asymptotically independent of all confounders, in a finite sample there will be a non-zero correlation between the instrument and confounders. This correlation biases the IV estimator towards the observational confounded association.

If the instrument is strong, then the difference in mean exposure between genetic subgroups will be mainly due to the genetic instrument, and the difference in outcome (if any) will be due to this difference in exposure. However if the instrument is weak, that is it explains little variation in the exposure, the chance difference in confounders may explain more of the difference in mean exposure between genetic subgroups than the instrument. If the effect of the instrument is near zero, then the estimate of the "causal effect" approaches the association between exposure and outcome resulting from changes in the confounders, which is the observational confounded association [Bound et al., 1995].

7.3.3 Sampling variation within genetic subgroups

Finally, we offer a graphical explanation of weak instrument bias. To do this, we simulate data with a negative causal effect of the exposure on the outcome, but with positive confounding giving a strong positive observational association between the exposure and outcome. We generate 1000 simulated datasets with 600 subjects divided equally into three genetic subgroups

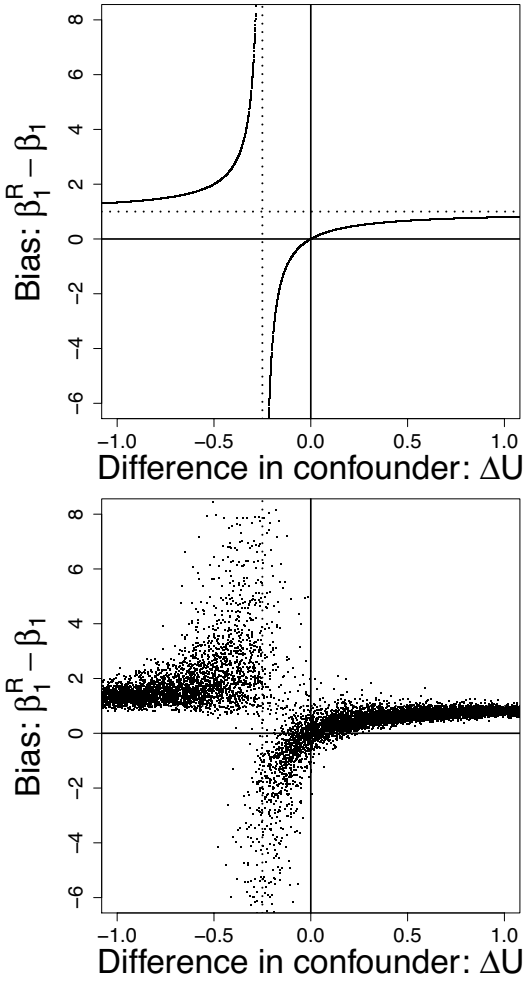


FIGURE 7.2
Bias in IV estimator as a function of the difference in mean confounder between groups ($\alpha_1 = 0.25$, $\alpha_2 = \beta_2 = 1$). Horizontal dotted line is at the confounded association $\frac{\beta_2}{\alpha_2}$, and the vertical dotted line at $\Delta U = -\frac{\alpha_1}{\alpha_2}$ where β_1^R is not defined. Top panel: no independent error in X or Y; bottom panel: $\Delta\varepsilon_X, \Delta\varepsilon_Y \sim \mathcal{N}(0, 0.1^2)$ independently.

($G = 0$, 1, or 2):

$$\begin{aligned} x_i &= \alpha_1 g_i + u_i + \varepsilon_{Xi} \\ y_i &= \beta_1 x_i + u_i + \varepsilon_{Yi} \\ u_i &\sim \mathcal{N}(0, \sigma_U^2); \varepsilon_{Xi} \sim \mathcal{N}(0, \sigma_X^2); \varepsilon_{Yi} \sim \mathcal{N}(0, \sigma_Y^2) \text{ independently.} \end{aligned} \tag{7.3}$$

We set $\beta_1 = -0.4$, $\sigma_U^2 = 1^2$, $\sigma_X^2 = 0.2^2$, and $\sigma_Y^2 = 0.2^2$, and take four values for the strength of the IV ($\alpha_1 = 0.5$, 0.2, 0.1, and 0.05) corresponding to expected F statistics of 100, 16, 4.7, and 2.0. The mean levels of exposure and outcome for each genetic subgroup from each simulated dataset are plotted (Figure 7.3), representing joint density functions for each subgroup. To examine the sampling distribution of the IV estimate, we 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 differences in exposure between the subgroups due to variation in the IV will generally lead to estimating a negative effect of exposure on outcome. When the instrument is weak, the differences in exposure between the subgroups due to the IV are small and the positively confounded observational association is more likely to be recovered.

7.4 Properties of IV estimates with weak instruments

In the previous section, we showed that IV estimates are biased in finite samples. In this section, we consider the magnitude of the bias in IV estimates, as well as the coverage of IV methods with weak instruments.

7.4.1 Bias of IV estimates

The bias of an estimator is the difference between the expectation of the estimator and the true value of the parameter. In IV analysis, the relative mean bias is the ratio of the bias of the IV estimator ($\hat{\beta}_{IV}$) to the bias of the observational association ($\hat{\beta}_{OBS}$) found by linear regression of the outcome on the exposure:

$$\text{Relative mean bias} = \frac{\mathbb{E}(\hat{\beta}_{IV}) - \beta_1}{\mathbb{E}(\hat{\beta}_{OBS}) - \beta_1}. \tag{7.4}$$

The relative mean bias from the 2SLS method is asymptotically approximately equal to $1/\mathbb{E}(F)$, where $\mathbb{E}(F)$ is the expected F statistic in the regression of the exposure on the IV [Staiger and Stock, 1997]. This approximation is only valid when the number of IVs is at least three. The rule-of-thumb of $F < 10$ indicating weak instruments (Section 4.5.2) derives from this expression. This rule approximately limits the bias in the IV estimate to less than

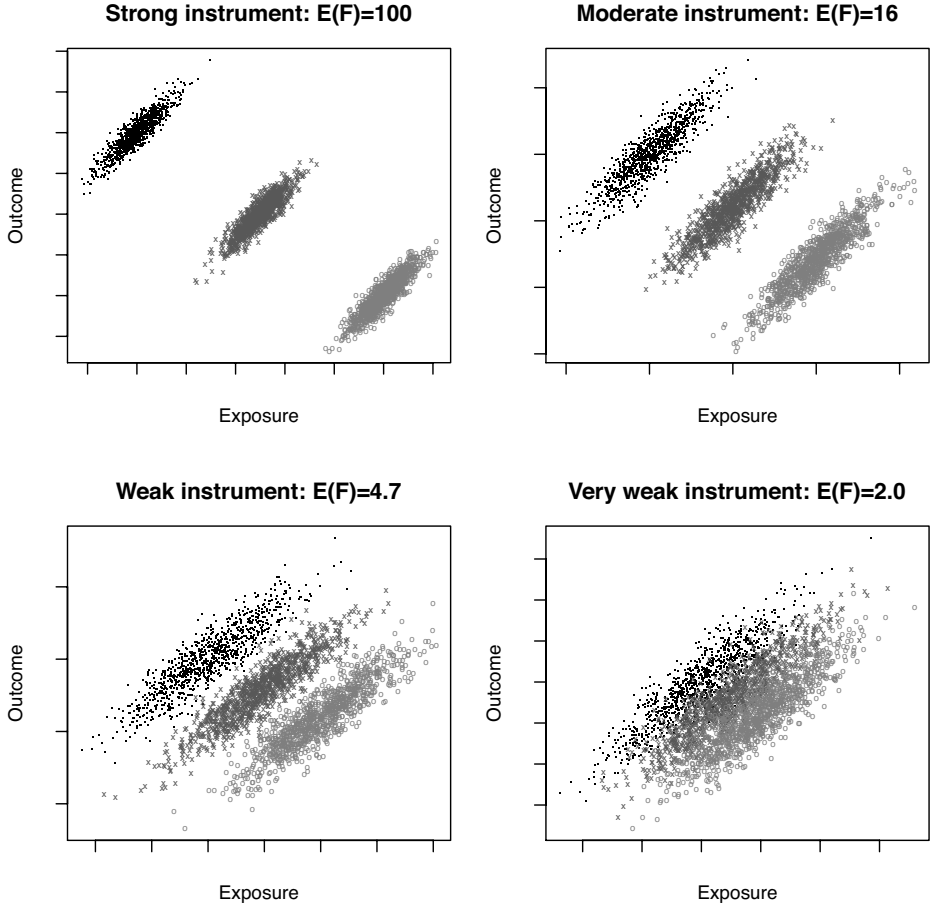


FIGURE 7.3
Distribution of mean outcome and mean exposure levels in three genetic subgroups (indicated by different symbols and shades of grey) for various strengths of the instrument, with expected values of the F statistic. One point of each colour comes from each of 1000 simulated datasets. The IV estimate in each simulation is the gradient of the line through the three points.

10% of the bias in the observational association estimate. However, weak instrument bias depends in a graded way on the F statistic, and such cut-offs are not always helpful or sensible. Biases less than this, corresponding to greater F statistics, can be important in practice. Moreover, as explained later in this chapter, there is an important distinction between the expected F statistic, on which the magnitude of the bias depends, and the F statistic observed in a particular dataset.

With a single IV, the expected value of the 2SLS estimate, and hence the bias, is undefined (Section 4.1.6). Simulations have shown that the median bias of the 2SLS method with a single IV (or equivalently the ratio or LIML method) is close to zero even for IVs with expected F statistics around 5, where the median bias is defined as the difference between the median estimate and the true value [Burgess and Thompson, 2011].

Other methods, such as likelihood-based methods, are less susceptible to bias. Although the mean bias of the LIML estimate is undefined (Section 4.3.2), the median bias is close to zero [Angrist and Pischke, 2009]. Simulations for Bayesian methods for IVs with expected F statistics around 5 have shown mean and median bias close to zero [Burgess and Thompson, 2012].

7.4.2 Coverage of IV estimates

In addition to problems of bias, IV estimates with weak instruments can have underestimated coverage [Stock and Yogo, 2002; Mikusheva and Poi, 2006]. As seen in Figure 7.1, the distribution of the IV estimate has long tails, and so is poorly approximated by a normal distribution. This means that asymptotically derived confidence intervals may underestimate the true uncertainty in the causal effect. This underestimation is especially severe when confounding is strong. Simulations for the 2SLS method have shown coverage as low as 75% for a nominal 95% confidence interval [Burgess and Thompson, 2012]. Similar results have been observed for the LIML method when there is a large number of IVs; while a correction is available (Bekker standard errors [Bekker, 1994]), this leads to inefficient estimates [Davies et al., 2014]. Confidence intervals from Fieller's theorem (Section 4.1.5), which are not constrained to be symmetric (or even finite), or those which do not rely on asymptotic assumptions, such as credible intervals from a Bayesian posterior distribution drawn from Monte Carlo Markov chain (MCMC) sampling, result in better coverage properties [Imbens and Rosenbaum, 2005]. Alternatively, confidence intervals from inverting a test statistic, such as the Anderson–Rubin test statistic [Anderson and Rubin, 1949] or the conditional likelihood ratio test statistic [Moreira, 2003] give appropriate confidence levels under the null hypothesis with weak instruments [Mikusheva, 2010].

7.4.3 Lack of identification

For semi-parametric approaches to IV analysis, such as the generalized method of moments (GMM) or structural mean models (SMM), there is no guarantee that a unique parameter estimate will be obtained, as the estimating equations may have no or multiple solutions (Section 4.4.3*). This is a common problem when the instrument is weak. Even when there is a unique solution, if the gradient of the graph of the objective function from the estimating equations against parameter values is close to zero in the neighbourhood of the parameter estimate, or if the objective function cannot be well approximated by a quadratic function, then identification is said to be weak, and problems of bias and coverage as explained above are likely to occur. Simulations suggest that the probability of obtaining a unique solution to the estimating equations with a binary outcome and a log-linear model is not especially sensitive to the sample size, depending more on the coefficient of determination (R^2, the proportion of variance in the exposure explained by the IV(s)) than the F statistic. With R^2 of 2% or less, lack of identification in a multiplicative GMM (or equivalently a multiplicative SMM) model was observed in over 50% of simulated datasets even when the F statistic was in the hundreds or even thousands [Burgess et al., 2014c].

7.5 Bias of IV estimates with different choices of IV

Including more instruments, where each instrument explains extra variation in the exposure, should give more information on the causal parameter (see Chapter 8). However, 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 IV estimates.

7.5.1 Multiple candidate IVs in simulated data

In order to investigate how using multiple instruments affects the bias of IV estimates, we perform simulations in a model [Burgess and Thompson, 2011] where, for each participant indexed by i, the exposure x_i depends linearly on six dichotomous IVs ($g_{ik}, k = 1, \ldots, 6$), a normally distributed confounder u_i, and an independent normally distributed error term ε_{Xi}. Outcome y_i is a linear combination of exposure, confounder, and an independent error term

ε_{Yi}:

$$x_i = \sum_{k=1}^{6} \alpha_{1k}\, g_{ik} + \alpha_2\, u_i + \varepsilon_{Xi} \tag{7.5}$$
$$y_i = \beta_1\, x_i + \beta_2\, u_i + \varepsilon_{Yi}$$
$$u_i, \varepsilon_{Xi}, \varepsilon_{Yi} \sim \mathcal{N}(0, 1^2) \text{ independently.}$$

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

Table 7.2 shows the median and 95% range of the estimates from the 2SLS and LIML methods and the mean estimate 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 statistics in each simulation, as well as using all IVs with an F statistic greater than 10 in univariate regressions of exposure on each IV.

Using 2SLS, as the number of IVs increases, the bias increases, despite the mean F statistic remaining fairly constant. This is because there is a greater risk of imbalances in confounders between the greater number of genetic subgroups defined by the instruments. The data are being subdivided in more different ways, and so there is more chance of these divisions giving genetic subgroups with different average levels of confounders. However, the variability of the IV estimator decreases. This is because a greater proportion of the variance in the exposure is modelled. The greatest increase in median bias is from one IV to two IVs, and coincides with the greatest increase in precision. With the 2SLS method, we therefore have a bias–variance trade-off in deciding how many IVs to use [Zohoori and Savitz, 1997].

While LIML provides estimates which are slightly more variable than 2SLS, a similar increase in precision with the number of IVs is observed, but no increase in bias. For 2SLS, the mean estimates are slightly smaller than the median estimates presented. In the case of a single IV, the theoretical mean is infinite (Section 4.1.6). For LIML, the mean bias is infinite for all numbers of IVs (Section 4.3.2).

Using the single IV with the greatest F statistic gives markedly biased results, despite a mean F statistic 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 exposure. If the IVs are chosen to be included in an analysis because they explain a large proportion of the variation in the exposure 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. In the notation of Section 7.3.1, ΔU is large and, having the same sign as α_1, leads to an estimate biased in the direction of $\frac{\beta_2}{\alpha_2}$. Conversely, 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, as shown in Table 7.2.

So we see that if the F statistic is used either to choose between instruments, or via a rule such as only including an IV in the analysis if $F > 10$, this procedure itself introduces a selection bias which can be greater in magnitude than the bias from weak instruments [Hall et al., 1996]. In a more realistic example, IVs would not all have the same true strength. However, the large sampling variation in F statistics means that choosing between IVs on the basis of a single measured F statistic is unreliable. One solution to this in practice is to use the strength of the IVs in an independent dataset to determine the IVs to include in an applied analysis, or to use an allele score to summarize multiple variants as a single IV (see Chapter 8).

IVs used	Median 2SLS	2.5% to 97.5% quantiles 2SLS	Median LIML	2.5% to 97.5% quantiles LIML	Mean 2SLS [1]	Mean F statistic
1 IV	0.00	−1.12 to 0.53			–	10.2
2 IVs	0.02	−0.54 to 0.39	0.00	−0.64 to 0.39	0.00	10.4
3 IVs	0.03	−0.39 to 0.33	0.00	−0.48 to 0.32	0.02	10.6
4 IVs	0.03	−0.31 to 0.30	0.00	−0.40 to 0.28	0.02	10.8
5 IVs	0.04	−0.26 to 0.27	0.00	−0.34 to 0.26	0.03	11.0
6 IVs	0.04	−0.23 to 0.26	0.00	−0.31 to 0.23	0.03	11.3
Greatest F	0.14	−0.30 to 0.52			–	23.9
Least F	−0.32	−2.57 to 0.58			–	6.7
IVs with $F > 10$	0.11	−0.20 to 0.39	0.10	−0.22 to 0.39	0.11	16.4

TABLE 7.2
Evaluation of bias: Median and 95% range of estimates of $\beta_1 = 0$ using 2SLS and LIML methods, mean estimate using 2SLS method and mean F statistic across 100 000 simulations using combinations of six uncorrelated instruments, using the instrument with the greatest/least F statistic, and using all instruments with univariate F statistics greater than 10.

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

7.5.2 Multiple candidate IVs in the Framingham Heart Study

As a further illustration, we consider the Framingham Heart Study, a cohort study measuring CRP and fibrinogen at baseline with complete data on 1500 participants for nine SNPs in the *CRP* gene. The observational estimate of the log(CRP)–fibrinogen (μmol/l) association is 1.13 (95% CI 1.05 to 1.22). We calculate the causal estimate of the association using the 2SLS method with different numbers of SNPs as an instrument, using a per allele additive model. Figure 7.4 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 effect reduces as we include more instruments, but the median causal estimate across the different combinations of IVs increases. The 2SLS estimate using all nine SNPs in an per allele additive model is -0.01 (95% CI: -0.72 to 0.71, $p = 0.99$, $F_{9,1490} = 3.34$). If we relax the assumptions of a per allele genetic model with additivity between SNPs to instead use a fully saturated model with one coefficient for each of the 49 genotypes represented in the data, the 2SLS estimate is 0.79 (95% CI 0.42 to 1.16, $p < 0.001$, $F_{48,1451} = 1.66$). Using LIML, the estimate from the saturated genetic model is 0.05 (95% CI -0.71 to 0.81, $p = 0.89$) – much less biased than the 2SLS estimate.

This illustrates the bias in the 2SLS method due to the use of multiple instruments, showing how an estimate close to the observational association can be obtained by injudicious choice of instrument. In the extreme case, if each of the individuals in a study were placed into separate genetic subgroups, then the IV estimate would be exactly the observational association. The LIML method with the saturated genetic model gives a substantially different answer to the 2SLS method, an indication that the 2SLS estimate may be biased.

7.6 Minimizing the bias of IV estimates

To provide guidance for epidemiological applications, we now list specific ways by which bias from weak instruments can be minimized in the design and analysis of Mendelian randomization studies.

7.6.1 Increasing the F statistic

As stated previously, the bias in 2SLS IV estimates depends on the expected F statistic in the regression of the exposure on the IV. This means that bias can be reduced by increasing the expected F statistic. The F statistic is

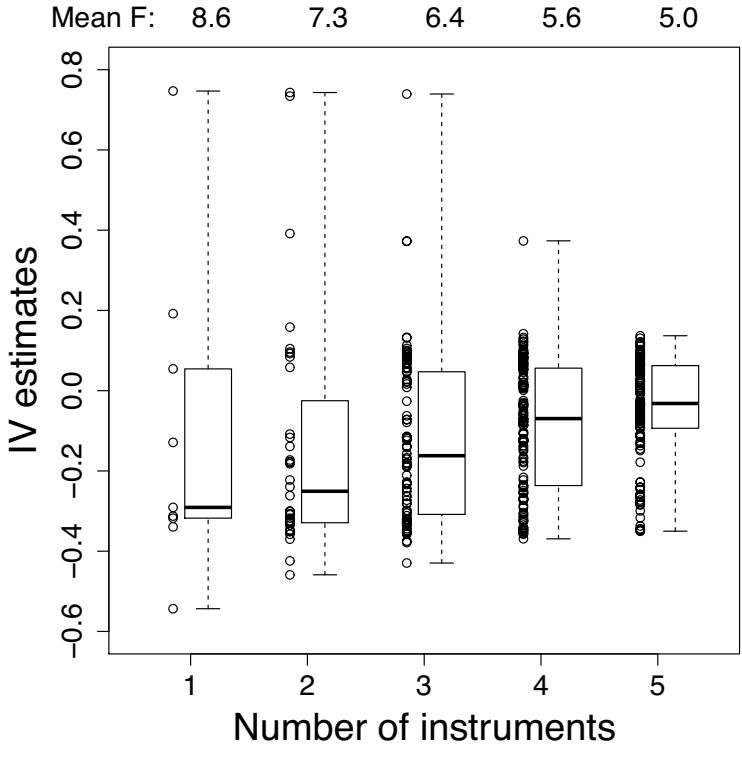


FIGURE 7.4
2SLS IV estimates for causal effect in the Framingham Heart Study of log(CRP) on fibrinogen (μmol/l) using all combinations of varying numbers of SNPs as IVs. Point estimates, associated box plots (median, inter-quartile range, range) and mean F statistics across combinations are displayed.

related to the proportion of variance in the exposure 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, bias can be reduced by increasing the sample size. Similarly, if there are instruments that are not contributing much to explaining the variation in the exposure, then excluding these instruments will increase the F statistic. 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 [Pierce et al., 2011; Palmer et al., 2011a]. Simulations have shown that, even when the true model is only approximately linear in the IV, a per allele genetic model reduces bias [Burgess and Thompson, 2011].

However, it is not enough to simply rely on an F statistic measured from data to inform us about bias [Hall et al., 1996]. Returning to the example from Section 7.2.1 where we divided the Copenhagen General Population Study into 16 equally sized substudies with mean F statistic 10.8, Figure 7.5 shows the estimates of these 16 substudies using the 2SLS method with their corresponding F statistics. We see that the substudies which have greater estimates are the ones with larger F statistics; the correlation between F statistics and point estimates is 0.83. The substudies with higher F statistics 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 0.23 (SE 0.14, $p = 0.09$) to 0.43 (SE 0.16, $p = 0.006$). 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 exposure, then the pooled estimate increases to 0.28 (SE 0.15, $p = 0.06$). So neither of these approaches are useful in reducing bias.

Although the expectation of the F statistic is a good indicator of bias, the observed F statistic shows considerable variation. In the 16 substudies of Figure 7.5, the measured 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 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 7.5, 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. Echoing the comments of Section 7.5 regarding the inclusion of IVs in a model, any guidance that relies on providing a threshold (such as $F > 10$) as an inclusion criterion is flawed and may introduce more bias than it prevents.

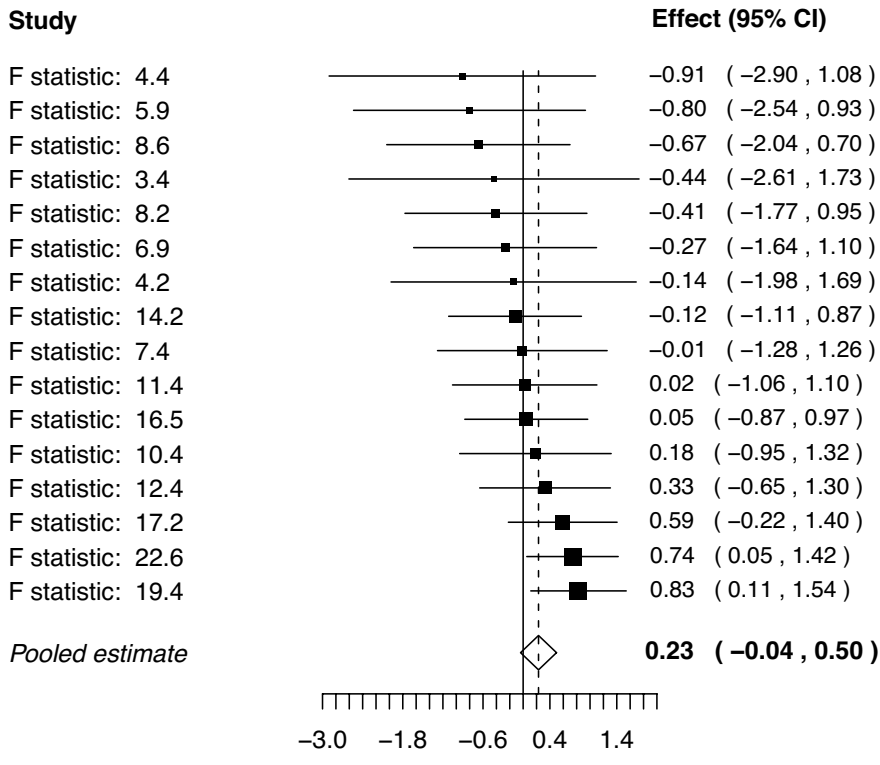


FIGURE 7.5
Forest plot of causal estimates of log(CRP) on fibrinogen (μmol/l) using data from the Copenhagen General Population Study divided randomly into 16 equally sized substudies (each $N \simeq 2230$). Studies ordered by causal estimate. F statistic from regression of exposure on three IVs. Size of markers is proportional to weight in a fixed-effect meta-analysis.

7.6.2 Adjustment for measured covariates

If we can find measured covariates that explain variation in the exposure, and that are not on the causal pathway between exposure and outcome, then we can incorporate these covariates in our model. This will increase precision in the genetic association with the exposure and reduce weak instrument bias. Simulations have shown that we may also see an increase in the precision of the IV estimator if these covariates are additionally used to explain variation in the outcome [Burgess et al., 2011b].

As an example, we consider data on interleukin-6 (IL6), a cytokine which is involved in the inflammation process upstream of CRP and fibrinogen [Hansson, 2005]. Elevated levels of IL6 lead to elevated levels of both CRP and fibrinogen, so IL6 is correlated with short-term variation in CRP [Kaptoge et al., 2010], but is independent of underlying genetic variation in CRP [CCGC, 2011]. 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.

We use data from the Cardiovascular Health Study, a cohort study from the CCGC measuring CRP, IL6 and fibrinogen at baseline, as well as three SNPs (rs1205, rs1417938, and rs1800947) on the CRP gene, with complete data for 4137 subjects. The proportion of variance in log(CRP) explained by log(IL6) is 26%. We calculate the 2SLS IV estimate of the CRP–fibrinogen association for each SNP separately and for all the SNPs together in an per allele additive model, both without and with adjustment for log(IL6) in the first- and second-stage regressions. Results are given in Table 7.3. We see that after adjusting for log(IL6) the causal estimate in each case has decreased (reflecting reduced weak instrument bias), its standard error has reduced (reflecting increased precision), and the F statistic has increased. With adjustment for a covariate, the relevant F statistic is a partial F statistic, representing the variation in the exposure explained by the IVs once the variation explained by the covariate has been accounted for. This is calculated from an analysis of variance (ANOVA) model.

7.6.3 Borrowing information across studies

The IV estimator would be unbiased if we knew the true values for the average exposure in different genetic subgroups. In a meta-analysis context [Thompson et al., 2005], we can combine the estimates of genotype–exposure association from different studies to give more precise estimates of exposure levels in each genetic subgroup. In the 2SLS method, an individual participant data (IPD) fixed-effect meta-analysis for data on individual i in study m with exposure x_{im}, outcome y_{im} and g_{ikm} for number of minor alleles (0, 1, or 2) of genetic

	Not adjusted		Adjusted	
IV estimate	Estimate (SE)	F statistic	Estimate (SE)	F statistic
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 3 SNPs	0.186 (0.194)	24.4	0.127 (0.188)	32.2

TABLE 7.3
2SLS estimates and standard errors (SE) of the 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 per allele additive model, without and with adjustment for log(IL6) in the Cardiovascular Health Study.

variant k ($k = 1, 2, \ldots K_m$) is:

$$\begin{aligned} x_{im} &= \alpha_{0m} + \sum_{k=1}^{K_m} \alpha_{km}\, g_{ikm} + \varepsilon_{Xim} \\ y_{im} &= \beta_{0m} + \beta_1\, \hat{x}_{im} + \varepsilon_{Yim} \\ \varepsilon_{Xim} &\sim \mathcal{N}(0, \sigma_X^2); \varepsilon_{Yim} \sim \mathcal{N}(0, \sigma_Y^2) \text{ independently.} \end{aligned} \tag{7.6}$$

The exposure levels are regressed on the IVs using a per allele additive linear model separately in each study, and then the outcome levels are regressed on the fitted values of exposure ($\hat{x}_{im}$). The terms α_{0m} and β_{0m} are study-specific intercept terms. Here we assume homogeneity of variances across studies; we can use Bayesian methods to allow for possible heterogeneity (see Section 9.6).

If the same genetic variants are measured in each study and are assumed to have the same effect on the exposure, we can use common genetic effects (i.e. $\alpha_{km} = \alpha_k$) across studies by replacing the first line in equation (7.6) with:

$$x_{im} = \alpha_{0m} + \sum_{k=1}^{K} \alpha_k\, g_{ikm} + \varepsilon_{Xim} \tag{7.7}$$

If the assumption of common genetic effects is correct, this will improve the precision of the fitted values ($\hat{x}_{im}$) and reduce weak instrument bias.

To illustrate this, 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 [CCGC, 2008]. In case-control studies, we use the data from controls alone since these better represent cross-sectional population studies. These five studies measured the same three SNPs on the *CRP* gene: rs1205, rs1130864 and rs3093077 (or rs3093064, which is in complete linkage disequilibrium with rs3093077). We estimate the causal effect

Study	N	F	df	Causal estimate (SE)	Observational estimate (SE)
CCHS	7999	29.6	(3, 7995)	−0.286 (0.373)	1.998 (0.030)
EAS	650	6.9	(3, 646)	0.754 (0.327)	1.115 (0.056)
HPFS	405	5.3	(3, 401)	0.758 (0.423)	1.048 (0.081)
NHS	385	6.1	(3, 381)	−0.906 (0.636)	0.562 (0.114)
SHEEP	1044	10.5	(3, 1040)	0.088 (0.345)	1.078 (0.051)
Different genetic effects		14.4	(15, 10463)	0.021 (0.195)	
Common genetic effects		56.6	(3, 10475)	−0.093 (0.225)	
Study-level estimates				0.234 (0.174)	

TABLE 7.4
Estimates of effect of log(CRP) on fibrinogen (μmol/l) from each of five studies separately and from meta-analysis of studies: number of participants (N), F statistic (F) with degrees of freedom (df) from per allele additive regression of exposure on three SNPs used as IVs, causal estimate using 2SLS with standard error (SE), observational estimate with SE. Fixed-effect meta-analyses conducted using individual-level data with different study-level genetic effects, common pooled genetic effects, and combining study-level estimates with inverse-variance weighting.

using the 2SLS method with different genetic effects (model 7.6), common genetic effects (model 7.7) and by a fixed-effect meta-analysis of estimates from each study.

Table 7.4 shows that the studies analysed separately have apparently disparate causal estimates with large SEs. The meta-analysis estimate assuming common genetic effects across studies is further from the confounded observational estimates and closer to the IV 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 pooled estimate from the study-level meta-analysis is greater than those from the individual-level meta-analyses. 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 are known to be underestimated when the correlation due to confounding is strong, especially with weak instruments (Section 7.4.2) [Stock and Yogo, 2002]. Also, Figure 7.5 showed that causal estimates nearer to the observational association have lower variance. So a study-level meta-analysis may be biased due to overestimated weights in the studies with more biased estimates.

Returning to the example of data from the Copenhagen General Population Study considered in Section 7.2.1, if we use the IPD (model 7.6) to

Substudies	Meta-analysis	p-value	Different genetic effects	p-value	Common genetic effects	p-value
1	−0.05 (0.15)	0.76				
5	−0.01 (0.15)	0.95	−0.03 (0.15)	0.85	−0.05 (0.15)	0.75
10	0.09 (0.14)	0.54	0.04 (0.14)	0.80	−0.05 (0.15)	0.76
16	0.23 (0.14)	0.09	0.15 (0.14)	0.26	−0.05 (0.15)	0.75
40	0.46 (0.13)	< 0.001	0.30 (0.13)	0.02	−0.04 (0.15)	0.77
100	0.83 (0.11)	< 0.001	0.68 (0.11)	< 0.001	−0.04 (0.15)	0.77
250	1.27 (0.08)	< 0.001	1.15 (0.08)	< 0.001	−0.04 (0.15)	0.78

TABLE 7.5
2SLS estimates of causal effect (standard error) of log(CRP) on fibrinogen from the Copenhagen General Population Study divided randomly into substudies and combined: using fixed-effect meta-analysis of substudy estimates, and using individual patient data (IPD) with different or common genetic effects across substudies.

combine the substudies in the meta-analysis rather than combining estimates from each substudy, then the pooled estimates are somewhat less biased (Table 7.5). If we additionally assume common genetic effects across studies (model 7.7), then we recover close to the original estimate based on analysing the full dataset as one study: weak instrument bias has been eliminated.

7.7 Discussion

This chapter has demonstrated the effect of weak instrument bias on causal estimates in real and simulated data. The magnitude of this bias depends on the statistical strength of the association between instrument and exposure.

Weak instrument bias can reintroduce the problem that IVs were developed to solve. It is misleading not solely because it biases estimates, but because estimates suffering from the bias do not provide a valid test of the null hypothesis. Weak instruments may convince a researcher that an observational association that they have estimated is in fact causal. The reason for the bias is that the variation in the exposure explained by the IV is not large enough to dominate the variation in the exposure caused by chance correlation between the IV and confounders.

While the magnitude of the bias depends on the instrument strength through the expected or mean F statistic, for a study of fixed size and underlying instrument strength, an observed F statistic greater than its expected value corresponds to an estimate closer to the observational association with

greater precision; conversely an observed F statistic less than the expected value corresponds with an estimate further from the observational association with less precision. Simply relying on an F statistic from an individual study is over-simplistic and threshold rules such as ensuring $F > 10$ may cause more bias than they prevent.

7.7.1 Bias–variance trade-off

Using the 2SLS method, we demonstrated a bias–variance trade-off for the 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 with the number of instruments, but the precision was slightly lower than for 2SLS. When using 2SLS, we seek parsimonious models of genetic association, for example using per allele additive models and including only IVs with a known association with the exposure, based on biological knowledge and external information. Provided the data are not severely misrepresented, these should provide the best estimates of the causal effect. Again, *post hoc* use of observed F statistics to choose between instruments may cause more bias than it prevents.

7.7.2 Combatting weak instrument bias in practice

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 values 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.

Testing the association between the outcome and each IV in turn (without estimating a causal effect) is a valid test of a causal relationship even with weak instruments. If there is a single IV, then an expected F statistic of 5 corresponds to a p-value in the regression of the exposure on the IV of around 0.03. It is perhaps unlikely that an IV would be considered for use in a dataset if the expected p-value were much greater than 0.03, and so bias from weak instruments would not be expected to be an issue in practice with a single IV. If there are multiple IVs, LIML or Bayesian methods could be used in the analysis, as the estimates from these are less biased than the 2SLS estimate. A difference between the 2SLS and LIML IV estimates is evidence of possible bias from weak instruments. The use of Fieller's theorem, the Anderson–Rubin test statistic or a Bayesian posterior distribution for inference is recommended.

It is also possible to summarize multiple SNPs into a single variable to reduce weak instrument bias using an allele score. Details about how to construct such a score are given in Chapter 8.

Adjustment for covariates helps reduce weak instrument bias. Including predictors of the exposure in the first-stage regression, or predictors of the outcome in the second-stage regression, also increases precision of the causal estimate. The former will also increase the F statistic for the IVs, and thus reduce weak instrument bias.

This chapter has considered bias in a one-sample Mendelian randomization setting. If the genetic associations with the exposure and outcome are estimated in non-overlapping sets of individuals, then bias from weak instruments will act in the direction of the null (Section 9.8.2). Although bias is never welcome, the direction of bias in a two-sample Mendelian randomization analysis means that a non-null causal effect estimate will not simply be an artefact of weak instrument bias.

7.7.3 Bias in study-level meta-analysis

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 estimates and their variances which results in studies with effects closer to the observational estimate being over-weighted. By using a single IPD model, we can 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 made in Section 7.6.3 are overly restrictive in practice; more reasonable extensions of IV methods to a meta-analysis context are discussed in Chapter 9.

7.7.4 Caution about validity of IVs

Finally, we recall 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 [Small and Rosenbaum, 2008].

7.8 Key points from chapter

- Bias from weak instruments can result in seriously misleading estimates of causal effects. Studies with instruments having large expected F statistics are less biased on average. However, if a study by chance has a larger

observed F statistic than expected, then the causal estimate will be more biased.

- Coverage levels with weak instruments can be poorly estimated by methods which rely on assumptions of asymptotic normality.
- Data-driven choice of instruments or analysis can exacerbate bias. In particular, any threshold guideline such as ensuring that an observed F statistic is greater than 10 is misleading. Methods, instruments, and data to be used should be specified prior to data analysis. Meta-analyses based on study-specific estimates of causal effect are susceptible to bias.
- Bias can be alleviated by use of measured covariates and parsimonious modelling of the genetic association (such as a per allele additive SNP model rather than one coefficient per genotype). This should be accompanied by sensitivity analyses to assess potential bias, for example from model misspecification.
- Bias can be reduced substantially by using LIML, Bayesian and allele score (see next chapter) methods rather than 2SLS, and bias in practice with a single IV should be minimal. Nominal coverage levels can be maintained by the use of Fieller's theorem with a single IV, and confidence intervals from the Anderson–Rubin test statistic or Bayesian MCMC methods with multiple IVs.

8

Multiple instruments and power

In the next two chapters, we consider extensions to IV methods to efficiently analyse data typically available in Mendelian randomization investigations. The first extension is the inclusion of multiple instrumental variables in a single analysis model, and the statistical issues arising. We consider the impact on statistical power, and discuss the practical issue of missing data, which can limit power gains.

8.1 Introduction

Although instrumental variable (IV) methods give estimates which are consistent for the causal effect, their variance is typically much larger than the variance of the estimate from an observational analysis [Davey Smith and Ebrahim, 2004]. This is because the variation in the exposure explained by the IV is usually small. If there are multiple IVs available, a more precise causal effect estimate can be obtained by incorporating data on all the IVs simultaneously to estimate a single causal effect [Palmer et al., 2011a]. However, two problems arising from including multiple IVs in an analysis are weak instruments and missing data.

When there are large numbers of genetic variants, several IV methods give estimates which are biased in the direction of the observational estimate with incorrectly sized confidence intervals (see Chapter 7). Allele scores are a convenient way of summarizing a large number of genetic variants associated with an exposure. Using a univariate allele score as a single IV rather than each genetic variant as a separate IV helps resolve problems in IV estimation resulting from weak instruments.

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, methods for incorporating participants with partially missing data can be employed.

In this chapter, we address the construction and use of allele scores in Mendelian randomization (Section 8.2). We investigate the power of an IV analysis, demonstrating the gain in power from using IVs which explain a greater proportion of the variance in the exposure (which can be achieved by including more genetic variants in an analysis) (Section 8.3). We then show that subjects with partially missing genetic data can be included in an analysis, enabling multiple IVs to be employed without reducing the available sample size even if data on each IV is incomplete (Section 8.4). Finally, we discuss other issues relating to the use of multiple IVs in Mendelian randomization analyses (Section 8.5).

8.2 Allele scores

As explained in Chapter 7, the use of large numbers of IVs can result in bias and poor coverage properties. An allele score (also called a genetic risk score, gene score, or genotype score) is a single variable summarizing multiple genetic variants in a univariate score. An unweighted allele score is constructed as the total number of exposure-increasing alleles present in the genotype of an individual. A weighted allele score can also be considered, where each allele contributes a weight reflecting the effect of the corresponding genetic variant on the exposure. These weights can be derived internally from the data under analysis, or externally from prior knowledge or an independent data source.

If an individual i has g_{ik} copies of the exposure-increasing allele for each variant $k = 1, \ldots, K$, then their unweighted score is $\sum_{k=1}^{K} g_{ik}$. This score takes integer values between 0 and $2K$. If the weight for variant k is w_k, then their weighted score is $\sum_{k=1}^{K} w_k g_{ik}$. Either score can then be used in an IV analysis using the ratio method, which, as we saw in Chapter 7, has median bias close to zero.

Another reason for using an allele score is simplicity. With large numbers of variants, the validity of the IV assumptions can be partially assessed by testing the association of each variant with a set of measured covariates. Additionally, the association of the allele score with the covariates can be tested. Assessment of IV violations will be clearer with a single score variable, rather than many variants, and power to detect a violation will be improved if several variants have pleiotropic associations with the same risk factor.

The use of an allele score in Mendelian randomization requires the assumption that the allele score is an instrumental variable. This means that each variant which contributes to the allele score must satisfy the assumptions of an instrumental variable, except that it is not necessary for all the variants to be associated with the exposure (a variant not associated with the exposure but satisfying the second and third IV assumptions will not invalidate the score, but neither will it add any information to the score). Additionally,

several parametric assumptions are made in specifying the allele score, such as additivity in the genetic model with no interactions between variants.

8.2.1 Choosing variants to include in an allele score

In Section 7.5.1, we saw that criteria for selecting IVs based on the data under analysis led to bias. The phenomenon that the magnitude of effect of the variant with the strongest association is typically over-estimated is known as the "winner's curse" (also the Beavis effect) [Taylor et al., 2014]. The choice of variants to include in an allele score should be made prior to analysis, or on the basis of external (independent) data. This is particularly important if there are several candidate variants with similar magnitudes of association with the exposure. Additionally, the inclusion of variants which are highly correlated with each other (in high linkage disequilibrium) will not give extra information compared to including any one of these variants, and may lead to inefficiency if the correlation is not taken into account in determining the weights.

8.2.2 Choosing weights in a weighted allele score

If the weights in a weighted allele score are the estimates from a regression of the exposure on the genetic variants using the data under analysis, then an IV analysis using an allele score gives precisely the same answer as a two-stage least squares (2SLS) analysis using each of the variants as separate IVs. In this case, there is no advantage in using an allele score over a conventional multiple IV analysis. Weights can be derived from external data, although simulations have shown that estimates using an unweighted score are unbiased [Burgess and Thompson, 2013]. Although there is some loss of power associated with using an unweighted rather than a weighted score, this loss is not large if the genetic variants have fairly similar magnitudes of association with the exposure. A similar loss of power is suffered in using a weighted score approach if the weights are imprecisely estimated.

If external data are not available, weights can instead be estimated using the data under analysis in a cross-validation approach, by dividing the data into equal sized parts, and constructing an allele score using weights in each part estimated using the data from all the other parts. For example, in a 10-fold cross-validation, 10 sets of weights are estimated. Weights used for constructing the allele score in each tenth of the sample are obtained from the remaining 90% of the sample. In this way, there is no correlation between the weights and the data for each individual, and a weighted allele score can be assigned to each individual in the study using the appropriate set of weights. A single IV estimate can be obtained using the weighted allele score across the whole dataset. Alternatively, separate IV estimates can be obtained for each tenth of the data, and then these estimates can be combined, for example using a fixed-effect meta-analysis model.

In general, a cross-validation approach would be preferred if external weights are not available or are thought to be not fully relevant to the data under analysis. Otherwise, whichever approach gave more precisely estimated weights would be preferred.

8.2.3 Performance of an allele score in IV estimation

Simulations have shown that the use of an allele score improves bias and coverage properties of IV estimates compared with estimates from the 2SLS and LIML methods, especially when large numbers of variants are included in the score [Burgess and Thompson, 2013]. They were also more efficient than 2SLS and LIML estimates in a simulation example [Davies et al., 2014]. The bias and coverage properties seem to be robust to misspecifications of the score, such as the presence of gene–gene and gene–environment interactions, departures from additivity in the genetic model, and mismeasurement of weights in a weighted score approach. However, as stated above, they are not robust to naive procedures which use the data under analysis to construct the score, nor to the inclusion of invalid IVs in a score.

One important conclusion from this is that the procedure for constructing an allele score in an applied Mendelian randomization analysis should be described fully and clearly (see Section 5.4).

8.3 Power of IV estimates

In this section, we initially investigate the power of an IV analysis with a single IV, and then consider the potential benefits of using multiple IVs.

8.3.1 Power with a single IV, continuous outcome

With a single IV and a continuous outcome, the asymptotic variance of the IV estimate of the causal effect of the exposure X on the outcome Y with a single IV G is given by the formula:

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}(R_Y^{IV})}{N \ \text{var}(X)\, \rho_{GX}^2} \tag{8.1}$$

where N is the sample size, $R_Y^{IV} = Y - \beta_1 X$ is the residual of the outcome on subtraction of the causal effect of the exposure, and ρ_{GX}^2 is the square of the correlation between the exposure X and the IV G [Nelson and Startz, 1990]. The coefficient of determination (R^2) in the regression of the exposure on the IV is an estimate of ρ_{GX}^2. The IV in these calculations could either be a single genetic variant or an allele score. This formula corresponds to the first term

from the delta method expansion (equation 4.9), and ignores uncertainty in the IV–exposure association; the subsequent calculations therefore represent the power to detect an association between the IV and the outcome.

The asymptotic variance of the conventional regression (ordinary least squares, OLS) estimate of the association between the exposure X and the outcome Y is given by the formula:

$$\text{var}(\hat{\beta}_{OLS}) = \frac{\text{var}(R_Y^{OLS})}{N \text{ var}(X)} \tag{8.2}$$

where $R_Y^{OLS} = Y - \beta_{OLS}X$ is the residual of the outcome on subtraction of the observational association of the exposure. The sample size necessary for an IV analysis to demonstrate a given magnitude of causal effect is therefore approximately equal to that for a conventional epidemiological analysis to demonstrate the same magnitude of association divided by the parameter ρ_{GX}^2 for the IV [Wooldridge, 2009].

If the two-sided significance level is α and the power desired to test the null hypothesis is β, then (assuming approximate normality of the IV estimate) the sample size required to test a causal effect of size β_1 is [Freeman et al., 2013]:

$$N = \frac{(z_{(1-\frac{\alpha}{2})} + z_\beta)^2 \text{ var}(R_Y^{IV})}{\text{var}(X)\, \beta_1^2\, \rho_{GX}^2} \tag{8.3}$$

where the quantile function z_a is the $100a$ percentile point of a standard normal distribution. If the significance level is 0.05 and the power is 0.8, then the sample size required to test a standardized causal effect of β_{1s} (measured in units of standard deviations in Y per standard deviation increase in X) is approximately:

$$N = \frac{7.848}{\beta_{1s}^2\, \rho_{GX}^2}. \tag{8.4}$$

This assumes that the variance of Y is approximately equal to the variance of R_Y^{IV}, which will be true if the causal effect of X does not explain much of the variation in Y.

For a given sample size N, the power to detect a standardized causal effect (in the same direction as the true effect) can be calculated as:

$$\text{Power } = \Phi(\beta_{1s}\, \rho_{GX}\, \sqrt{N} - z_{(1-\frac{\alpha}{2})}) \tag{8.5}$$

where Φ is the cumulative distribution function of the standard normal distribution. This is the inverse function of the quantile function ($\Phi(z_a) = a$).

We use these formulae to construct power curves for Mendelian randomization using a two-sided significance level $\alpha = 0.05$. In Figure 8.1, we fix the squared correlation ρ_{GX}^2 at 0.02, meaning the variant explains 2% of the variance of the exposure, and vary the size of the standardized causal effect $\beta_{1s} = 0.05$ to 0.3 and the sample size $N = 1000$ to 10 000. In Figure 8.2, we fix the size of the standardized causal effect at $\beta_{1s} = 0.2$ and vary the

squared correlation $\rho^2_{GX} = 0.005$ to 0.03 and the sample size as before. In each of the figures, the power to detect a positive causal effect is displayed; this tends to 0.025 as the sample size tends to zero. Sample sizes of several thousands are required to achieve adequate power in settings typical for many Mendelian randomization studies (modest causal effects, low correlation of genetic variants with exposure). Code to customize these calculations for a given scenario is available [Brion et al., 2013] together with an online calculator (http://glimmer.rstudio.com/kn3in/mRnd/).

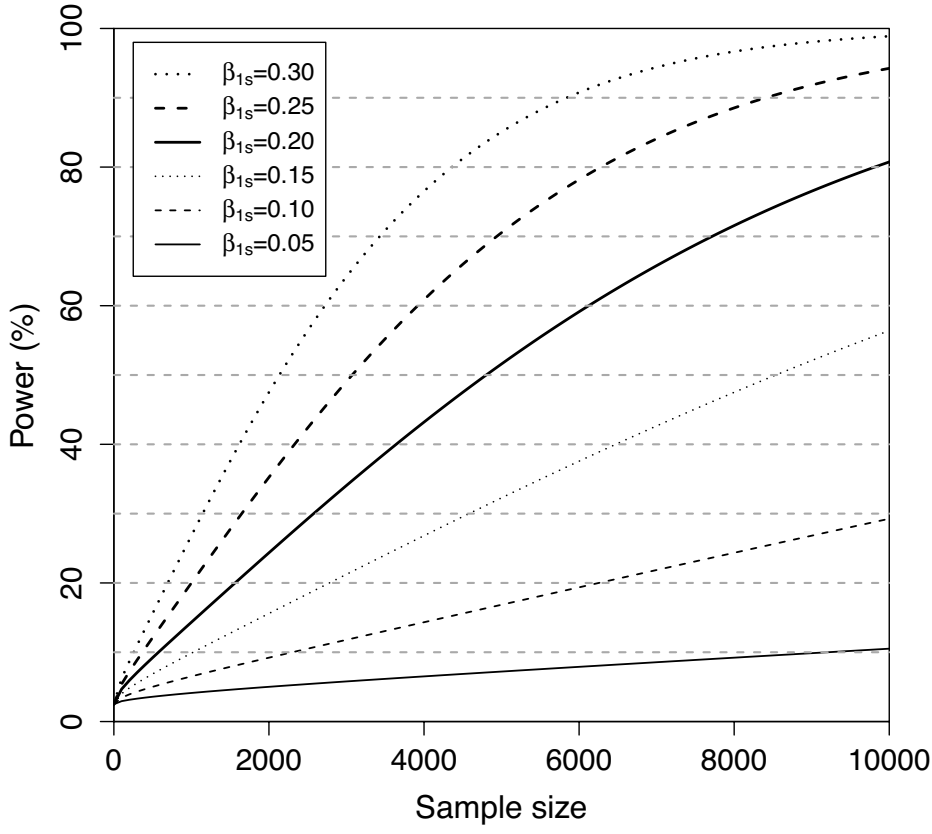


FIGURE 8.1
Power curves with two-sided significance level $\alpha = 0.05$ varying the sample size for a fixed value of the IV strength ($\rho^2_{GX} = 0.02$) and different values of the size of the standardized causal effect ($\beta_{1s} = 0.05$ to 0.3) with a single IV.

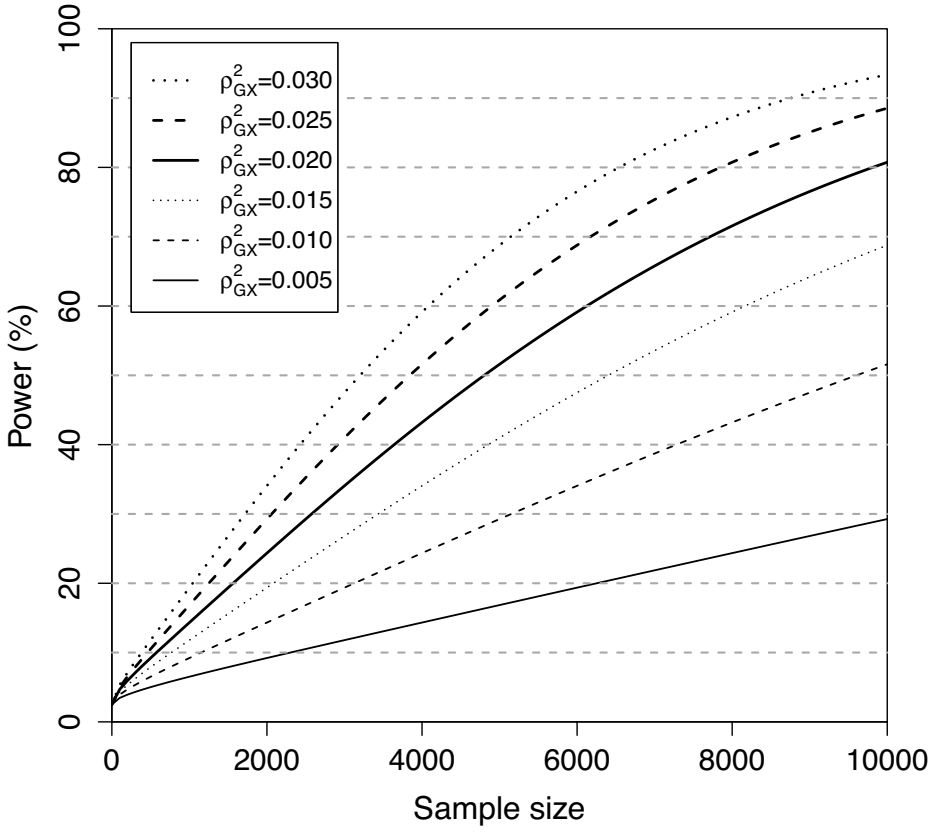


FIGURE 8.2
Power curves with two-sided significance level $\alpha = 0.05$ varying the sample size for a fixed size of standardized causal effect ($\beta_{1s} = 0.2$) and varying the value of the IV strength ($\rho^2_{GX} = 0.005$ to 0.3) with a single IV.

8.3.2 Power with a single IV, binary outcome

With a single IV, the asymptotic variance of the IV estimate of the causal effect of the exposure X on the outcome Y with a single IV G can be approximated using the delta method for the ratio estimate. The leading term in the expansion is:

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}(\hat{\beta}_{Y|G})}{\hat{\beta}^2_{X|G}} \tag{8.6}$$

where $\hat{\beta}_{Y|G}$ and $\hat{\beta}_{X|G}$ are the genetic association estimates with the outcome and exposure respectively.

The sample size for an IV analysis can therefore be approximated by considering the variance of the coefficient $\hat{\beta}_{Y|G}$. Assuming the outcome is binary ($Y = 0$ or 1) and using a logistic regression model to obtain $\hat{\beta}_{Y|G}$, the variance of the IV estimate is approximately:

$$\text{var}(\hat{\beta}_1) = \frac{1}{N \text{ var}(X) \rho^2_{GX} \mathbb{P}(Y = 1) \mathbb{P}(Y = 0)} \tag{8.7}$$

where $\mathbb{P}(Y = 1)$ and $\mathbb{P}(Y = 0)$ are the probabilities of the two outcomes for Y in the sample population (so the proportions of cases and controls in a case-control study).

The sample size required to detect a standardized causal effect of size β_{1s} (the log odds ratio per standard deviation increase in X) with 80% power and a two-sided significance level of $\alpha = 0.05$ is therefore:

$$N = \frac{7.848}{\beta^2_{1s} \rho^2_{GX} \mathbb{P}(Y = 1) \mathbb{P}(Y = 0)}. \tag{8.8}$$

If there are to be an equal number of cases and controls, $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = 0.5$, and:

$$N = \frac{31.39}{\beta^2_{1s} \rho^2_{GX}}. \tag{8.9}$$

The corresponding power to detect a standardized causal effect of size β_{1s} with a two-sided significance level of 0.05 is:

$$\text{Power } = \Phi(\beta_{1s} \rho_{GX} \sqrt{N \mathbb{P}(Y = 1) \mathbb{P}(Y = 0)} - 1.96). \tag{8.10}$$

We use these formulae to calculate the number of cases needed to obtain 80% power at $\alpha = 0.05$ in a Mendelian randomization analysis with a binary outcome for different values of β_{1s} and ρ^2_{GX}, assuming a 1:1 ratio of cases to controls. The results are displayed in Figure 8.3. It is evident that in most realistic Mendelian randomization contexts (moderate causal odds ratio, low correlation of genetic variants with exposure), many thousands of cases are required to achieve adequate power.

These formulae can be used by investigators planning a Mendelian randomization study, or to assess whether their study has adequate power to

detect a causal effect of a given magnitude. Code to customize these calculations for a given scenario is available [Burgess, 2014] together with an online calculator (http://spark.rstudio.com/sb452/power/).

8.3.3 Power with multiple IVs

Simulation studies have been performed to estimate power and sample sizes required with multiple IVs using the 2SLS method [Pierce et al., 2011]. An advantage of the use of simulation studies in this context is the reliance of analytical methods on simplifications and approximations. For example, the expression (8.6) does not take into account uncertainty in the genetic association with the exposure. Asymptotic approximations for the variance of the ratio estimator assume that IV estimates follow normal distributions. This is known to underestimate the variability of estimates, particularly if the IV is weak. However, comparisons between analytical and simulation approaches have generally shown a good level of agreement [Freeman et al., 2013]. In the absence of confounding, when the coefficient of determination (R^2) in the regression of the exposure on the IVs is constant, varying the number of variants does not seem to affect the power [Pierce et al., 2011]. When there is confounding, using additional variants increases weak instrument bias, making the comparison of power levels using the 2SLS method problematic.

However, in practice, using multiple variants will also increase the proportion of the variance in the exposure explained by the IVs. As shown in the previous two sections, gains in power from increasing the strength of the IV are substantial, giving motivation to researchers to find and use multiple variants in Mendelian randomization analyses.

8.4 Multiple variants and missing data

We illustrate the gain in precision from using multiple genetic variants and the problems of missing data (particularly genetic data) using the British Women's Heart and Health Study (BWHHS), one of the constituent studies of the CRP CHD Genetics Collaboration (CCGC).

8.4.1 Data from the British Women's Heart and Health Study

We examine the causal effect of C-reactive protein (CRP) on fibrinogen using three single nucleotide polymorphisms (SNPs) in the *CRP* gene coding region as IVs: rs1205, rs1130864, and rs1800947. Although log(CRP) and fibrinogen are positively correlated ($r = 0.45$), it is not thought that long-term variation

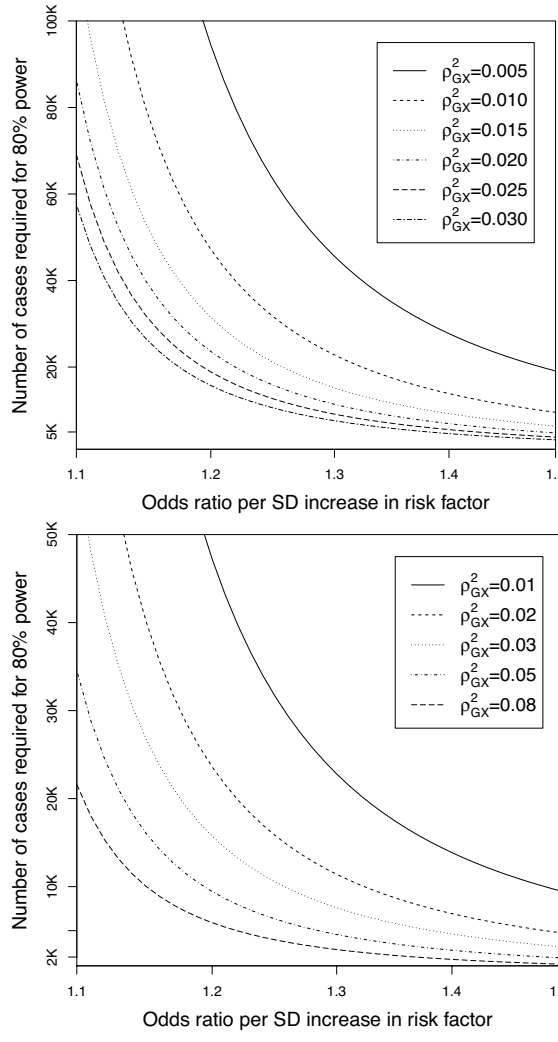


FIGURE 8.3
Number of cases required (assuming an equal number of controls) in a Mendelian randomization analysis with a binary outcome and a single instrumental variable for 80% power with a 5% significance level varying the size of the standardized causal effect (odds ratio per standard deviation increase in exposure) for different values of IV strength (ρ^2_{GX}: 0.005 to 0.03 in top panel, 0.01 to 0.08 in bottom panel).

in CRP is causally associated with levels of fibrinogen. As CRP has a skewed distribution, a linear association is assumed between log-transformed CRP and fibrinogen.

Each of the SNPs has some missing data. We use cross-sectional baseline data on 3693 participants with CRP and fibrinogen data, who have complete or partial data for the three SNPs. There is missingness in 10.8% of participants for rs1205, 1.9% for rs1130864, and 2.6% for rs1800947. Genotyping was undertaken on two separate occasions for SNP rs1205, and then for SNPs rs1130864 and rs1800947. 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 participants have complete data on all the SNPs. In these complete data, the F statistic in a multiple regression of log(CRP) on all the SNPs is 16.7. The Sargan overidentification test (Section 4.5.3) gives $p = 0.72$, indicating that there is no more heterogeneity between the causal estimates using different IVs than would be expected by chance.

Table 8.1 gives the estimates of causal effect from a Bayesian method using an additive per allele genetic model (Section 4.3.3). For each SNP, the causal effect is given both using all participants with data on the given SNP, and for the 3188 individuals with complete data on all three SNPs. We see that, considering the data on participants with complete data, using all the SNPs as the IV gives the most precise estimate, with at least a 34% reduction in standard error compared to the estimate using any of the SNPs individually. However, a substantial proportion of the data has been discarded in the complete-case analysis. 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 the standard error of the causal estimate. Although the gain in precision is not uniform across all SNPs, with a slight loss of precision in the causal estimate using SNP rs1800947 as the IV despite a sample size increase of 396, these results motivate us to use methods for incorporating individuals with missing data.

8.4.2 Power and missing data

Power can be increased in a study by including individuals with partially missing data in an analysis. Although missing data is not a problem which is unique to Mendelian randomization, missing genetic data is a specific problem in this context. Mendelian randomization studies often have limited power, and so excluding participants due to the presence of missing data is not the best strategy if they provide information on the causal effect. Additionally, if there are multiple genetic variants which can be used as IVs, the aim would be to include all available genetic variants, but not to exclude participants with missing data on some of the available SNPs.

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,

SNP	N	Participants with complete data on SNP (sample size = N)	Participants with complete data on all SNPs (sample size = 3188)
rs1205	3283	0.03 (0.40)	0.02 (0.49)
rs1130864	3609	−0.15 (0.34)	−0.27 (0.43)
rs1800947	3584	−0.22 (0.43)	−0.17 (0.41)
All three	3188		−0.10 (0.27)

TABLE 8.1
Estimate and standard error of causal effect of unit increase in log(CRP) on fibrinogen (μmol/l) using various SNPs as IVs: analyses for participants (N) with complete data on SNP used as IV in analysis, and for participants with complete data on all SNPs.

DNA extracted may be of insufficient quality or quantity for analysis, or the reading from a genotyping platform may be difficult to interpret. In the first three cases, no genetic data would be available for the individual, and they would not contribute greatly to the estimation of the causal effect. In the fourth case, data may be available for several individuals on some variants, but a missing result may be reported for one or more variants. By imputing missing genetic data, we can include all participants in an IV analysis using all the genetic variants as IVs, while appropriately acknowledging uncertainty in the imputation. If the genetic variants are highly correlated (in high linkage disequilibrium, LD), then the imputation of missing genetic data may be possible with little uncertainty, and there may be little loss of precision compared to a hypothetical complete-data analysis if all the data were available (or little over-precision if the uncertainty in the imputation procedure is ignored).

8.4.3 Methods for incorporating missing data

Here we present a brief description of IV methods for handling missing data; further details for interested readers are available elsewhere [Burgess et al., 2011a]. A difficulty with the two-stage method is that uncertainty in the first-stage regression is not acknowledged in the second-stage regression even without missing data (Section 4.3.5). There is no clear way to account for uncertainty in imputed data in the first-stage regression. This is not a difficulty in likelihood-based methods, such as in a Bayesian framework. Likelihood-based methods usually assume that data are "missing at random" (MAR), meaning that the probability that a data value is missing depends only on the observed data values of the measured variables [Little and Rubin, 2002].

Genetic data can be imputed using many software packages, including Beagle [Browning, 2006; Browning and Browning, 2007] and fastPHASE [Scheet and Stephens, 2006]. Output from these packages can be obtained in the form

of posterior probabilities of the number of variant alleles for each SNP in individuals, or as imputed datasets randomly drawn from the same posterior distributions. Either multiple imputed datasets (multiple imputations method), or the posterior probabilities (SNP imputation method) can be used as inputs in an analysis model. Both approaches acknowledge uncertainty in the imputation process; however, as the imputation and analysis models are performed separately, there is no feedback between the two stages. Alternatively, imputation can be performed as part of the analysis model using a latent variable approach, modelling the haplotypes using a multivariate normal distribution (latent variable method) [Lunn et al., 2006], or by modelling the probability of an individual having a given set of haplotypes directly using knowledge about the structure of the data and the prevalence of known haplotype patterns (haplotype imputation method).

8.4.4 Results of missing data analyses

We apply the four imputation methods sketched out above. Each of the methods gives fairly similar answers; the point estimates are all nearer zero than that from the complete-case analysis (Table 8.2). The reduction in the standard error for all missing data methods compared to the complete-case analysis is 8–12%. Assuming that the precision (= 1/variance) of the causal estimate increases proportionally to the sample size, this corresponds to a 17–29% increase in effective sample size, slightly more than the true increase in sample size of 16% (3693 compared to 3188 individuals).

It is perhaps surprising to find a gain in precision more than anticipated from the gain in sample size. However, the increase in sample size within each of the genetic subgroups is not uniform. In this example, 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. Although this may be simply good fortune, heterozygotes and minor homozygotes are less easy to determine from the output of genotyping platforms, and so this may not be an isolated case.

8.5 Discussion

In this chapter, we have considered using multiple instruments in IV analyses. Using multiple instruments has the potential to reduce the standard error of causal estimates, but if there are sporadically missing genetic data, this increase is offset by a decrease in sample size in a complete-case analysis.

Imputation method	Effect (SE)	95% confidence interval
Complete case analysis	−0.10 (0.27)	−0.70, 0.38
Multiple imputations	−0.09 (0.25)	−0.62, 0.36
SNP imputation	−0.07 (0.25)	−0.61, 0.37
Latent variable method	−0.04 (0.24)	−0.55, 0.40
Haplotype imputation	−0.06 (0.25)	−0.59, 0.39

TABLE 8.2
Estimate, standard error (SE) and 95% confidence interval of the causal effect for a unit increase in log(CRP) on fibrinogen (μmol/l) in a complete-case analysis ($N = 3188$) and in the entire study population ($N = 3693$) using different imputation methods for missing genetic data in the British Women's Heart and Health Study.

8.5.1 Heterogeneity and supplementary analyses

In using multiple genetic variants to estimate a single causal effect, the assumption is made that the causal effect identified by each of these variants is the same (known as 'no treatment effect heterogeneity'). This may not be true, even if the variants are all valid IVs. Differences may occur if there are multiple mechanisms by which the exposure affects the outcome. For example, variants may be associated with body mass index (BMI) by various mechanisms, such as suppressing appetite or increasing metabolic rate. If genetic variants can be categorized as associated with one or other of these mechanisms, then separate Mendelian randomization estimates can be obtained using each category of variants. A Mendelian randomization estimate constructed using variants associated with BMI through appetite suppression more closely represents the causal effect of intervening on BMI via appetite suppression [Hernán and Taubman, 2008]. Differences in the causal estimates using genetic variants associated with different mechanisms may be informative in understanding the aetiology of the disease, and may highlight specific mechanisms to prioritize for pharmacological intervention (Section 6.3.1).

8.5.2 Subsample Mendelian randomization

Especially when the outcome is binary, the sample size required in a Mendelian randomization experiment may be prohibitively large due to the expense of collecting data on the exposure. In this case, a subsample IV approach may be a cost-effective approach. Rather than collecting data on the exposure from the entire study sample, exposure data can be measured for a random subsample of (control) participants. As the association between the IV and the exposure is typically stronger than that between the IV and the outcome, the precision of the IV estimate may not be noticeably affected by reducing the sample size

on which the exposure is measured. Simulations have shown that a subsample IV analysis with exposure data on only 10% of participants may retain 90% of the power of the full-sample IV analysis [Pierce and Burgess, 2013]. Estimates and confidence intervals with a single IV can be calculated using the ratio method and Fieller's theorem (Section 4.1.5). With multiple IVs, a modified version of the two-stage least squares method can be used [Inoue and Solon, 2010].

8.5.3 Relevance to epidemiological practice

The conclusion of Chapter 7 was that problems due to weak instruments, while potentially serious, are surmountable. Bearing in mind the advice of Chapter 7, multiple instrumental variables provide an opportunity to obtain more precise estimates of causal effects. One particular way of incorporating multiple instrumental variables into an analysis which avoids the danger of weak instrument bias is the use of an allele score, although care must be taken in the construction of the score so as not to introduce bias.

8.6 Key points from chapter

- Use of multiple instrumental variables in Mendelian randomization leads to more precise estimates of causal effects.
- Sporadically missing genetic data may offset this gain, but missing data methods can recover much of the loss.
- Parsimonious models of genetic association, and in particular allele scores, can alleviate the problems of weak instruments which may arise when using large numbers of instrumental variables.
- The procedure for constructing an allele score to be used in an analysis should be made clear, and in particular how variants and weights for the score are chosen, as this has a considerable impact on bias.

9

Multiple studies and evidence synthesis

In this chapter, we consider extensions to a simple Mendelian randomization analysis to include data from multiple studies. We provide methods for combining the information provided by each study in an efficient way to produce a single causal estimate. Also, we consider how to combine summarized data on genetic associations from multiple variants in a single study.

9.1 Introduction

In general, the variation in the exposure of interest explained by genetic variants in Mendelian randomization is small, and so adequately powered investigations typically require large sample sizes. This often demands synthesis of evidence from multiple, possibly heterogeneous studies.

In this chapter, we first consider assessment of the causal relationship using data from multiple studies (Section 9.2) before proceeding to methods for estimating a pooled causal effect. Methods are presented in order of the homogeneity and detail of data required from each constituent study. A study-level meta-analysis requires the least detail, combining the causal effect estimates obtained in each study (Section 9.3). However, in order to estimate such a pooled causal effect, each study needs to measure data on genetic variants, the exposure and the outcome. A summary-level meta-analysis requires more detailed information from studies, including information which may not be routinely reported in a published paper but is increasingly being made available by large consortia (Section 9.4). An individual-level meta-analysis requires individual participant data (IPD) from studies (Section 9.5). However, individual-level models are the most flexible for addressing the heterogeneity of data available in each study. The methods are illustrated and compared using real data (Section 9.6). We discuss extensions to the meta-analysis model for binary outcomes (Section 9.7), and conclude with a discussion of application of the methods presented in practice (Section 9.8).

9.2 Assessing the causal relationship

In Section 3.3, we drew a distinction between assessing a causal relationship and estimating a causal effect. If assessment of a causal relationship is sufficient, with a single genetic variant or an allele score as the sole instrumental variable (IV), a causal relationship can be inferred by undertaking a meta-analysis of the IV–outcome regression coefficients from each of the studies. Standard inverse-variance weighted methods for meta-analysis are described in many basic texts; software for performing such analyses is available and well-documented [Borenstein et al., 2009]. A pooled estimate away from the null is indicative of a causal relationship.

If genetic variants used as IVs (G) in each study have different magnitudes of association with the exposure (X), the causal effect of the exposure on the outcome (Y) can be examined visually by plotting a graph of the regression estimates for the G–Y association against the regression estimates for the G–X association [Minelli et al., 2004]. The points on this graph will be subject to error in both associations and the gradient of the graph will show the causal X–Y association. This will be similar to Figure 6.1, although the points will represent different genetic variants in multiple studies.

9.3 Study-level meta-analysis

If it is possible to estimate the causal effect in each study, a study-level meta-analysis can be performed directly on these estimated causal effects, for example using inverse-variance weighting. However, if the genetic association with the exposure is small or is measured imprecisely, the asymptotic variance estimates from each study used in the meta-analysis may be unreliable measures of uncertainty (Section 7.4.2). Additionally, meta-analysis based on study-level causal effect estimates tends to exaggerate weak instrument bias (Section 7.7.3).

9.4 Summary-level meta-analysis

It may be that some studies only provide data on one of the exposure or the outcome, and so a causal effect cannot be estimated from that study alone. Even if some study-level causal estimates can be calculated, a more precise estimate of the pooled causal effect can be obtained using summary-level data,

such as estimates of the associations between the genetic variant(s) and each of the exposure and outcome, or the mean level of the exposure and outcome in each genetic subgroup. A pooled estimate can be evaluated by combining data in a hierarchical model, which we now describe.

9.4.1 Multiple genetic variants in a single study

Before considering the possibility of using summary-level data in a meta-analysis, we explore their use in a single study with multiple genetic variants.

We assume that the estimate of association for genetic variant $k = 1, \ldots, K$ with the exposure is $\hat{\beta}_{Xk}$ with standard error σ_{Xk}, and the estimate of association with the outcome is $\hat{\beta}_{Yk}$ with standard error σ_{Yk}. (The standard error parameters are assumed to be estimated without uncertainty.) When a study is used in estimating both the gene–exposure and the gene–outcome associations, these estimates will be correlated. We assume that these association estimates can be modelled by a bivariate normal distribution, with correlation ρ assumed to be the same for each variant:

$$\begin{pmatrix} \hat{\beta}_{Xk} \\ \hat{\beta}_{Yk} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \xi_k \\ \eta_k \end{pmatrix}, \begin{pmatrix} \sigma^2_{Xk} & \rho\,\sigma_{Xk}\,\sigma_{Yk} \\ \rho\,\sigma_{Xk}\,\sigma_{Yk} & \sigma^2_{Yk} \end{pmatrix} \right). \tag{9.1}$$

A linear association is assumed between the underlying unmeasured means ξ_k and η_k. As the genetic association with the outcome is assumed to be zero if the association with the exposure is zero (from the IV assumptions), we have:

$$\eta_k = \beta_1\,\xi_k. \tag{9.2}$$

The causal effect β_1 is assumed to be the same for all variants (Section 8.5.1). This and subsequent models in this chapter can be estimated either by numerical maximization of the log-likelihood function or by Bayesian methods [Thompson et al., 2005], the latter for example using WinBUGS [Spiegelhalter et al., 2003] or MLwiN [Rasbash et al., 2009].

The correlation ρ can be estimated as part of the analysis, but there is likely to be little information on the parameter in the data [Riley et al., 2007]. We recommend that the value of the parameter be specified as part of the model, and a sensitivity analysis performed to assess the effect of varying this parameter value on estimates; its value should be similar to the observational correlation between the exposure and outcome.

By combining the estimates of association from multiple variants into a single estimate of the causal effect, an assumption is made that the variants provide independent information on the causal effect. If the association estimates are derived from the same data, then they will not be independent. However, if the variants are independently distributed (that is, they are not in linkage disequilibrium, LD), correlation between these estimates should be low unless the sample size is particularly small. Simulations using independently distributed variants have shown that estimates from the summary-level

data model (9.1) and (9.2) are well-behaved even in the presence of statistical interactions between the effects of the genetic variants (gene–gene interactions). Similar weak instrument bias was observed to estimates from the two-stage least squares (2SLS) method, and confidence intervals were appropriately sized with correct coverage rates. The efficiency of estimates based on summary-level data was similar to that of estimates based on individual-level data [Burgess et al., 2013].

If the genetic variants are correlated in their distributions (that is, they are in LD), then the association estimates $\hat{\beta}_{Xk}$ $(k = 1, \ldots, K)$ will be correlated, as will $\hat{\beta}_{Yk}$ $(k = 1, \ldots, K)$. This can be accounted for in the likelihood model by a multivariate normal distribution for the genetic association estimates from each variant using estimates of the correlations between the variants (which will be the same as the correlations between the association estimates) [Burgess et al., 2014e]. This extension is not discussed further here. If the variants are correlated, then estimates from equation (9.1) will overstate precision.

For a single study, a further method has been developed for combining summary-level data on multiple genetic variants not in LD [Johnson, 2011]. This method combines the ratio estimates $\frac{\hat{\beta}_{Yk}}{\hat{\beta}_{Xk}}$ from each variant in an inverse-variance weighted meta-analysis using asymptotic variances calculated from the delta method for the ratio of two random variables [Dastani et al., 2012]. This variance is:

$$\frac{\sigma^2_{Yk}}{\hat{\beta}^2_{Xk}}. \tag{9.3}$$

This differs from the formula for the standard error of the ratio estimate given in equation (4.9) as the uncertainty in the genetic association with the exposure is assumed to be zero.

The combined inverse-variance weighted (IVW) estimate $\hat{\beta}_{IVW}$ is:

$$\hat{\beta}_{IVW} = \frac{\sum_k \hat{\beta}_{Xk}\hat{\beta}_{Yk}\sigma^{-2}_{Yk}}{\sum_k \hat{\beta}^2_{Xk}\sigma^{-2}_{Yk}}. \tag{9.4}$$

The approximate standard error of the estimate is:

$$\text{se}(\hat{\beta}_{IVW}) = \sqrt{\frac{1}{\sum_k \hat{\beta}^2_{Xk}\sigma^{-2}_{Yk}}} \tag{9.5}$$

As the method assumes the ratio estimates are normally distributed, and as the uncertainty in the genetic associations with the exposure is not accounted for, the precision of IVW estimates is overstated. However, simulations have shown that the underestimation of confidence intervals may be slight, with an average 93% coverage probability for a nominal 95% confidence interval in a plausibly realistic scenario [Burgess et al., 2013]. Therefore the IVW method may be a reasonable simpler alternative to a likelihood-based model when

the genetic associations with the exposure are estimated precisely. However, likelihood-based models should be preferred for use in practice where possible, particularly if the genetic associations with the exposure are estimated imprecisely.

9.4.2 Single genetic variant in multiple studies

If a single genetic variant (potentially different between studies) is measured in multiple studies, the same likelihood-based model as above can be used to combine the genetic association estimates from each study into a single pooled causal estimate.

For each study $m = 1, \ldots, M$, the estimated G–X association $\hat{\beta}_{Xm}$ is assumed to be normally distributed with mean ξ_m and variance σ^2_{Xm} and the estimated G–Y association $\hat{\beta}_{Ym}$ is normally distributed with mean η_m and variance σ^2_{Ym}. The correlation ρ between $\hat{\beta}_{Xm}$ and $\hat{\beta}_{Ym}$ is assumed to be independent of m. This is identical to equations (9.1) and (9.2) except for the change in the subscripted index.

$$\begin{pmatrix} \hat{\beta}_{Xm} \\ \hat{\beta}_{Ym} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \xi_m \\ \eta_m \end{pmatrix}, \begin{pmatrix} \sigma^2_{Xm} & \rho\,\sigma_{Xm}\,\sigma_{Ym} \\ \rho\,\sigma_{Xm}\sigma_{Ym} & \sigma^2_{Ym} \end{pmatrix} \right) \qquad (9.6)$$

$$\eta_m = \beta_1\,\xi_m$$

Alternatively, the IVW method can be used, although this is equivalent to a study-level meta-analysis rather than a summary-level meta-analysis.

9.4.3 Single common genetic variant in multiple studies

If the same single genetic variant has been measured in each of the studies then, in principle, the same within-study model (9.6) could be used to combine the genetic association estimates from each study. However, this would not take into account the fact that the genetic variant is the same in each study. A hierarchical model is therefore proposed, whereby the G–X and G–Y association parameters are additionally pooled in a second-level (or between-study) model. This approach also allows the inclusion of studies where only one of the exposure or outcome have been measured.

Initially, we assume a fixed-effect meta-analysis model; random-effects models are considered later. For each study $m = 1, \ldots, M$ measuring both the G–X and G–Y associations, the estimated G–X association $\hat{\beta}_{Xm}$ is assumed to be normally distributed with mean ξ (the same for each study) and variance σ^2_{Xm} and the estimated G–Y association $\hat{\beta}_{Ym}$ is normally distributed with mean $\eta = \beta_1\xi$ and variance σ^2_{Ym}.

$$\begin{pmatrix} \hat{\beta}_{Xm} \\ \hat{\beta}_{Ym} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \xi \\ \eta \end{pmatrix}, \begin{pmatrix} \sigma^2_{Xm} & \rho\,\sigma_{Xm}\,\sigma_{Ym} \\ \rho\,\sigma_{Xm}\sigma_{Ym} & \sigma^2_{Ym} \end{pmatrix} \right) \qquad (9.7)$$

To include studies where only one of the G–X and G–Y associations has

been reported, we use the marginal distribution of $\hat{\beta}_{Xm}$ or $\hat{\beta}_{Ym}$ as appropriate. For example:

$$\hat{\beta}_{Xm} \sim \mathcal{N}(\xi, \sigma^2_{Xm}). \tag{9.8}$$

Estimation proceeds by direct maximization of the log-likelihood function or by Bayesian methods, as before [Thompson et al., 2005].

9.4.4 Multiple genetic variants in multiple studies – Genetic associations

If multiple, potentially different genetic variants are measured in multiple studies, equation (9.6) can be extended to a hierarchical model:

$$\begin{pmatrix} \hat{\beta}_{Xkm} \\ \hat{\beta}_{Ykm} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \xi_{km} \\ \eta_{km} \end{pmatrix}, \begin{pmatrix} \sigma^2_{Xkm} & \rho\,\sigma_{Xkm}\,\sigma_{Ykm} \\ \rho\,\sigma_{Xkm}\,\sigma_{Ykm} & \sigma^2_{Ykm} \end{pmatrix} \right) \tag{9.9}$$
$$\eta_{km} = \beta_1\,\xi_{km}$$

where $k = 1, \ldots, K_m$ indexes genetic variants (first level of the hierarchical model) and m indexes studies (second level). We initially assume that the causal effect β_1 takes the same value in each study (fixed-effect model); in other words the same parameter β_1 is estimated regardless of which genetic variants are measured and of how many variants are available in each study.

9.4.5 Multiple genetic variants in multiple studies – Genetic subgroups

An alternative way of modelling based on summary-level data is to partition the population into genetic subgroups, each of which contains all the individuals in the study with the same genotype for the measured variants. We index genetic subgroups by the subscript $j = 1, \ldots, J_m$. For each study m, the mean of the exposure ($\bar{X}_{jm}$) in each subgroup is assumed to come from a normal distribution with mean ξ_{jm} and known variance σ^2_{Xjm}. Similarly, the mean of the outcome ($\bar{Y}_{jm}$) in each subgroup is assumed to come from a normal distribution with mean η_{jm} and known variance σ^2_{Yjm}.

$$\begin{pmatrix} \bar{X}_{jm} \\ \bar{Y}_{jm} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \xi_{jm} \\ \eta_{jm} \end{pmatrix}, \begin{pmatrix} \sigma^2_{Xjm} & \rho\,\sigma_{Xjm}\,\sigma_{Yjm} \\ \rho\,\sigma_{Xjm}\,\sigma_{Yjm} & \sigma^2_{Yjm} \end{pmatrix} \right) \tag{9.10}$$
$$\eta_{jm} = \beta_{0m} + \beta_1\,\xi_{jm}$$

This model is appropriate even if the genetic variants are in LD, as the genetic subgroups are defined using all the variants, and so the mean values of the exposure and outcome in the genetic subgroups are independent. However, unlike estimates of genetic association, data on the mean values of the exposure and outcome in genetic subgroups are unlikely to be routinely reported in publications.

9.4.6 Fixed- and random-effects meta-analysis

The models given so far in this chapter represent fixed-effect meta-analyses, as the same value of ξ (equation 9.7) or β_1 (equations 9.9 and 9.10) is assumed for each study. For a random-effects meta-analysis, we allow study-specific parameters ξ_m or β_{1m} to vary between studies, but model them as coming from a common distribution. This acknowledges the possibility that the 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.

Equation (9.7) can be extended to a random-effects model by allowing the study-specific parameters ξ_m and η_m to come from normal distributions:

$$\begin{aligned}\begin{pmatrix} \hat{\beta}_{Xm} \\ \hat{\beta}_{Ym} \end{pmatrix} &\sim \mathcal{N}_2 \left(\begin{pmatrix} \xi_m \\ \eta_m \end{pmatrix}, \begin{pmatrix} \sigma^2_{Xm} & \rho\,\sigma_{Xm}\,\sigma_{Ym} \\ \rho\,\sigma_{Xm}\,\sigma_{Ym} & \sigma^2_{Ym} \end{pmatrix} \right) \\ \xi_m &\sim \mathcal{N}(\mu_\xi, \tau^2_\xi) \\ \eta_m &\sim \mathcal{N}(\mu_\eta, \tau^2_\eta).\end{aligned} \tag{9.11}$$

Pooling of these parameters assumes that the gene–exposure and gene–outcome associations are similar in each study. The causal effect estimate is the ratio of the means of the random-effects distributions for ξ_m and η_m: $\hat{\beta}_1 = \frac{\hat{\mu}_\eta}{\hat{\mu}_\xi}$ [Thompson et al., 2005]. The variance parameters τ^2_ξ and τ^2_η are measures of between-study heterogeneity.

If different genetic variants are measured in each study, then a more generalizable way of modelling heterogeneity between studies is by allowing study-specific causal effect parameters β_{1m} to come from a common distribution; in particular, a normal distribution with mean β_1 and variance τ^2. In equation (9.9), a random-effects meta-analysis model is achieved by replacing the last line by:

$$\begin{aligned}\eta_{km} &= \beta_{1m}\,\xi_{km} \\ \beta_{1m} &\sim \mathcal{N}(\beta_1, \tau^2).\end{aligned} \tag{9.12}$$

In equation (9.10), for a random-effects meta-analysis the last line is replaced by:

$$\begin{aligned}\eta_{jm} &= \beta_{0m} + \beta_{1m}\,\xi_{jm} \\ \beta_{1m} &\sim \mathcal{N}(\beta_1, \tau^2).\end{aligned} \tag{9.13}$$

Additionally, the correlation parameter ρ could be replaced by study-specific parameters ρ_m, which could be specified or estimated separately in each study, and combined in a random-effects distribution if required. If $\tau = 0$, then a fixed-effect model is recovered.

We generally advocate random-effects models rather than fixed-effect models in applied investigations, as the assumption of homogeneity for a parameter across studies is usually unrealistic. If there is not much heterogeneity

between the studies, then the value of the heterogeneity parameter will be close to zero, and the random-effects analysis will approximate a fixed-effect analysis. If there is considerable heterogeneity, then this provides evidence against the fixed-effect model and in favour of the random-effects model. A fixed-effect model may be used if there is a strong argument why a parameter may be similar across studies (for example, if the separate studies were in fact centres in a clustered investigation using the same protocol and sampling individuals from the same population). If there are few studies then it may be difficult to obtain a precise estimate of heterogeneity, and either a fixed-effect model or an informative prior on the heterogeneity parameter in a Bayesian analysis may be employed. A fixed-effect model may also be useful in comparing between meta-analysis coefficients from different analyses to ensure that the differences were not simply due to changes in the heterogeneity estimate.

The hierarchical nature of the above models is now clear: at the first (within-study) level, a causal parameter (β_{1m}) is specified in each study, and at the second (between-study) level, these causal parameters are pooled to provide a single causal estimate (β_1). By evaluating the estimates using a likelihood-based method in a single model, the meta-analysis is performed in a single step. This is in contrast with a two-step meta-analysis, as described in Section 9.3, in which the causal estimates are first estimated in each study, and then the estimates are combined. By performing the analysis in a single step, uncertainty in the model is correctly acknowledged and feedback is allowed between the two levels of the model.

9.4.7 Using published summary-level data

Several consortia with large numbers of participants, such as CARDIoGRAMplusC4D for coronary artery disease [Schunkert et al., 2011] and DIAGRAM for type 2 diabetes [Morris et al., 2012], have published summary-level data on the association of catalogues of genetic variants with either risk factors or disease status. These provide precise estimates of genetic associations which can be used to obtain causal estimates, provided the genetic variants included in the analysis are restricted to those for which the IV assumptions are valid.

The main advantages of using published data in Mendelian randomization are their size and scope. Large meta-analyses of genome-wide association studies (GWAS) have discovered many genetic variants associated with various risk factors which are candidate instrumental variables. The associations of these variants with the exposure and outcome in large consortia are likely to be more precisely estimated than in a single study or a more limited meta-analysis of available studies. However, it is unlikely that published data are available on the genetic associations with the exposure and with the outcome on the same set of studies. This may necessitate a two-sample Mendelian randomization analysis strategy (see Section 9.8.2), in which data on the genetic associations with the exposure and with the outcome are estimated on non-overlapping sets of individuals [Angrist and Krueger, 1992]. This simplifies

the models such as equation (9.1), as the correlation between the genetic association estimates with the exposure and with the outcome (ρ) would be zero.

It may be that gene–exposure and gene–outcome association estimates taken from the literature are not estimated from a single study, but themselves represent pooled estimates from meta-analyses. These can be combined across variants using the methods of Section 9.4.1. However, the heterogeneity across studies will not be modelled as faithfully as in a hierarchical model using the study-specific association estimates.

9.4.8 Advantages of summary-level meta-analysis

Summary-level meta-analysis provides a compromise between study-level and individual-level meta-analysis. In some cases, it may not be possible or feasible for a researcher to share individual-level data. By sharing summary-level data, evidence can be included even if a study only provides data on genetic association estimates or on genetic subgroups. These data may be available from published work, and should contribute information on the parameter representing the causal effect, as well as helping to avoid weak instrument bias by providing an alternative to a study-level analysis.

9.5 Individual-level meta-analysis

If individual participant data (IPD) are available, then rather than modelling summary-level data, we can model the individual-level data on the exposure and outcome directly. This enables us to consider the model of genetic association between the genetic variants and the exposure in more detail.

9.5.1 Modelling in a single study

We drop the subscript m where possible to improve readibility. We initially consider estimates using individual-level data in a single study, although there is a natural extension to a meta-analysis identical to the summary-level models considered in the previous section by pooling studies in a hierarchical model on the causal effect parameter. We index individuals by the subscript $i = 1, \ldots, N$ or $i = 1, \ldots, N_j$ as appropriate.

In the summary-level analyses, the assumption of exact knowledge of variances for each genetic association or mean value in a genetic subgroup is not strictly appropriate. Indeed, using genetic subgroups if $N_j = 1$, a group-specific estimate of variance cannot even be calculated. It is then preferable to base the analysis on the variance of the exposure (σ^2_X) and the outcome

(σ_Y^2) in the whole population, using an individual-based model. For example, equation (9.10) becomes

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \xi_j \\ \eta_j \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\,\sigma_X\,\sigma_Y \\ \rho\,\sigma_X\,\sigma_Y & \sigma_Y^2 \end{pmatrix} \right) \qquad (9.14)$$
$$\eta_j = \beta_0 + \beta_1\,\xi_j$$

where j indexes genetic subgroups and ρ denotes the observational correlation between the exposure and outcome.

9.5.2 Model of genetic association

In equation (9.14), separate parameters ξ_j and η_j are included for each genetic subgroup. If a specific model of genetic association is assumed, such as an additive per allele model, this can be included in the analysis model. If the model is correct, it should help to provide more precise estimates of the unknown parameters in the model and should reduce weak instrument bias (Section 7.5.2). If g_{ik} $(= 0, 1, 2)$ is the number of copies of the minor allele for genetic variant $k = 1, \ldots, K$ for individual i, we can write the model as:

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\,\sigma_X\,\sigma_Y \\ \rho\,\sigma_X\,\sigma_Y & \sigma_Y^2 \end{pmatrix} \right) \qquad (9.15)$$
$$\xi_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k\,g_{ik}$$
$$\eta_i = \beta_0 + \beta_1\,\xi_i$$

dropping the subscript j and the division into genetic subgroups. In model (9.14), there are up to 3^K genetic subgroups and an equal number of ξ_j parameters in the model of genetic association with the exposure (assuming each variant is a biallelic SNP and so takes three values). In model (9.15), there are $K + 1$ α_k parameters (one parameter for each genetic variant and an intercept parameter), meaning that the genetic association with the exposure is modelled more parsimoniously.

9.5.3 Common genetic variants

In a meta-analysis context, an additive per allele model for genetic association in each study m can be written as:

$$\xi_{im} = \alpha_{0m} + \sum_{k=1}^{K_m} \alpha_{km}\,g_{ikm} \qquad (9.16)$$

When the same set of genetic variants has been used in several studies, we can combine the estimates of genetic association α_{km} across studies, in the

same way as the parameter ξ_m was combined in equation (9.11). 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 (Section 7.6.3). Due to possible heterogeneity between populations, we propose a random-effects model, where we impose a multivariate normal distribution on the study level parameters $\boldsymbol{\alpha}_m = (\alpha_{km},\, k = 1, \ldots, K_m)$ with mean vector $\boldsymbol{\mu}_\alpha$ and variance-covariance matrix Ψ. Note that the intercept parameters α_{0m} are not pooled, as these depend on the characteristics of each study population and would not necessarily be similar across studies.

$$\begin{aligned} \xi_{im} &= \alpha_{0m} + \sum_{k=1}^{K} \alpha_{km}\, g_{ikm} \\ \boldsymbol{\alpha}_m &\sim \mathcal{N}_K(\boldsymbol{\mu}_\alpha, \Psi) \end{aligned} \tag{9.17}$$

9.5.4 Lack of exposure or outcome data

Where a study has not measured the exposure but has genetic data in common with other studies, we can use the random-effects distributions for the genetic association parameters defined above as a predictive distribution or implicit prior for the unknown parameters. This requires an assumption that the mean difference in exposure per additional allele is similar (i.e. can be drawn from the same random-effects distribution) to that in the other studies. For identifiability, we set $\alpha_{0m} = 0$ as with no data on the exposure, this parameter cannot be estimated. Alternatively, the exposure and IVs could be centered in each study, so that the intercept parameter α_{0m} would be equal to zero in each study by design. Studies without data on the outcome can be included in a meta-analysis in a similar way.

9.5.5 Advantages of individual-level meta-analysis

There are several advantages of analysing individual-level data in Mendelian randomization studies. The most important reasons are not related to the estimation of a causal effect, but rather concern the assessment of the assumptions necessary for the validity of the genetic variants as IVs. For example, with individual-level data on covariates, the associations of each genetic variant with measured potential confounders that might bias causal estimates can be tested. While these assessments can be performed using summary-level data, it is not usually possible to do this in a systematic way with a range of covariates.

Individual-level data enable more complex modelling of the genetic association with the exposure. This allows the pooling of genetic association parameters across studies and inclusion of studies without complete information on all of the genetic variants, exposure and outcome. As we shall see in the subsequent applied analysis of the CCGC dataset (Chapter 10), this

enables large gains in precision of causal effect estimates by the inclusion of additional data.

9.5.6 Combining summary- and individual-level data

In a practical setting, it may be the case that some studies are able to provide individual-level data and others are able to provide only summary-level data. The parameters for the genetic association with the exposure (α parameters) and the parameters for the exposure association with the outcome (β parameters) are the same in the summary- and individual-level models [Sutton et al., 2008]. Hence, if for example summary-level data are available for each of the genetic subgroups, the hierarchical model can include individual-level data from those studies for which they are available, and summary-level data from all other studies.

9.6 Example: C-reactive protein and fibrinogen

In Section 7.6.3, we provided estimates of the causal effect of C-reactive protein (CRP) on fibrinogen from five studies. However, the model used for analysis required the homogeneity of variance of the exposure and outcome in each study. We re-evaluate the same data using a hierarchical meta-analysis model. Mean levels of log-transformed CRP and fibrinogen in each of the genetic groups for the studies are shown in Figure 9.1. Due to the small number of studies, we use fixed-effect meta-analysis models for the causal effect parameters. An additive per allele genetic model was used throughout. For the individual-level models, the parameters of genetic association were estimated in three ways: with different, study-specific parameters; with parameters common across studies; and with parameters drawn from a random-effects distribution.

Estimation of the hierarchical models (for summary-level data, equation (9.10); for individual-level data, equation (9.15) with different genetic effects, equation (9.16) with common genetic effects, and equation (9.17) with random genetic effects) was performed in WinBUGS using vague priors (normal with mean zero and variance 1000^2, uniform on the interval [0,20] for positive valued parameters), except for the standard deviation parameter in the random-effects distributions for the parameters of genetic association, where a uniform prior distribution on the interval [0,1] was used.

Results are shown in Table 9.1. For comparison, the confounded observational association was 1.568 from a fixed-effect meta-analysis of the study-level observational estimates. We see that the study-level meta-analyses give the most positive causal effect estimates with the narrowest confidence intervals, in line with the comments of Section 7.6.3 on the effect of weak instruments on study-level meta-analysis. Using the LIML method rather than 2SLS

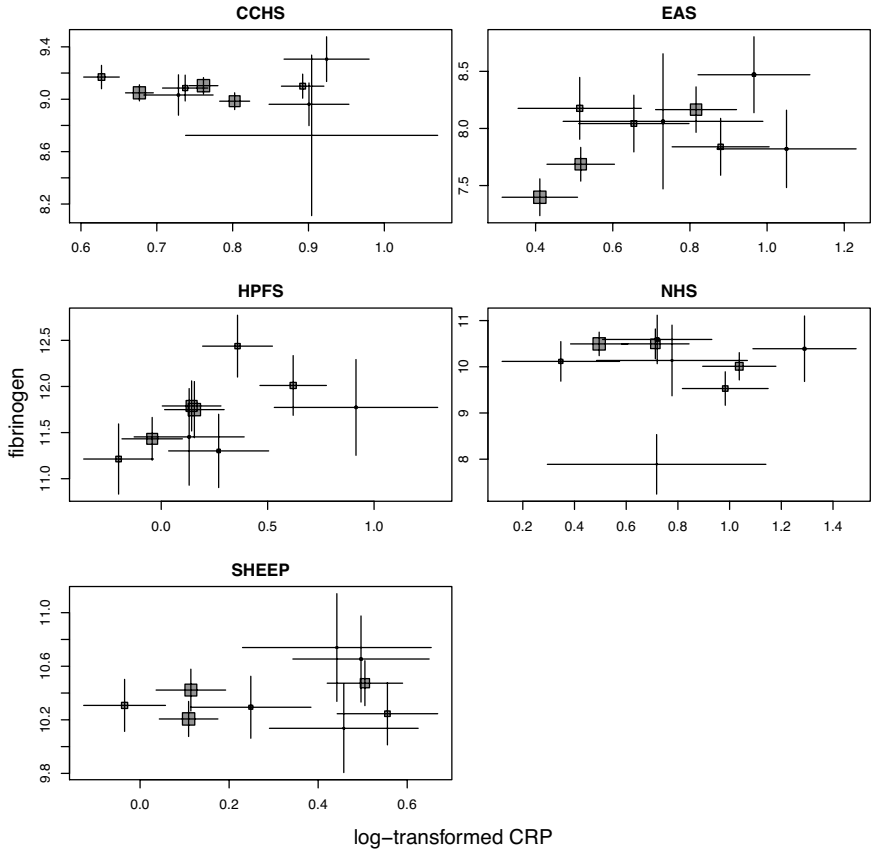


FIGURE 9.1
Summary plot of mean fibrinogen (μmol/l) against mean log(CRP) (lines are 95% confidence intervals) in each genetic subgroup for five studies. Subgroups with less than 5 subjects have been omitted; the size of the shaded squares is proportional to the number of subjects in each subgroup.

Meta-analysis model	Estimate	95% CI	DIC
Study-level model (2SLS)	0.234	−0.107 to 0.575	
Study-level model (LIML)	0.182	−0.172 to 0.536	
Summary-level model	0.058	−0.394 to 0.437	
Individual-level: different genetic effects	0.108	−0.301 to 0.479	70119
Individual-level: common genetic effects	−0.123	−0.733 to 0.348	70125
Individual-level: random genetic effects	0.072	−0.352 to 0.440	70112

TABLE 9.1
Estimates of causal effect (95% confidence intervals, CI) of log(CRP) on fibrinogen (μmol/l) from meta-analysis of five studies using study-level (data on study-specific causal effects combined by inverse-variance weighting), summary-level (data on genetic subgroups), and individual-level (without and with pooling of parameters of genetic association): hierarchical models with deviance information criterion (DIC) for individual-level models.

results in a pooled estimate further from the observational estimate, although the estimate is still closer to the positive confounded association than those of the individual-level models. The point estimates in the summary-level and individual-level model move away from the confounded association towards the slight negative association estimated in the largest study of the collaboration (Table 7.4) as the pooling of the parameters of genetic association becomes more restrictive. The deviance information criterion (DIC) is a Bayesian measure of model adequacy, where a lower value indicates a better predictive fit [Spiegelhalter et al., 2002]. Out of the models considered, the model with the lowest DIC best predicts a replicate dataset which has the same structure as that currently observed. Only models with the same data structure can be compared, hence the DIC is only given for individual-level models. A difference in DIC of 5 to 10 is considered substantial. Using the DIC to assess model adequacy, the model with random genetic effects is preferred.

9.7 Binary outcomes

Often in Mendelian randomization the outcome of interest is binary. We can modify the above methods to assume a logistic-linear association between the outcome and the exposure, thus estimating an odds ratio parameter.

9.7.1 Using summary-level data

We again drop the study-level subscript m for clarity. If we have summary-level data on the genetic associations with the exposure from linear regression, and with the outcome from logistic regression, then these coefficients can be included in a model such as equation (9.9). In this case, the causal effect parameter β_1 represents a causal log odds ratio.

If we have summary-level data on genetic subgroups, a binomial distribution in the outcome model can be assumed for the number of individuals in genetic subgroup j with events n_j (that is, the number with $Y = 1$) out of the total number of individuals in the subgroup (N_j). A linear association can be assumed between the mean level of the exposure (ξ_j) and the linear predictor (η_j), which in the example below is the logit of the probability of an event in the subgroup ($\text{logit}(\pi_j)$):

$$\begin{aligned} \bar{X}_j &\sim \mathcal{N}(\xi_j, \sigma^2_{X_j}) \\ n_j &\sim \text{Binomial}(N_j, \pi_j) \\ \eta_j = \text{logit}(\pi_j) &= \beta_0 + \beta_1 \, \xi_j. \end{aligned} \tag{9.18}$$

A log-linear regression model for the genetic associations with the outcome, or a log-linear model for relating the outcome and exposure could also be considered; in this case a causal log relative risk parameter would be estimated.

9.7.2 Using individual-level data

Similarly, we can consider modelling the probability of an event (π_i) for each individual i. The outcome Y_i takes the values 0 (no event) or 1 (event):

$$\begin{aligned} X_i &\sim \mathcal{N}(\xi_i, \sigma^2_x) \\ Y_i &\sim \text{Binomial}(1, \pi_i) \\ \eta_i = \text{logit}(\pi_i) &= \beta_0 + \beta_1 \, \xi_i. \end{aligned} \tag{9.19}$$

A hierarchical model for meta-analysis can be introduced as in the continuous outcome case.

9.7.3 Combining incident and prevalent cases in a longitudinal study

In a longitudinal cohort study with a binary outcome, if individuals are not excluded from study entry at baseline due to history of disease, each participant has two windows of opportunity to have an event: one before study entry and one after. We want to include participants in such longitudinal studies up to twice in the analysis, once in the study viewed retrospectively and once prospectively. A retrospective analysis is performed by viewing the baseline data as a cross-sectional case-control 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 new incident 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. However, while we do not want to include the individual's exposure measurement twice, we want to ensure that the same odds ratio parameter is estimated in both analyses.

In the corresponding model (9.20), we consider genetic subgroup j, containing 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.

$$\begin{aligned} X_{ij} &\sim \mathcal{N}(\xi_j, \sigma^2) \text{ for } i = 1, \ldots N_{2j} \text{ non-prevalent individuals} \\ n_{1j} &\sim \text{Binomial}(N_{1j}, \pi_{1j}) \\ n_{2j} &\sim \text{Binomial}(N_{2j}, \pi_{2j}) \\ \text{logit}(\pi_{1j}) &= \eta_{1j} = \beta_{01} + \beta_1 \, \xi_j \\ \text{logit}(\pi_{2j}) &= \eta_{2j} = \beta_{02} + \beta_1 \, \xi_j \end{aligned} \tag{9.20}$$

This model ensures that the same fitted values of the exposure are used in both logistic regressions without including individuals twice in the regression of the exposure on the genetic variants. The causal log odds ratio parameter is β_1, which is assumed to be the same in the retrospective and prospective analyses.

9.8 Discussion

In this chapter, we have presented a flexible set of models for meta-analysis of multiple studies in a hierarchical framework. This allows for the efficient synthesis of summary-level and/or individual-level data from different sources. Although study-level causal estimates can be combined in a conventional inverse-variance weighted meta-analysis, such an analysis has a number of technical deficiencies, and does not allow for the inclusion of extra information from studies where a causal estimate cannot be obtained. More detailed data present a number of advantages to the researcher, including the incorporation of information from studies where the exposure or outcome is not measured, and the efficient estimation of the genetic model of association where the same genetic variants are measured in multiple studies.

An advantage of the hierarchical structure is that the whole meta-analysis can be performed in one step. This keeps each study distinct within the hierarchical model, only combining information from 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 gene–exposure and gene–outcome associations [Thompson et al., 2005].

9.8.1 Precision of the causal estimate

To obtain a precise estimate of the causal effect, one needs to have precise estimates of both the gene–exposure and gene–outcome associations. A precise estimate of the gene–exposure association comes from a study with many participants, such as baseline data in a cohort study. For a binary outcome, a precise estimate of the gene–outcome association comes from a study with many events, such as a case-control study. The proposed hierarchical methods are able to borrow strength across such studies measuring common genetic variants to provide precise estimates of the genetic associations in all studies, and therefore obtain a more precise estimate of the causal effect.

9.8.2 Two-sample Mendelian randomization

An extreme example of the above is two-sample Mendelian randomization, in which the associations between the genetic variant(s) and exposure and between the variant(s) and outcome are estimated from non-overlapping sets of individuals. Although this may simply reflect the absence of information on the exposure and outcome associations in the same participants, it is also a potentially efficient design strategy for Mendelian randomization, particularly in view of the increasing public availability of summarized data on genetic associations with risk factors and disease outcomes from large consortia.

An important assumption therefore, to ensure the validity of the analysis, is that the two sets of individuals represent samples taken from the same underlying population. If this is not the case, then inferences may be misleading, as the association of the genetic variants with the exposure may not be replicated in the set of individuals in which the association with the outcome is estimated.

A further feature of a two-sample analysis is that any bias due to weak instruments does not act in the direction of the observational confounded association, but rather in the direction of the null [Inoue and Solon, 2010]. This means that the use of large numbers of genetic variants should not result in misleading causal claims. If the data sources for the gene–exposure and gene–outcome association estimates are partially overlapping (subsample Mendelian randomization, see Section 8.5.2), then the direction of bias will depend on the degree of overlap. If the overlap is substantial, then bias will be similar to a one-sample Mendelian randomization analysis, in the direction of the observational association. If the overlap is not substantial, then bias will be similar to a two-sample Mendelian randomization analysis, in the direction of the null.

9.8.3 Relevance to epidemiological practice

Evidence synthesis is particularly necessary in Mendelian randomization to obtain sufficiently precise estimates of causal effects to be clinically relevant.

Hierarchical models can be used to combine evidence from multiple sources in an efficient way. An example of such an analysis using a Bayesian model with vague priors is given in Chapter 10 for the pooled association of C-reactive protein on coronary heart disease risk.

9.9 Key points from chapter

- A pooled causal effect estimate can be obtained by combining study-level, summary-level or individual-level data.
- A single causal effect can be estimated from published data on genetic associations with the exposure and with the outcome, either taken from a single study or from separate sources.
- If the same genetic variants have been measured in several studies, the parameters of genetic association can be pooled in a hierarchical model across studies.
- Studies with common genetic variants can contribute to a pooled causal effect estimate even if data on one of the exposure or the outcome has not been measured.

10

Example: The CRP CHD Genetics Collaboration

Much of this book has been motivated and illustrated by data collected by the CRP CHD Genetics Collaboration [CCGC, 2008]. In this chapter, we analyse the entirety of the CCGC data to estimate the causal effect of C-reactive protein (CRP) on coronary heart disease (CHD) risk as an illustration of the Mendelian randomization approach, as well as several of the methodological issues highlighted in this book. We first give an overview of the complete dataset and address the validity of the genetic variants as instrumental variables (IVs) (Section 10.1). We then analyse a single study, exemplifying some features of the data (Section 10.2), before continuing to present an analysis of the full dataset (Section 10.3). We conclude this chapter with a discussion, including the interpretation of the results of this analysis (Section 10.4). A more detailed analysis of these data is available in a published paper [Burgess et al., 2012].

10.1 Overview of the dataset

The CCGC collated data from 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 commonly measured as a marker of systemic inflammation. As discussed in Section 1.3, it is known that CRP is observationally associated with CHD, but it is not established whether this association is causal. Studies from the collaboration measured CRP levels, genetic variants relating to CRP, and CHD events. We restrict attention to participants of European descent, excluding the four studies with no European descent participants from the analysis. This is to ensure greater homogeneity of the genetic associations in the different study populations and to mitigate potential violations of the IV assumptions due to population stratification. A list of study abbreviations for the 43 studies with European-descent participants in the CCGC is provided in Table 10.1.

Table 10.2 lists the major statistical features of the studies of the CCGC.

AGES	The Reykjavik Study of Healthy Aging for the New Millennium
ARIC	Atherosclerosis Risk in Communities Study
BHF-FHS	British Heart Foundation Family Heart Study
BRHS	British Regional Heart Study
BWHHS	British Women's Heart and Health Study
CAPS	Caerphilly Study
CCHS	Copenhagen City Heart Study
CGPS	Copenhagen General Population Study
CHAOS	Cambridge Heart Antioxidant Study
CHS	Cardiovascular Health Study
CIHDS	Copenhagen Ischaemic Heart Disease Study
CUDAS	Carotid Ultrasound Disease Assessment Study
CUPID	Carotid Ultrasound in Patients with Ischaemic Heart Disease
DDDD	Die Deutsche Diabetes Dialyse (4D) Trial
EAS	Edinburgh Artery Study
ELSA	English Longitudinal Aging Study
EPICNL	European Prospective Investigation in Cancer and Nutrition, Netherlands Centre
EPICNOR	European Prospective Investigation in Cancer and Nutrition, Norfolk Centre
FRAMOFF	Framingham Offspring Study
GISSI	Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico
HEALTHABC	Health Aging and Body Composition Study
HIFMECH	The Hypercoagulability and Impaired Fibrinolytic Function Mechanisms Study
HIMS	Health in Men Study
HPFS	Health Professionals Follow Up Study
HVHS	Heart and Vascular Health Study
INTHEART	INTERHEART Study
ISIS	International Study of Infarct Survival
LURIC	The Ludwigshafen Risk and Cardiovascular Health Study
MALMO	Malmo Diet and Cancer Study
MONICA/ KORA	Monitoring of Trends and Determinants in Cardiovascular Disease/ Cooperative Health Research in the Region of Augsburg Study
NHS	Nurses Health Study
NPHSII	Northwick Park Heart Study II
NSC	Northern Swedish Cohort Study
PENNCATH	University of Pennsylvania Catheterization Study
PROCARDIS	Precocious Coronary Artery Disease Study
PROSPER	Prospective Study of Pravastatin in the Elderly at Risk
ROTT	Rotterdam Study
SHEEP	Stockholm Heart Epidemiology Program
SPEED	Speedwell Study
UCP	Utrecht Cardiovascular Pharmacogenetics Study
WHIOS	Women's Health Initiative Observational Study
WHITE2	Whitehall II Study
WOSCOPS	West of Scotland Coronary Prevention Study

TABLE 10.1
Abbreviations for the 43 studies with subjects of European descent in the CCGC.

Further details on the individual studies can be found in the main published paper from the collaboration [CCGC, 2011]. We discuss below issues relating to the study design for studies in the collaboration, as well as relevant details for the analysis about the exposure, genetic variants, outcome, and various covariates.

10.1.1 Study design

The collaboration includes prospective studies: cohort studies, and nested case-control studies (both matched and unmatched); and retrospective studies: case-control studies (unmatched). In some prospective studies, CRP measurements were not made 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 have prevalent events in the baseline-transformed study. Four of the studies in the collaboration did not provide individual-level but only summary-level data on the numbers of individuals with and without CHD events in each genetic subgroup.

10.1.2 Exposure data: C-reactive protein

The exposure CRP was measured in each study using a high-sensitivity assay. Some of the studies did not measure CRP for all individuals, and others did not measure it for any individuals. In retrospective case-control studies, CRP measurements for cases were excluded from the analysis, as they were measured after the CHD event, to prevent bias in the causal effect due to reverse causation. In nested (prospective) case-control studies, blood was drawn and stored at baseline, to enable pre-CHD event measurement of CRP. However, in both nested and retrospective case-control studies, oversampling of cases into the study population compared to the general population biases the associations between the genetic variant(s) and the exposure. Hence analysis of CRP measurements is restricted to the controls, who form a more representative sample of the population as a whole [Bowden and Vansteelandt, 2011]. 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 the analysis of the exposure. Table 10.2 lists the number of individuals in each study with a CRP measurement suitable for use in the IV analysis according to the criteria above. As CRP has a skewed distribution, log-transformed CRP is used as the exposure.

10.1.3 Genetic data

The 43 studies in the collaboration with European descent participants measured different genetic information in the form of single nucleotide

Study	Study type	Total participants	Number of subjects with: Incident CHD	Prevalent CHD	CRP data [2]	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	10 259	680	241	9503	✓	✓		✓
CGPS	Cohort with prevalent cases	32 038	188	899	30 491	✓	✓		✓
CHS	Cohort with prevalent cases	4511	793	447	4051	✓	P	✓	
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	✓	P	✓	
ROTT	Cohort with prevalent cases	5406	259	614	4524	✓	✓		P
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	✓	✓	✓	P
NHS	Nested matched case-control	684	196		387	✓	✓	✓	P
NSC	Nested matched case-control	1673	577		969	✓	✓	✓	✓
CAPS	Nested unmatched case-control	1157	198		783		✓	✓	✓
DDDD	Nested unmatched case-control	897	269		614	✓	✓	✓	P
EPICNL	Nested unmatched case-control	3478	426		3215	✓	✓	✓	P
WHIOS	Nested unmatched case-control	3756	1339		1725	✓	✓	✓	P
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	2261		632	859	✓	P		P
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 10.1.3)			
LURIC	Unmatched case-control	2747		1137	1599	✓	✓	✓	P
PROCARDIS	Unmatched case-control	6464		3126	3302	✓	✓	✓	P
SHEEP	Unmatched case-control	2671		1113	1083	✓	✓		✓
WHITE2	Unmatched case-control	5515		31	4800	✓	✓		✓
CIHDS	Unmatched case-control (CRP in controls only)	6716		2236	4415	✓	✓		✓
BHF-FHS	Unmatched case-control (no CRP data)	4548		2146	0	✓	✓	✓	P
CHAOS	Unmatched case-control (no CRP data)	2475		623	0	✓	✓		✓
GISSI	Unmatched case-control (no CRP data)	4034		3054	0	✓	✓	✓	✓
HVHS	Unmatched case-control (no CRP data)	4407		1040	0	✓	P	✓	✓
INTHEART	Unmatched case-control (no CRP data)	4188		1883	0	✓	✓	✓	✓
UCP	Unmatched case-control (no CRP data)	2011		922	0	✓	✓	✓	P
AGES	Tabular data	3219		800	0	✓	✓	✓	✓
HEALTHABC	Tabular data	1660		584	0	✓	✓	✓	✓
MONICA/KORA	Tabular data	1675		272	0	✓	✓	✓	✓
PENNCATH	Tabular data	1509		1022	0	✓	✓	✓	✓
Total		162 416	8392	28 089	103 039				

TABLE 10.2
Summary of studies in the CRP CHD Genetics Collaboration with subjects of European descent.

[1] g1 = rs1205, g2 = rs1130864, g3 = rs1800947, g4 = rs3093077 or equivalent proxies (P indicates use of a proxy).
[2] In case-control studies, CRP data was taken in controls only; in prospective cohort studies, in subjects without prevalent CHD.

polymorphisms (SNPs) in the *CRP* gene region. Only SNPs located in this region were considered as potential IVs to ensure maximal plausibility of the IV assumptions. The region is on chromosome 1 and is responsible for the production of CRP and its regulation. The number of relevant 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 as the instrumental variables to be used in the analysis: rs1205, rs1130864, rs1800947, and rs3093077 [CCGC, 2008]. These four SNPs show varying degrees of correlation and give rise to five haplotypes which comprise at least 99% of the genetic variation exhibited in European descent populations. Indeed, over 99% of individuals in the CCGC had a genotype which was compatible with these haplotypes. Only 11 studies measured all four of the pre-specified SNPs. Some studies measured SNPs which are in complete linkage disequilibrium (LD) with one of the pre-specified SNPs ($r^2 > 0.97$ in European populations in the HapMap database), and which are used as proxies for these SNPs. 20 measured all four SNPs or proxies thereof and an additional 17 measured some three out of these four. Five of the remaining studies measured fewer than this, and the final study ISIS measured no SNPs which correspond to any of these four (in ISIS a single SNP rs2808628, also in the *CRP* gene region, was used as an IV).

Proxy SNPs are treated 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. Overall minor allele frequencies were 0.34 for g1, 0.30 for g2, 0.06 for g3, and 0.06 for g4.

There was some sporadic missingness in the genetic data in most of the studies, although this was rarely greater than 10% per SNP and usually much less. Table 10.2 lists the pre-specified SNPs measured in each study. We found that an additive per allele model of association was the most appropriate, with similar coefficients for the per allele increase in the exposure in each study [Burgess et al., 2012].

10.1.4 Outcome data: coronary heart disease

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 as it could not be separated from other CHD events in the data available. Only the first CHD event was included, so an individual could not contribute more than one event to the analysis. 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.

10.1.5 Covariate data

Data on various covariates were measured in 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), low-density lipoprotein cholesterol (LDL-C), triglycerides, apolipoprotein A1 (apo A1), and apolipoprotein B (apo B); and inflammation markers, such as fibrinogen and interleukin-6. Figure 10.1 summarizes the pooled associations of the four SNPs with CRP levels and with a wide range of 21 covariates from meta-analyses across all studies in the collaboration reporting measurements on each SNP and covariate in turn. The associations represent the standard deviation change in the covariate per allele change in the SNP. These show strong associations for CRP ($p < 10^{-30}$ for each of the four SNPs), but no more significant associations with any other covariates than would be expected by chance. 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 $p < 0.05$. We conclude that there is no indication of violation of the IV assumptions due to pleiotropic associations with measured covariates for any of the SNPs.

10.1.6 Validity of the SNPs used as IVs

Although conclusive proof is never possible, there is strong evidence for the validity of the SNPs as IVs [CCGC, 2011]. First, the SNPs are taken from the *CRP* gene region. Scientific knowledge about this genetic region gives strong plausibility to the specific association of the SNPs with CRP. Secondly, the genes from the *CRP* gene region are not known to be in linkage disequilibrium (LD) with functional variants in genes outside this region. Thirdly, the empirical associations of the SNPs with a range of potential confounders are no stronger than would be expected by chance. These potential confounders comprise the major known predictors of CHD risk. Fourthly, the genetic associations with CRP are consistent (up to chance variation) across studies. Fifthly, the frequencies of genetic variants are consistent (up to chance variation) across studies. The last two observations support the homogeneity of the European descent populations, and hence that combining estimates across different populations is reasonable and meaningful. They are also consistent with the associations of the genetic variants being due to the effects of the variants themselves, and not due to the distributions of confounders, which may differ in each population.

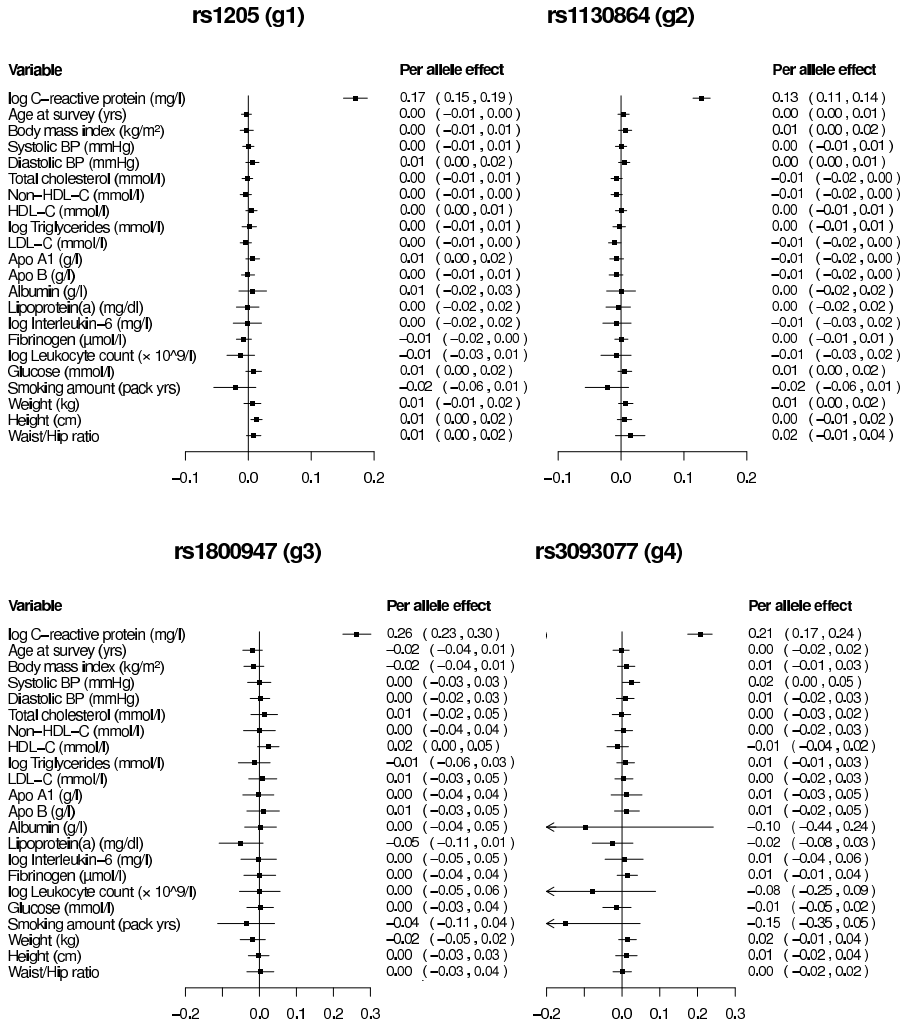


FIGURE 10.1

Pooled estimates of standard deviation change in covariate per CRP-increasing allele change in SNP for a range of covariates and the SNPs used for IV analysis. Estimates and 95% confidence intervals presented are based on random-effects meta-analyses of study-specific associations.

10.2 Single study: Cardiovascular Health Study

We first analyse a prospective cohort study, the Cardiovascular Health Study (CHS) [Fried et al., 1991], in detail as a worked example before considering the other studies. As some individuals entered the study having suffered a previous CHD event, we analyse the study in two ways for illustrative purposes: retrospectively as a case-control study, where cases are those with a prevalent CHD event and controls are all other individuals; and prospectively as a longitudinal study, excluding those with prevalent events and including only healthy individuals at baseline, where cases are those with an incident CHD event. In all analyses, we use a logistic model of association so that an odds ratio parameter is estimated in prospective and retrospective analyses; limitations of this approach in the prospective setting are discussed in Section 10.4.2.

10.2.1 Results

We analyse the data separately retrospectively and prospectively using some of the methods of Chapter 4: two-stage, Bayesian, generalized method of moments (GMM), and structural mean model (SMM) methods. Results are given in Table 10.3 using each SNP individually as an IV (analyses using an additional SNP rs2808630, labelled g5, are also presented here, although this SNP is not used in the overall meta-analysis). We see that the results from different methods are similar throughout, with differences between estimates small compared to their uncertainty. CHS suggests a significantly positive causal effect of CRP in some of the prospective analyses; this is not representative of the totality of the data (Section 10.3).

10.2.2 Posterior distributions from Bayesian methods

To illustrate the Bayesian method, the prior (normal with mean zero and variance 1000^2) and posterior distributions of the causal effect (β_1) for the retrospective logistic analyses using SNPs g1, g2 and g3 separately as IVs are shown in Figure 10.2, and the distributions using g5 in Figure 10.3. We see that the posterior distributions using g1, g2 and g3 are very different to the prior distribution, but that in the case of g5, much of the information in the posterior distribution comes from the prior. Variant g5 is only weakly associated with CRP in the CHS dataset (F statistic = 0.1, $p = 0.70$). Indeed, due to the weakness of the variant, convergence in the Monte Carlo Markov chain (MCMC) algorithm for g5 was not achieved even after a million iterations, as can be seen by the heavy tails of the posterior distribution. Convergence was assessed by sampling from multiple chains using different starting values in the MCMC algorithm, and examining the Gelman–Rubin plots to

Prospective analyses ($N = 4064$, $n = 793$)				
SNP used as IV	Two-stage	Bayesian	GMM	SMM
rs1205 (g1)	0.758 (0.295)	0.784 (0.320)	0.844 (0.438)	0.773 (0.319)
rs1417938 (g2)	0.671 (0.475)	0.728 (0.559)	0.721 (0.625)	0.680 (0.494)
rs1800947 (g3)	0.723 (0.556)	0.830 (0.704)	0.834 (0.894)	0.726 (0.579)
rs2808630 (g5) [1]	1.889 (6.546)			
all	0.725 (0.252)	0.717 (0.264)	0.791 (0.355)	0.737 (0.272)
Retrospective analyses ($N = 4511$, $n = 447$)				
rs1205 (g1)	0.388 (0.366)	0.408 (0.382)	0.388 (0.388)	0.388 (0.388)
rs1417938 (g2)	−0.527 (0.671)	−0.531 (0.696)	−0.553 (0.766)	−0.506 (0.687)
rs1800947 (g3)	0.627 (0.620)	0.864 (0.893)	0.666 (0.806)	0.634 (0.669)
rs2808630 (g5) [1]	3.521 (2.614)			
all	0.352 (0.322)	0.309 (0.326)	0.314 (0.329)	0.342 (0.330)

TABLE 10.3
Causal log odds ratios (standard errors) of CHD per unit increase in log(CRP) in prospective and retrospective analyses of the Cardiovascular Health Study (N = sample size, n = number of events) using two-stage, Bayesian, generalized method of moments (GMM), and structural mean model (SMM) methods.

[1] In the Bayesian, GMM, and SMM analyses, the estimates using g5 as an IV failed to converge.

compare between- and within-chain variance. Similarly, a single causal estimate was not obtained in the semi-parametric (GMM and SMM) approaches. This corresponds to the example of Figure 4.3, for which the confidence interval from Fieller's theorem would be unbounded. The two-stage estimates using g5 should therefore be viewed with suspicion, as the data using g5 as an IV appear to give little information on a causal effect.

10.3 Meta-analysis of all studies

Having discussed causal estimation in a single dataset, we proceed to consider the causal estimate based on the whole CCGC dataset, applying the meta-analysis methods of Chapter 9. First, we look at estimation of the causal effect using a single SNP as the IV; then we present results using all the pre-specified SNPs from study-level meta-analyses of two-stage estimates and from individual-level meta-analyses using Bayesian hierarchical models.

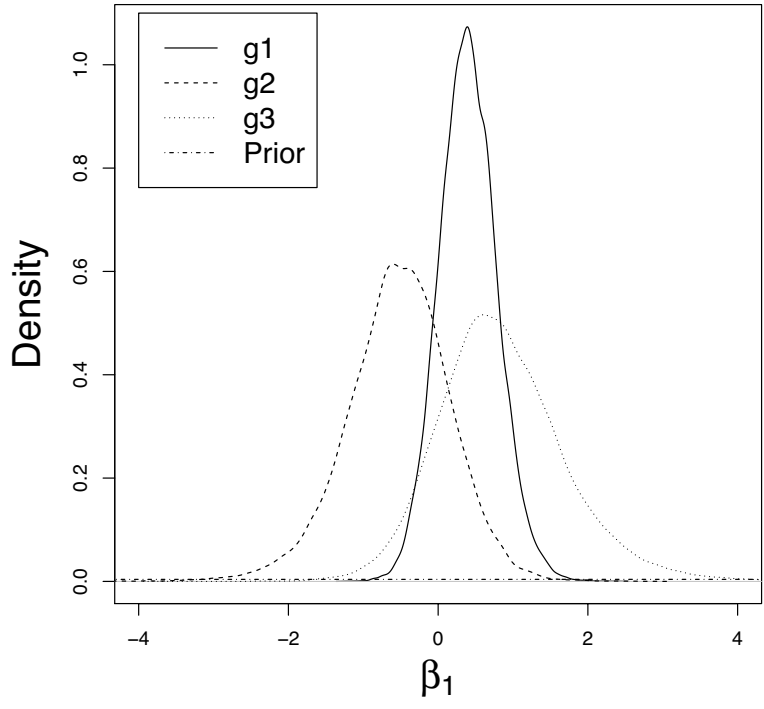


FIGURE 10.2
Prior and posterior distributions of causal log odds ratio parameter (β_1) for retrospective logistic IV analyses of the Cardiovascular Health Study using SNPs rs1205 (g1), rs1417938 (g2) and rs1800947 (g3). On this horizontal scale, the prior appears as a flat line at close to zero density.

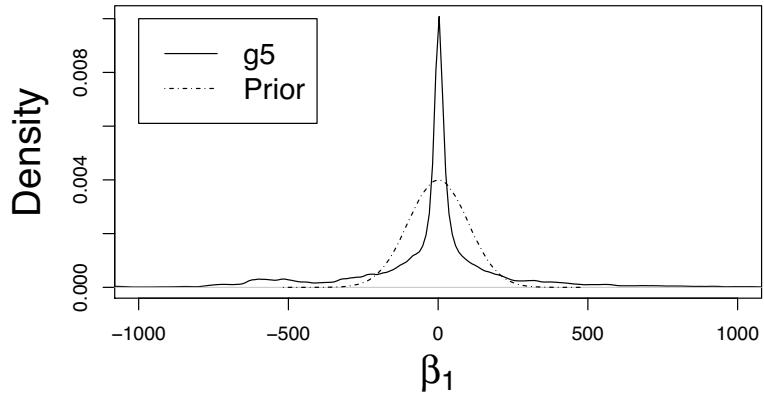


FIGURE 10.3
Prior and posterior distributions of causal log odds ratio parameter (β_1) for retrospective logistic IV analysis of the Cardiovascular Health Study using SNP rs2808630 (g5).

10.3.1 Using SNPs one at a time

We calculate causal estimates using each SNP in turn as the sole IV (Table 10.4). Pooled estimates for the G–X and G–Y associations (beta-coefficients and standard errors) are obtained from inverse-variance weighted meta-analyses using a moment estimate for the heterogeneity parameter. The causal X–Y effect estimates (odds ratios and 95% confidence intervals) are obtained from the summary-level study-specific G–X and G–Y association estimates by IV analyses in a Bayesian analysis framework using the hierarchical model of equation (9.11) and allowing for heterogeneity in the genetic association parameters using random-effects models. The correlation parameter ρ is taken as 0 and the point estimate is the mean of the posterior distribution.

The causal estimates using each SNP are similar and all compatible with a null effect; heterogeneity in the causal effect estimates would be potential evidence against the validity of one or more of the genetic variants as IVs. As the genetic variants are correlated (in linkage disequilibrium), the causal estimates are correlated, and so cannot be naively combined in a meta-analysis without considering the individual- or summary-level data. As none of these analyses uses the totality of the genetic data, an integrated approach is preferred including all of the SNPs in a single analysis.

	SNP	Number of studies	Pooled effect (SE)	p-value	Heterogeneity (I^2 and 95% CI)
G–X	g1	29	0.170 (0.010)	2×10^{-78}	58% (37–72%)
	g2	32	0.128 (0.007)	1×10^{-75}	29% (0–54%)
	g3	17	0.263 (0.019)	7×10^{-43}	14% (0–51%)
	g4	24	0.198 (0.012)	3×10^{-57}	8% (0–41%)
G–Y	g1	39	0.014 (0.013)	0.29	31% (0–54%)
	g2	42	0.001 (0.010)	0.91	2% (0–37%)
	g3	26	0.004 (0.024)	0.86	0% (0–41%)
	g4	34	−0.003 (0.023)	0.90	4% (0–32%)
	SNP	Number of studies	Causal estimate (95% CI)		
X–Y	g1	39	1.08 (0.93, 1.26)		
	g2	42	1.00 (0.84, 1.19)		
	g3	26	1.02 (0.83, 1.24)		
	g4	34	0.99 (0.78, 1.25)		

TABLE 10.4
Pooled estimates from univariate 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, and heterogeneity (I^2 represents the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error [Higgins et al., 2003]); causal estimates (X–Y association) for odds ratio of CHD per unit increase in log(CRP) from meta-analysis using each SNP as the sole IV from the method of equation (9.11).

10.3.2 Using all SNPs

We perform meta-analyses of both study-level and individual-level data. In the study-level data meta-analysis, we undertake a two-stage analysis in each study with information on genetic variants, the exposure and the outcome using all the pre-specified SNPs measured in that study as IVs. In cohort studies, two separate estimates are calculated using the two-stage method with prevalent and with incident events; the two estimates are then combined for each study using an inverse-variance weighted fixed-effect meta-analysis to give a study-specific effect estimate. The study-level estimates are combined in an inverse-variance weighted random-effects meta-analysis.

In the individual-level meta-analyses, a Bayesian hierarchical model with vague priors (uniform on the interval $[0, 10]$ for standard deviation and heterogeneity parameters, normal with mean zero and variance 1000^2 for all other parameters) is used as described in Section 9.5 (in particular, equations 9.15 and 9.12). Analyses are presented based on the same data as the study-level meta-analysis, as well as on the totality of the data. By pooling the parameters

of genetic association (Section 9.5.3), an additional 10 studies and over 10 000 extra CHD cases were able to be included in the analysis. These additional studies either did not measure CRP levels, or only provided summary-level data. Studies were divided into four groups based on the SNPs measured in that study, and the parameters of genetic association were pooled across studies within these groups. Prospective and retrospective analyses of cohort studies were combined as described in Section 9.7.3. Heterogeneity was acknowledged by the use of random-effects models in both the genetic association and causal effect parameters.

Table 10.5 shows the pooled estimates of association. We see that the causal effect in each analysis is close to the null. When the same data are used, the point estimates in the two-stage and Bayesian methods are very similar and the 95% confidence/credible intervals (CIs) are of similar width, with the Bayesian interval slightly wider. The analysis with pooled genetic association parameters based on the same data here gave a slight reduction in precision of the causal estimate because of an increase in the between-study heterogeneity, but for the analysis based on all of the data, the precision of the causal effect estimate increased.

These analyses rule out even a small causal effect of long-term CRP levels on CHD risk. The upper bound of the 95% CI in the final analysis using the totality of the data available corresponds to an odds ratio of 1.10 for a unit increase in log(CRP), which is close to a 1 standard deviation increase in log(CRP).

Method used	Studies	Events	Causal estimate	$\hat{\tau}$
Study-level meta-analysis of two-stage estimates	33	24 135	1.02 (0.91 to 1.15)	0.121
Individual-level Bayesian meta-analysis without pooling	33	24 135	1.02 (0.89 to 1.16)	0.132
Individual-level Bayesian meta-analysis with pooling (same data)	33	24 135	1.01 (0.87 to 1.16)	0.153
Individual-level Bayesian meta-analysis with pooling (all data)	43	36 463	0.99 (0.89 to 1.10)	0.106

TABLE 10.5

Causal estimates of odds ratio of CHD per unit increase in log(CRP) using all available pre-specified SNPs as IVs in random-effects meta-analyses: number of studies and CHD events included in analysis, estimate of causal effect (95% confidence/credible interval), heterogeneity estimate ($\hat{\tau}$, the between-study standard deviation of the causal log odds ratios); pooling refers to pooling of the genetic associations with log(CRP) across studies.

10.4 Discussion

This chapter has illustrated methods for the synthesis of Mendelian randomization data comprising a variety of study designs and measuring a variety of genetic variants. Studies with differing design can be analysed separately and then combined in a study-level meta-analysis, or alternatively analysed together in an individual-level meta-analysis using a hierarchical model.

10.4.1 Precision of the causal estimate

The individual-level hierarchical method is able to include an additional 10 studies and 50% more events compared to a study-level meta-analysis. A more precise estimate of the causal effect is obtained. This is illustrated by the width of the 95% CI of the causal parameter on the log odds ratio scale reducing from 0.306, 0.343, 0.401 and 0.468 using a single SNP as the IV (Table 10.4), or 0.232 and 0.260 using the two-stage or hierarchical methods with data on 33 studies (Table 10.5), down to 0.209 in the final hierarchical method with data on 43 studies (Table 10.5) due to the borrowing of information across studies and inclusion of studies without measured exposure levels. The use of the final method represents more than a 110% gain in efficiency compared to the single SNP analyses of Table 10.4, and more than a 20% gain compared to the two-stage estimate in this example.

10.4.2 Limitations of this analysis

The main limitation of the methods used is the reliance on parametric assumptions and explicit specification of distributions, such as normality of the exposure, homogeneity of its variance across genetic subgroups, and additive per allele models of the IV–exposure association. While there is no evidence against these assumptions in this example, sensitivity analyses can be used to quantify the potential impact of violation of these assumptions. For example, with the CHS study, the semi-parametric approaches (GMM and SMM) gave similar estimates to the fully-parametric methods.

One particular assumption was that all of the studies could be analysed using a logistic model of association, although sensitivity analyses have been performed using alternative models (such as a log-linear model and a Cox model for a survival outcome, see Section 11.1.1) [Burgess, 2012b]. Studies with different designs could be analysed using different regression models, such as conditional logistic regression for matched case-control studies, or proportional hazards regression models for prospective cohort studies. If this were done, an additional assumption would have to be made in the meta-analysis, that estimates of somewhat different parameters from studies of different designs can be combined in a single meta-analysis model.

10.4.3 Assessing the IV assumptions

The combination of individual-level data from multiple studies enables more detailed assessment of the IV assumptions than summary-level data or data from a single study. With individual-level data, the associations of the SNPs used as IVs with numerous measured covariates can be tested in a systematic way. With data from multiple studies, in addition to the gains in power from the increased sample size, the IV assumptions can be assessed by inspection of the homogeneity of genetic associations and haplotype frequencies across studies. The combination of statistical assessment and scientific knowledge helps to justify the validity of the genetic variants as IVs in this analysis.

10.4.4 Interpretation of the results

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 (or decrease) in CRP levels 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 level of CRP necessary for the formation of atherosclerotic plaques is so small that no practical intervention can lower CRP to a level where the CHD risk is reduced. The biochemical notion of causation is not necessarily relevant to the consequences of an intervention targeted at CRP. The interpretation of the Mendelian randomization estimate is in terms of the effect of a long-term change in usual levels of CRP on CHD risk.

While the null association from the Mendelian randomization analysis of CRP on CHD risk does not preclude a causal effect of a small magnitude, the rationale for proposing CRP as a target for clinical intervention to reduce CHD risk is diminished. The results from this analysis add to a growing body of evidence that CRP is a bystander of CHD, rather than a causal agent [Keavney, 2011]. The estimate from a Mendelian randomization analysis is not attenuated by measurement error or within-individual variation, and represents the effect of long-term exposure to elevated levels of CRP, so may be greater in magnitude than that of any potential intervention even if there were a small causal effect of CRP. However, the results do not preclude the possibility of short-term acutely elevated levels of CRP being part of the process leading to a CHD event.

10.4.5 Relevance to epidemiological practice

This analysis demonstrates the feasibility of Mendelian randomization for addressing a clinically important question, but also the great efforts required to achieve a clinically relevant estimate. The gains in efficiency of the more sophisticated analyses do not come from additional assumptions, but from the synthesis of evidence from multiple IVs and multiple studies to give a single causal estimate based on the totality of the data available. Although it may seem disappointing to go to all this effort to demonstrate a null finding, not all analyses using Mendelian randomization have given negative results (Chapter 5), and in many cases, including this one, the demonstration of no clinically relevant causal effect still has considerable scientific importance.

10.5 Key points from chapter

- The analyses presented in this chapter exemplify the assessment of the assumptions required to perform a Mendelian randomization analysis and the estimation of an overall causal effect.
- The integrated analyses presented based on the totality of available data give a precise enough causal estimate to rule out even a moderately-sized causal effect of C-reactive protein on coronary heart disease risk.

Part III

Prospects for Mendelian randomization

11

Future directions

In this final chapter, we consider the future of Mendelian randomization within the wider context of genetic epidemiology. We divide the chapter into two sections. First, we discuss methodological developments in instrumental variable techniques which enable more sophisticated Mendelian randomization analyses. Secondly, we discuss applied developments, such as advances in genotyping and other high-throughput cell biology techniques, which widen the scope for future Mendelian randomization analyses.

11.1 Methodological developments

We consider here areas in need of further methodological development, alongside recent innovations in instrumental variable (IV) methods, as well as IV methods which are established in the econometrics literature but have not yet been applied to the context of Mendelian randomization.

11.1.1 Survival data

For disease incidence in a longitudinal study, rather than the outcome being a binary indicator of the presence or absence of disease, survival data (also called time-to-event data) may be available on the length of time each individual was enrolled in the study prior to a disease event [Collett, 2003]. Typically, such data are analysed using a proportional hazards model (known as a Cox model) to investigate the relationship between covariates and disease risk, although other approaches are also available. The relevant estimate from a proportional hazards model is a (log) hazard ratio.

As with other forms of data, a causal relationship can be assessed by testing for an association between the IV and the outcome. With survival data, this may be performed in a Cox regression model of the survival outcome on the IV. Parameter estimation is more troublesome, particularly in view of the non-collapsibility of the hazard ratio (Section 4.2.3*). The precise definition of a causal hazard ratio, and so the target parameter to be estimated in an

IV analysis using survival data, is not clear. Additionally, issues of competing risks or informative censoring of follow-up times may have to be addressed.

Although *ad hoc* methods for IV estimation with survival data have been considered (such as a ratio estimate, the coefficient from the gene–outcome Cox regression model divided by the coefficient from the gene–exposure linear regression model, as in Section 5.3.3), a more principled approach should be possible. A potential approach for this is an accelerated failure-time model, as this has proved to be a good choice in other aspects of causal modelling [Robins, 1992]. A pragmatic alternative is to ignore the time-to-event component and just consider a binary outcome with a log-linear or logistic model (as in Chapter 10).

11.1.2 Non-linear exposure–outcome relationships

Although semi-parametric methods (Section 4.4*) are able to weaken the distributional assumptions made in a fully parametric IV analysis model, a parametric model is still necessary for the association between the exposure and the outcome. In many cases, such as with the association of obesity with all-cause mortality, the relationship between the exposure and the outcome is non-linear and may even not be monotone. Non-linear parametric methods have been considered for use in IV analyses, but inferences based on such methods have been shown to be highly sensitive to the choice of parameterization [Mogstad and Wiswall, 2010; Horowitz, 2011].

One reason for this is that most genetic variants have a small effect on the exposure. For example, the variant (located in the *FTO* gene region) explaining the most variation in body mass index (BMI) has an association of less than 1 kg/m^2 per additional allele [Speliotes et al., 2010], whereas BMI in most populations typically ranges from about 17 to 40 kg/m^2 across individuals. So a Mendelian randomization analysis using this genetic variant would compare subgroups which differ only slightly in their average level of BMI. Non-linearities on this reduced scale would not be apparent, and would not address the clinically relevant question of whether being underweight leads to an increased risk of death.

The estimate from a linear IV analysis, such as using the ratio method (Section 4.1), approximates a population-averaged causal effect [Angrist et al., 1996]. With a linear exposure–outcome relationship, this is the average change in the outcome for a uniform change (usually a 1 unit increase) in the distribution of the exposure across the whole population. For a non-linear relationship, a linear IV estimate approximates the same population-averaged causal effect when the change in the distribution of the exposure associated with the IV is small, and the linear IV estimate is scaled to represent the effect of a change in the exposure of similar magnitude to that associated with a change in the IV. For example, in the case of BMI, each additional copy of the variant in the *FTO* gene region in a European population was estimated to be associated with a 0.4 kg/m^2 increase in BMI [Burgess et al., 2014b]. Hence the linear IV

estimate using this variant, expressed as the causal effect on the outcome of a 0.4 kg/m^2 change in BMI, would approximate the average effect of increasing the BMI of every individual in the population by 0.4 kg/m^2.

If the exposure–outcome relationship is not monotone (for example, it is J- or U-shaped), then the true change in the outcome for a given change in the exposure may be in different directions for various members of the population; but the IV estimate is of the average change in the outcome across the population [Angrist et al., 2000]. Hence, standard IV methods can still be used to test for the presence of a causal effect even if the exposure–outcome relationship is non-linear, and the estimated parameter has a natural interpretation, but any single effect estimate will not tell the whole story of the causal relationship.

If the shape of the exposure–outcome causal relationship is of interest, local IV estimates can be obtained within strata of the exposure, such as deciles or quintiles. By plotting these estimates against the average level of the exposure in the strata, the shape of the causal relationship can be assessed graphically. However, if the exposure is stratified on directly, misleading results may be obtained. This is because the exposure lies on the causal pathway between the IV and the outcome, and so conditioning on the exposure induces an association between the IV and confounders. This can be circumvented by initially subtracting the effect of the IV on the exposure from the exposure measurement, to obtain the 'IV-free exposure'. This quantity, representing the expected value of the exposure for an individual if their IV took the value zero, can then be safely conditioned on. For the approach to be valid, it is necessary for the average genetic association with the exposure in the population to remain constant at different levels of the exposure [Burgess et al., 2014b]. Further methodological work is required to assess the robustness of this approach to violation of this assumption, as well as to stratifying directly on the exposure in situations where calculating the IV-free exposure may be problematic, such as if the exposure takes discrete values or has a natural maximum or minimum value.

11.1.3 Untangling the causal effects of related exposures

Low-density lipoprotein cholesterol (LDL-C) and high-density lipoprotein cholesterol (HDL-C) are both lipid fractions which have been shown to be observationally associated with coronary heart disease (CHD) risk. The causal effects of both lipid fractions on CHD risk have been estimated in Mendelian randomization investigations using genetic variants specifically associated with each of LDL-C and HDL-C in turn, and neither associated with the other, nor with another lipid fraction, triglycerides [Voight et al., 2012] (see Section 5.4). However, these analyses exclude the majority of variants associated with HDL-C and LDL-C, and cannot be performed for triglycerides, due to a lack of variants associated with triglycerides and not associated with either HDL-C and LDL-C.

Instrumental variable analyses can be performed for multiple exposure variables simultaneously. If variants that describe variation in the exposures of interest have pleiotropic effects, but these effects are restricted to the set of exposures under investigation, then the causal effects of each of the exposures can be estimated.

Formally, the assumptions necessary are:

i. the set of genetic variants must be associated with each of the exposures (it is not necessary for each variant to be associated with every exposure),

ii. each variant must not be associated with confounders of any exposure–outcome association, and

iii. all causal pathways from a variant to the outcome must pass through one of the exposures.

This situation is analogous to a factorial randomized controlled trial, where multiple randomized interventions are simultaneously assessed, and has been named 'multivariable Mendelian randomization' [Burgess and Thompson, 2014]. In the case of the lipid fractions above, such an analysis may have greater power to detect causal effects than analyses restricted to variants only associated with each of the individual lipid fractions in turn.

A multivariable Mendelian randomization analysis can be performed using the two-stage least squares method (Section 4.2). This is undertaken by first regressing the exposures on the genetic variants in a multivariate multiple linear regression (first stage; multiple dependent variables and multiple explanatory variables), and then by regressing the outcome linearly on the fitted values of each of the exposures in a univariate multiple regression (second stage; one dependent variable and multiple explanatory variables). Alternatively, a multivariable Mendelian randomization analysis can be performed using summarized data, as in Section 9.4.1, except considering a multivariate normal distribution for the genetic associations with each of the exposures and with the outcome. If there are causal effects between the exposures (say, of one exposure on another), then the causal effect estimates from a multivariable Mendelian randomization analysis represent direct causal effects of each exposure variable on the outcome, not including indirect effects via another exposure variable.

Inference of a causal effect in multivariable Mendelian randomization relies on the differential associations of multiple genetic variants with the exposures. Consequently, the intuitive appeal of using Mendelian randomization to infer a causal effect from a variant's sole associations with an exposure and outcome is somewhat reduced. Additionally, the assumption that the pleiotropic effects of variants can be completely characterized may be unrealistic. This approach should therefore only be considered for closely-related exposure variables, and is not a general purpose way of attempting to deal with pleiotropy.

11.1.4 Elucidating the direction of causal effect

If two distinct sets of genetic variants are available, each of which consists of valid IVs for a separate variable, then the direction of causal effect between the two variables (if any) can be judged by assessing whether each variable in turn has a causal effect on the other. For example, C-reactive protein (CRP) and BMI are observationally correlated. A genetic variant in the *CRP* gene region has been shown not to be associated with BMI, suggesting that elevated CRP is not a cause of changes in BMI levels; but a variant in the *FTO* gene region has been shown to be associated with CRP, suggesting that elevated BMI is a cause of increased CRP levels [Timpson et al., 2011]. In this way, the direction of the causal relationship between obesity and inflammation (in particular, between BMI and CRP) can be assessed [Welsh et al., 2010].

In additional to there existing no causal relationship or a unidirectional causal relationship, it is also possible for there to be a reciprocal causal relationship, where each of the variables is a cause of the other [Grassi et al., 2007]. This may occur due to variation in the causal effects of the variables across different periods of the life-span [Burgess et al., 2014a].

11.1.5 Investigating indirect and direct effects

Furthermore, if the direction of causation is known for two exposure variables (and one is the cause of the other), the direct and indirect effects of the primary exposure (in the example above, BMI) on an outcome via the secondary exposure (in the example above, CRP) can be considered. The indirect effect of BMI via CRP represents the effect of BMI on the outcome mediated by the effect of BMI on CRP. The direct effect of BMI represents the effect of BMI on the outcome via all other causal pathways, but not via CRP. If IVs for both the primary exposure and the secondary exposure (referred to as a mediator) are available, then direct and indirect effects can be calculated in the presence of unmeasured confounding on the assumption that all effects are linear without interactions [Burgess et al., 2014a].

It is also possible to consider the indirect and direct effects of a genetic variant on an outcome, with the exposure as a mediator. For example, genetic variants linked with smoking have also been shown to be associated with lung cancer, suggesting a causal effect of smoking on lung cancer risk. But the indirect effect of the variant mediated by a measure of smoking intensity, the number of cigarettes smoked per day, was close to zero and the direct effect via other pathways was similar in magnitude to the overall causal effect [VanderWeele et al., 2012]. Such a scenario indicates a violation of the IV assumptions, as a direct effect of the genetic variant on the outcome in a Mendelian randomization analysis is precluded. However, a more reasonable interpretation is that the genetic association with the outcome is mediated via a different pathway through another measure of smoking behaviour, such as via the amount of nicotine extracted from each cigarette [Le Marchand et al.,

2008]. More generally, mediation analysis can suggest the pathway by which a genetic variant is associated with the outcome, and hence be informative about causal mechanisms linking an exposure measure to the outcome.

There has been considerable recent research on mediation analysis, including technical definitions of direct and indirect effects [Pearl, 2001], and investigations into the assumptions necessary for valid estimation of these effects, in particular relating to unmeasured confounding [VanderWeele and Vansteelandt, 2009]. When the genetic variant can be assumed to be randomly assigned, as in Mendelian randomization for a valid genetic instrumental variable, the "no unmeasured confounding" assumptions relating to the associations between the genetic variant and the exposure and between the genetic variant and the outcome are automatically satisfied; however additional assumptions such as no unmeasured confounding between the mediator and exposure and no post-treatment confounding are still required [Emsley et al., 2010].

11.2 Applied developments

In this final section, we consider advances in genetic epidemiology leading to emerging directions for applied Mendelian randomization analyses.

11.2.1 High-throughput cell biology: -omics data

The term "-omics" covers a broad range of fields of study in cell biology and beyond resulting from developments in high-throughput analytical techniques. Examples of -omics data include gene expression data (genomics), methylation data (epigenomics), protein data (proteomics), transcription data (transcriptomics), and metabolites (metabolomics/metabonomics) [Relton and Davey Smith, 2012a]. Integration of multiple types of -omics data may give insight into the relations between basic biological biomarkers. Examples of such approaches have been named 'genetical genomics' (integration of genetic variants and gene expression data) [Jansen and Nap, 2001] and 'genetical epigenomics' (integration of genetic variants and epigenetic data) [Relton and Davey Smith, 2010]. A practical application of the integration of -omics data with phenotypic and disease data is an investigation into associations between cigarette smoking behaviours and disease outcomes with DNA methylation to search for mechanisms by which an increased risk of smoking-related diseases may persist even after cessation of smoking [Wan et al., 2012].

Relationships between epigenetic markers, proteins, transcription factors and metabolites can be affected by confounding and reverse causation in the same way as relationships between phenotypic exposures and outcomes. Although the causal network is generally high-dimensional and unknown, the

direction of potential causal relationships between types of -omics data can often be deduced from external biological knowledge (for example, from a genetic variant to gene expression to a protein). A similar analytical approach to the investigation of indirect and direct effects in Mendelian randomization has been proposed under the name 'two-step epigenetic Mendelian randomization' using separate genetic variants as instrumental variables for a phenotype (exposure) and an epigenetic marker (mediator), to investigate mediation of the causal effect of the exposure on the outcome [Relton and Davey Smith, 2012b].

A key difficulty here is finding separate genetic variants specifically associated with the phenotype and with the epigenetic marker if the two variables are closely biologically related. Additionally, obtaining relevant data may be problematic, as several variables (such as methylation and transcription data) are tissue-specific, and so must be measured in a specific cell type. However, as technologies for measuring -omics data improve, Mendelian randomization will be an important tool for understanding biological pathways, particularly as genetic variants are closer biologically to these cellular variables than they are to phenotypes such as BMI or blood pressure, and so genetic associations may be stronger.

11.2.2 Mendelian randomization with GWAS data

A genome-wide association study (GWAS) is a hypothesis-free examination of the whole genome of individuals in a study population to discover genetic variants associated with a particular trait. Such studies present difficulties due to the sheer number of genetic variants analysed and the corresponding number of association tests. Stringent levels for p-values, such as $p < 5 \times 10^{-8}$, have been used as a threshold for statistical significance to control the number of false positive findings. Such a stringent p-value means that the power to detect relevant variants may be low. However, GWAS investigations have been successful, discovering hundreds of genetic variants associated with exposures and disease outcomes [Manolio, 2010], and providing evidence regarding novel causal pathways and risk factors [Klein et al., 2005].

A GWAS can be used as a source of genetic variants for a Mendelian randomization analysis. However, when the function of these genetic variants is unknown, it may be that causal estimates are biased due to violations of the IV assumptions for one or more variants. Although it is possible for the associations of multiple genetic variants with the outcome all to be biased due to pleiotropy, if several genetic variants associated with an exposure are all concordantly associated with the outcome, then it is more implausible for all to be due to pleiotropy, particularly if there is a dose–response relationship in the genetic associations with the exposure and outcome (Section 3.2.6). This would increase confidence in the conclusion that the exposure is a cause of the outcome, or at least is a proxy measure of such a cause.

Although a hypothesis-free (agnostic) approach, in which the function of variants in an analysis is unknown, may give an indication of whether an exposure is a causal risk factor, neither positive nor negative results should be over-interpreted. This is particularly relevant when large numbers of genetic variants are included in the analysis. Analyses using variants from the whole genome of individuals regardless of the strength of association of the variant or its function are common in estimating the heritability of traits [Yang et al., 2011] and in risk prediction [Dudbridge, 2013], but have been shown to give misleading findings in Mendelian randomization. For example, such analyses have suggested that CRP is a causal risk factor for CHD risk, but BMI is not [Evans et al., 2013]. There is no strong justification for using large numbers of variants from genome-wide data for Mendelian randomization investigations. Indeed, investigations have shown that the proportion of variance in exposures explained by externally-derived allele scores can decrease (rather than increase as might be expected) as the p-value threshold for including a variant in such a score becomes more liberal [Burgess et al., 2014d].

11.2.3 Whole-genome sequencing and rare variants

Advances in genotyping technology known as 'next-generation sequencing' are enabling the measurement of increasing numbers of genetic markers, up to and including the whole genome of an individual (whole-genome sequencing). The measurement of rare genetic variants provides opportunities to discover 'better' tools for Mendelian randomization: such as variants with stronger associations with the exposure, or more specific associations with the exposure if a candidate gene region is pleiotropic. However, Mendelian randomization simply relies on the genetic variant being specifically associated with the exposure of interest; it is not necessary to find the 'causal variant' to perform a valid Mendelian randomization analysis (Figure 3.2).

The use of rare genetic variants in Mendelian randomization may pose problems. Genetic variants are not truly randomized in a population, but rather passed on through Mendelian inheritance. If the variants are fairly common, then it may reasonable to assume that the variants are randomly distributed with regard to potential confounding variables, and so can be regarded as being randomized (known as quasi-randomization, see Section 2.1.4). But rare variants will be clustered in families, and hence cannot be regarded as randomly distributed in the population. So while rare genetic variants are useful for functional genomics, their use in Mendelian randomization should be viewed with some caution. Additionally, if the variant is rare, the power to detect a causal effect may be low.

11.2.4 Published data and two-sample Mendelian randomization

Two-sample Mendelian randomization (Section 9.8.2) is the use of separate datasets to estimate the gene–exposure and gene–outcome associations in a Mendelian randomization analysis. While such analyses are not altogether novel, the increasing availability of published resources of genetic associations with traits (both exposures and disease outcomes) enables the assessment of causality for an exposure with many outcomes. For example, a functional variant in the *IL6R* gene region associated with interleukin-6 receptor is associated with CHD risk [Swerdlow et al., 2012]. Two-sample Mendelian randomization can be used to see whether the variant is also associated with a number of other outcomes, and so whether the interleukin-6 pathway is causal for those outcomes. For example, the variant is also associated with psoriatic arthritis and asthma, suggesting common causal risk factors and pathways underlying the disease outcomes. The availability of genetic associations in public repositories makes these 'phenome scans' of genetic associations with multiple risk factors and disease outcomes a practical option [Burgess et al., 2014e].

11.3 Conclusion

In conclusion, there are still areas of ongoing methodological research in Mendelian randomization, and work is needed to translate existing and future methodological developments into the context of Mendelian randomization for applied researchers. This is fueled to a large extent by increasing data availability: new exposure variables, increasing detail of genetic measurements, and publicly-available data resources. These are likely to provide further insights into causal mechanisms, and further scope for methodological and applied developments in the future.

Bibliography

Allin, K., Nordestgaard, B., Zacho, J., Tybjærg-Hansen, A., and Bojesen, S. 2010. C-reactive protein and the risk of cancer: a Mendelian randomization study. *Journal of the National Cancer Institute*, 102(3):202–206. (Cited on page 11.)

Almon, R., Álvarez-Leon, E., Engfeldt, P., Serra-Majem, L., Magnuson, A., and Nilsson, T. 2010. Associations between lactase persistence and the metabolic syndrome in a cross-sectional study in the Canary Islands. *European Journal of Nutrition*, 49(3):141–146. (Cited on page 11.)

Anderson, T. and Rubin, H. 1949. Estimators of the parameters of a single equation in a complete set of stochastic equations. *Annals of Mathematical Statistics*, 21(1):570–582. (Cited on pages 54 and 108.)

Angrist, J., Graddy, K., and Imbens, G. 2000. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies*, 67(3):499–527. (Cited on pages 48, 56, and 177.)

Angrist, J., Imbens, G., and Rubin, D. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455. (Cited on pages 38, 41, and 176.)

Angrist, J. and Krueger, A. 1992. 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*, 87(418):328–336. (Cited on page 146.)

Angrist, J. and Pischke, J. 2009. *Mostly harmless econometrics: an empiricist's companion. Chapter 4: Instrumental variables in action: sometimes you get what you need*. Princeton University Press. (Cited on pages 56, 57, 58, 61, 67, and 108.)

Basmann, R. 1960. On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association*, 55(292):650–659. (Cited on page 68.)

Baum, C., Schaffer, M., and Stillman, S. 2003. Instrumental variables and GMM: Estimation and testing. *Stata Journal*, 3(1):1–31. (Cited on pages 67, 68, and 69.)

Baum, C., Schaffer, M., and Stillman, S. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal*, 7(4):465–506. (Cited on page 67.)

Bech, B., Autrup, H., Nohr, E., Henriksen, T., and Olsen, J. 2006. Stillbirth and slow metabolizers of caffeine: comparison by genotypes. *International Journal of Epidemiology*, 35(4):948–953. (Cited on page 11.)

Beer, N., Tribble, N., McCulloch, L., et al. 2009. The P446L variant in *GCKR* associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Human Molecular Genetics*, 18(21):4081–4088. (Cited on page 95.)

Bekker, P. 1994. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, 62(3):657–681. (Cited on pages 61 and 108.)

Beral, V., Banks, E., Bull D., Reeves, G. (Million Women Study Collaborators) 2003. Breast cancer and hormone-replacement therapy in the Million Women Study. *The Lancet*, 362(9382):419–427. (Cited on page 4.)

Bochud, M., Chiolero, A., Elston, R., and Paccaud, F. 2008. A cautionary note on the use of Mendelian randomization to infer causation in observational epidemiology. *International Journal of Epidemiology*, 37(2):414–416. (Cited on page 32.)

Bochud, M. and Rousson, V. 2010. Usefulness of Mendelian randomization in observational epidemiology. *International Journal of Environmental Research and Public Health*, 7(3):711–728. (Cited on page 9.)

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. 2009. *Introduction to meta-analysis. Chapter 34: Generality of the basic inverse-variance method*. Wiley. (Cited on page 140.)

Bound, J., Jaeger, D., and Baker, R. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450. (Cited on pages 99 and 104.)

Bowden, J. and Vansteelandt, S. 2011. Mendelian randomisation analysis of case-control data using structural mean models. *Statistics in Medicine*, 30(6):678–694. (Cited on pages 51 and 159.)

Brennan, P., McKay, J., Moore, L., et al. 2009. Obesity and cancer: Mendelian randomization approach utilizing the *FTO* genotype. *International Journal of Epidemiology*, 38(4):971–975. (Cited on page 89.)

Brion, M.-J., Shakhbazov, K., and Visscher, P. 2013. Calculating statistical power in Mendelian randomization studies. *International Journal of Epidemiology*, 42(5):1497–1501. (Cited on page 128.)

Browning, S. 2006. Multilocus association mapping using variable-length Markov chains. *American Journal of Human Genetics*, 78(6):903–913. (Cited on page 134.)

Browning, S. and Browning, B. 2007. 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*, 81(5):1084–1097. (Cited on page 134.)

Buonaccorsi, J. 2005. *Encyclopedia of Biostatistics*, chapter Fieller's theorem, pages 1951–1952. Wiley. (Cited on page 53.)

Burgess, S. 2012a. *Statistical issues in Mendelian randomization: use of genetic instrumental variables for assessing causal associations. Chapter 4: Collapsibility for IV analyses of binary outcomes.* PhD thesis, University of Cambridge. Available at http://www.dspace.cam.ac.uk/handle/1810/242184. (Cited on page 59.)

Burgess, S. 2012b. *Statistical issues in Mendelian randomization: use of genetic instrumental variables for assessing causal associations. Chapter 8: Meta-analysis of Mendelian randomization studies of C-reactive protein and coronary heart disease.* PhD thesis, University of Cambridge. Available at http://www.dspace.cam.ac.uk/handle/1810/242184. (Cited on page 170.)

Burgess, S. 2014. Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. *International Journal of Epidemiology*, 43(3):922–929. (Cited on page 131.)

Burgess, S., Butterworth, A., and Thompson, S. 2013. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665. (Cited on pages 90 and 142.)

Burgess, S. and CCGC (CHD CRP Genetics Collaboration) 2013. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine*, 32(27):4726–4747. (Cited on page 59.)

Burgess, S., Daniel, R., Butterworth, A., Thompson, S., and EPIC-InterAct Consortium 2014a. Network Mendelian randomization: extending instrumental variable techniques. *International Journal of Epidemiology*, available online before print. (Cited on page 179.)

Burgess, S., Davies, N., Thompson, S., and EPIC-InterAct Consortium 2014b. Instrumental variable analysis with a non-linear exposure–outcome relationship. *Epidemiology*, available online before print. (Cited on pages 47, 176, and 177.)

Burgess, S., Granell, R., Palmer, T., Didelez, V., and Sterne, J. 2014c. Lack of identification in semi-parametric instrumental variable models with binary

outcomes. *American Journal of Epidemiology*, 180(1):111–119. (Cited on pages 66, 68, and 109.)

Burgess, S., Howson, J., Surendran, P., Thompson, S., and EPIC-CVD Consortium 2014d. Mendelian randomization in the post-GWAS era: the perils of causal inference without biological knowledge. Submitted for publication. (Cited on page 182.)

Burgess, S., Scott, R., Timpson, N., Davey Smith, G., Thompson, S., and EPIC-InterAct Consortium 2014e. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology*, available online before print. (Cited on pages 142 and 183.)

Burgess, S., Seaman, S., Lawlor, D., Casas, J., and Thompson, S. 2011a. Missing data methods in Mendelian randomization studies with multiple instruments. *American Journal of Epidemiology*, 174(9):1069–1076. (Cited on page 134.)

Burgess, S. and Thompson, S. 2011. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine*, 30(11):1312–1323. (Cited on pages 101, 108, 109, and 114.)

Burgess, S. and Thompson, S. 2012. Improvement of bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statistics in Medicine*, 31(15):1582–1600. (Cited on pages 60, 62, 73, and 108.)

Burgess, S. and Thompson, S. 2013. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, 42(4):1134–1144. (Cited on pages 125 and 126.)

Burgess, S. and Thompson, S. 2014. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology*, available online before print. (Cited on page 178.)

Burgess, S., Thompson, S., and CRP CHD Genetics Collaboration 2011b. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, 40(3):755–764. (Cited on pages 68 and 116.)

Burgess, S., Thompson, S., and CRP CHD Genetics Collaboration 2012. Methods for meta-analysis of individual participant data from Mendelian randomization studies with binary outcomes. *Statistical Methods in Medical Research*, available online before print. (Cited on pages 157 and 161.)

Cai, B., Small, D., and Ten Have, T. 2011. Two-stage instrumental variable methods for estimating the causal odds ratio: Analysis of bias. *Statistics in Medicine*, 30(15):1809–1824. (Cited on page 60.)

Casas, J., Bautista, L., Smeeth, L., Sharma, P., and Hingorani, A. 2005. Homocysteine and stroke: evidence on a causal link from Mendelian randomisation. *The Lancet*, 365(9455):224–232. (Cited on page 11.)

CCGC (CRP CHD Genetics Collaboration) 2008. Collaborative pooled analysis of data on C-reactive protein gene variants and coronary disease: judging causality by Mendelian randomisation. *European Journal of Epidemiology*, 23(8):531–540. (Cited on pages 9, 117, 157, and 161.)

CCGC (CRP CHD Genetics Collaboration) 2011. Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *British Medical Journal*, 342:d548. (Cited on pages 88, 116, 159, and 162.)

Chaussé, P. 2010. Computing generalized method of moments and generalized empirical likelihood with *R*. *Journal of Statistical Software*, 34(11):1–35. (Cited on page 72.)

Chen, L., Davey Smith, G., Harbord, R., and Lewis, S. 2008. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLoS Medicine*, 5(3):e52. (Cited on page 11.)

Cheung, B., Lauder, I., Lau, C., and Kumana, C. 2004. Meta-analysis of large randomized controlled trials to evaluate the impact of statins on cardiovascular outcomes. *British Journal of Clinical Pharmacology*, 57(5):640–651. (Cited on page 91.)

Christenfeld, N., Sloan, R., Carroll, D., and Greenland, S. 2004. Risk factors, confounding, and the illusion of statistical control. *Psychosomatic Medicine*, 66(6):868–875. (Cited on page 16.)

Clarke, P., Palmer, T., and Windmeijer, F. 2011. Estimating structural mean models with multiple instrumental variables using the generalised method of moments. The Centre for Market and Public Organisation 11/266, Centre for Market and Public Organisation, University of Bristol, UK. (Cited on page 72.)

Clarke, P. and Windmeijer, F. 2010. 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. (Cited on pages 38, 63, 65, and 66.)

Clarke, R., Peden, J., Hopewell, J., et al. 2009. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *New England Journal of Medicine*, 361(26):2518–2528. (Cited on pages 11 and 81.)

Cohen, J., Boerwinkle, E., Mosley Jr, T., and Hobbs, H. 2006. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *New England Journal of Medicine*, 354(12):1264–1272. (Cited on page 95.)

Collett, D. 2003. *Modelling survival data in medical research.* Chapman and Hall/CRC Press. (Cited on page 175.)

Collins, R., Armitage, J., Parish, S., Sleight, P., Peto, R. (Heart Protection Study Collaborative Group) 2002. MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20536 high-risk individuals: a randomised placebo-controlled trial. *The Lancet*, 360(9326):23–33. (Cited on page 4.)

Cox, D. 1958. *Planning of experiments. Section 2: Some key assumptions.* Wiley. (Cited on page 41.)

Danesh, J. and Pepys, M. 2009. C-reactive protein and coronary disease: is there a causal link? *Circulation*, 120(21):2036–2039. (Cited on page 7.)

Darwin, C. 1871. *The descent of man and selection in relation to sex.* Murray, London. (Cited on page 5.)

Dastani, Z., Hivert, M.-F., Timpson, N., et al. 2012. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: A multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genetics*, 8(3):e1002607. (Cited on page 142.)

Davey Smith, G. 2006. Randomised by (your) god: robust inference from an observational study design. *Journal of Epidemiology and Community Health*, 60(5):382–388. (Cited on page 88.)

Davey Smith, G. 2011. Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. *Genes & Nutrition*, 6(1):27–43. (Cited on pages 17, 35, and 89.)

Davey Smith, G. and Ebrahim, S. 2003. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22. (Cited on pages 4, 5, and 20.)

Davey Smith, G. and Ebrahim, S. 2004. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42. (Cited on pages 23 and 123.)

Davey Smith, G., Lawlor, D., Harbord, R., Timpson, N., Day, I., and Ebrahim, S. 2007. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Medicine*, 4(12):e352. (Cited on page 18.)

Davidson, R. and MacKinnon, J. 1993. *Estimation and inference in econometrics. Chapter 18: Simultaneous equation models.* Oxford University Press. (Cited on page 61.)

Davidson, R. and MacKinnon, J. 2014. Confidence sets based on inverting Anderson–Rubin tests. *The Econometrics Journal*, 17(2):S39–S58. (Cited on page 54.)

Davies, N., von Hinke Kessler Scholder, S., Farbmacher, H., Burgess, S., Windmeijer, F., and Davey Smith, G. 2014. The many weak instrument problem and Mendelian randomization. *Statistics in Medicine*, available online before print. (Cited on pages 61, 108, and 126.)

Davignon, J. and Laaksonen, R. 1999. Low-density lipoprotein-independent effects of statins. *Current Opinion in Lipidology*, 10(6):543–559. (Cited on page 91.)

Dawid, A. 2000. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424. (Cited on page 26.)

Dawid, A. 2002. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189. (Cited on page 37.)

Debat, V. and David, P. 2001. Mapping phenotypes: canalization, plasticity and developmental stability. *Trends in Ecology & Evolution*, 16(10):555–561. (Cited on page 31.)

Dekkers, O., von Elm, E., Algra, A., Romijn, J., and Vandenbroucke, J. 2010. How to assess the external validity of therapeutic trials: a conceptual approach. *International Journal of Epidemiology*, 39(1):89–94. (Cited on page 93.)

Didelez, V., Meng, S., and Sheehan, N. 2010. Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25(1):22–40. (Cited on pages 38, 48, and 51.)

Didelez, V. and Sheehan, N. 2007. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330. (Cited on pages 34, 37, 42, 43, and 51.)

Ding, E., Song, Y., Manson, J., et al. 2009a. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. *New England Journal of Medicine*, 361(12):1152–1163. (Cited on page 11.)

Ding, W., Lehrer, S., Rosenquist, J., and Audrain-McGovern, J. 2009b. The impact of poor health on academic performance: new evidence using genetic markers. *Journal of Health Economics*, 28(3):578–597. (Cited on page 11.)

Drukker, D. 2009. Generalized method of moments estimation in Stata 11. Technical report, Stata Corp. (Cited on page 71.)

Ducharme, G. and LePage, Y. 1986. Testing collapsibility in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2):197–205. (Cited on page 58.)

Dudbridge, F. 2013. Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, 9(3):e1003348. (Cited on page 182.)

Dunn, G., Maracy, M., and Tomenson, B. 2005. Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. *Statistical Methods in Medical Research*, 14(4):369–395. (Cited on page 65.)

Ebrahim, S. and Davey Smith, G. 2008. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics*, 123(1):15–33. (Cited on pages 11, 35, and 56.)

Efron, B. and Tibshirani, R. 1993. *An introduction to the bootstrap*. Chapman & Hall/CRC Press. (Cited on page 54.)

Ehret, G., Munroe, P., Rice, K., et al. (The International Consortium for Blood Pressure Genome-Wide Association Studies) 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478:103–109. (Cited on page 93.)

Elliott, P., Chambers, J., Zhang, W., et al. 2009. Genetic loci associated with C-reactive protein levels and risk of coronary heart disease. *Journal of the American Medical Association*, 302(1):37–48. (Cited on page 88.)

Emsley, R., Dunn, G., and White, I. 2010. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Statistical Methods in Medical Research*, 19(3):237–270. (Cited on page 180.)

Evans, D., Brion, M.-J., Paternoster, L., et al. 2013. Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genetics*, 9(10):e1003919. (Cited on page 182.)

Farnier, M. 2013. PCSK9 inhibitors. *Current Opinion in Lipidology*, 24(3):251–258. (Cited on page 95.)

Fieller, E. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 16(2):175–185. (Cited on page 52.)

Fischer-Lapp, K. and Goetghebeur, E. 1999. Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials*, 20(6):531–546. (Cited on page 65.)

Fisher, R. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433. (Cited on page 5.)

Foster, E. 1997. Instrumental variables for logistic regression: an illustration. *Social Science Research*, 26(4):487–504. (Cited on pages 58 and 64.)

Fox, J. 2006. Teacher's Corner: Structural Equation Modeling with the *sem* Package in R. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3):465–486. (Cited on page 72.)

Freeman, G., Cowling, B., and Schooling, M. 2013. Power and sample size calculations for Mendelian randomization studies. *International Journal of Epidemiology*, 42(4):1157–1163. (Cited on pages 127 and 131.)

Fried, L., Borhani, N., Enright, P., et al. 1991. The Cardiovascular Health Study: Design and rationale. *Annals of Epidemiology*, 1(3):263–276. (Cited on page 164.)

Frost, C. and Thompson, S. 2000. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(2):173–189. (Cited on page 19.)

Geiger, D., Verma, T., and Pearl, J. 1990. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534. (Cited on page 27.)

Gidding, S., Daniels, S., Kavey, R., and Expert Panel on Cardiovascular Health and Risk Reduction in Youth 2012. Developing the 2011 integrated pediatric guidelines for cardiovascular risk reduction. *Pediatrics*, 129(5):e1311–e1319. (Cited on page 84.)

Glymour, M., Tchetgen Tchetgen, E., and Robins, J. 2012. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339. (Cited on pages 34, 35, 68, and 95.)

Grassi, M., Assanelli, D., and Pezzini, A. 2007. Direct, reverse or reciprocal causation in the relation between homocysteine and ischemic heart disease. *Thrombosis Research*, 120(1):61–69. (Cited on page 179.)

Greenland, S. 1987. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, 125(5):761–768. (Cited on page 42.)

Greenland, S. 2000a. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729. (Cited on page 14.)

Greenland, S. 2000b. Causal analysis in the health sciences. *Journal of the American Statistical Association*, 95(449):286–289. (Cited on page 26.)

Greenland, S., Lanes, S., and Jara, M. 2008. Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. *Clinical Trials*, 5(1):5–13. (Cited on page 65.)

Greenland, S. and Robins, J. 1986. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419. (Cited on pages 14, 28, and 30.)

Greenland, S., Robins, J., and Pearl, J. 1999. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46. (Cited on page 58.)

Hahn, J., Hausman, J., and Kuersteiner, G. 2004. Estimation with weak instruments: accuracy of higher-order bias and MSE approximations. *Econometrics Journal*, 7(1):272–306. (Cited on page 61.)

Hall, A., Rudebusch, G., and Wilcox, D. 1996. Judging instrument relevance in instrumental variables estimation. *International Economic Review*, 37(2):283–298. (Cited on pages 111 and 114.)

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 50(4):1029–1054. (Cited on page 64.)

Hansson, G. 2005. Inflammation, atherosclerosis, and coronary artery disease. *The New England Journal of Medicine*, 352(16):1685–1695. (Cited on page 116.)

Hardin, J., Schmiediche, H., and Carroll, R. 2003. Instrumental variables, bootstrapping, and generalized linear models. *Stata Journal*, 3(4):351–360. (Cited on pages 71 and 72.)

Hayashi, F. 2000. *Econometrics*. Princeton University Press. (Cited on page 61.)

Hennekens, C., Buring, J., Manson, J., et al. 1996. Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *New England Journal of Medicine*, 334(18):1145–1149. (Cited on page 4.)

Hernán, M. and Robins, J. 2006. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372. (Cited on pages 17, 32, and 39.)

Hernán, M. and Taubman, S. 2008. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32:S8–S14. (Cited on page 136.)

Higgins, J., Thompson, S., Deeks, J., and Altman, D. 2003. Measuring inconsistency in meta-analyses. *British Medical Journal*, 327(7414):557–560. (Cited on page 168.)

Hill, A. B. 1965. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58(5):295–300. (Cited on page 35.)

Hingorani, A. and Humphries, S. 2005. Nature's randomised trials. *The Lancet*, 366(9501):1906–1908. (Cited on page 69.)

Holland, P. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960. (Cited on pages 25 and 26.)

Hooper, L., Ness, A., and Davey Smith, G. 2001. Antioxidant strategy for cardiovascular diseases. *The Lancet*, 357(9269):1705–1706. (Cited on page 4.)

Horowitz, J. L. 2011. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394. (Cited on pages 42 and 176.)

Imbens, G. and Rosenbaum, P. 2005. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):109–126. (Cited on pages 54 and 108.)

Inoue, A. and Solon, G. 2010. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561. (Cited on pages 137 and 155.)

Irons, D., McGue, M., Iacono, W., and Oetting, W. 2007. Mendelian randomization: A novel test of the gateway hypothesis and models of gene–environment interplay. *Development and Psychopathology*, 19(4):1181–1195. (Cited on page 11.)

Jansen, R. and Nap, J.-P. 2001. Genetical genomics: the added value from segregation. *Trends in Genetics*, 17(7):388–391. (Cited on page 180.)

Johnson, T. 2011. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Technical report, Queen Mary University of London. (Cited on page 142.)

Johnston, K., Gustafson, P., Levy, A., and Grootendorst, P. 2008. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, 27(9):1539–1556. (Cited on pages 64 and 65.)

Jones, E., Thompson, J., Didelez, V., and Sheehan, N. 2012. On the choice of parameterisation and priors for the Bayesian analyses of Mendelian randomisation studies. *Statistics in Medicine*, 31(14):1483–1501. (Cited on page 62.)

Kamstrup, P., Tybjaerg-Hansen, A., Steffensen, R., and Nordestgaard, B. 2009. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *Journal of the American Medical Association*, 301(22):2331–2339. (Cited on pages 11 and 80.)

Kaptoge, S., Di Angelantonio, E., Lowe, G. et al. (Emerging Risk Factors Collaboration) 2010. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *The Lancet*, 375(9709):132–140. (Cited on pages 7 and 116.)

Keavney, B. 2011. C reactive protein and the risk of cardiovascular disease. *British Medical Journal*, 342:d144. (Cited on page 171.)

Keavney, B., Danesh, J., Parish, S., et al. 2006. Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'. *International Journal of Epidemiology*, 35(4):935–943. (Cited on page 75.)

Khaw, K., Bingham, S., Welch, A., et al. 2001. Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study: a prospective population study. *The Lancet*, 357(9257):657–663. (Cited on page 4.)

Kinal, T. 1980. The existence of moments of k-class estimators. *Econometrica*, 48(1):241–249. (Cited on page 57.)

Kivimäki, M., Lawlor, D., Davey Smith, G., et al. 2008. Does high C-reactive protein concentration increase atherosclerosis? The Whitehall II Study. *PLoS ONE*, 3(8):e3013. (Cited on page 11.)

Kivimäki, M., Lawlor, D., Eklund, C., et al. 2007. Mendelian randomization suggests no causal association between C-reactive protein and carotid intima-media thickness in the young Finns study. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 27(4):978–979. (Cited on page 11.)

Kleiber, C. and Zeileis, A. 2014. *AER: Applied Econometrics with R.* R package version 1.2-2. (Cited on page 72.)

Kleibergen, F. and Zivot, E. 2003. Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics*, 114(1):29–72. (Cited on page 62.)

Klein, R., Zeiss, C., Chew, E., et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389. (Cited on page 181.)

Law, M., Morris, J., and Wald, N. 2009. Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. *British Medical Journal*, 338:b1665. (Cited on page 93.)

Law, M., Wald, N., and Rudnicka, A. 2003. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *British Medical Journal*, 326(7404):1423. (Cited on page 91.)

Lawlor, D., Harbord, R., Sterne, J., Timpson, N., and Davey Smith, G. 2008. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163. (Cited on pages 20, 21, 52, and 67.)

Le Marchand, L., Derby, K., Murphy, S., et al. 2008. Smokers with the *CHRNA* lung cancer–associated variants are exposed to higher levels of nicotine equivalents and a carcinogenic tobacco-specific nitrosamine. *Cancer Research*, 68(22):9137–9140. (Cited on page 179.)

Lewis, S. and Davey Smith, G. 2005. Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiology Biomarkers & Prevention*, 14(8):1967–1971. (Cited on page 23.)

Little, R. and Rubin, D. 2002. *Statistical analysis with missing data (2nd edition)*. Wiley. (Cited on page 134.)

Lunn, D., Whittaker, J., and Best, N. 2006. A Bayesian toolkit for genetic association studies. *Genetic Epidemiology*, 30(3):231–247. (Cited on page 135.)

Maldonado, G. and Greenland, S. 2002. Estimating causal effects. *International Journal of Epidemiology*, 31(2):422–429. (Cited on page 26.)

Manolio, T. 2010. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176. (Cited on page 181.)

Martens, E., Pestman, W., de Boer, A., Belitser, S., and Klungel, O. 2006. Instrumental variables: application and limitations. *Epidemiology*, 17(3):260–267. (Cited on pages 16 and 93.)

McPherson, J., et al. (The International Human Genome Mapping Consortium) 2001. A physical map of the human genome. *Nature*, 409(6822):934–941. (Cited on page 5.)

Mendel, G. 1866. Versuche über Pflanzen-hybriden. *Verhandlungen des naturforschenden Vereines in Brünn [Proceedings of the Natural History Society of Brünn]*, 4:3–47. (Cited on page 5.)

Mikusheva, A. 2010. Robust confidence sets in the presence of weak instruments. *Journal of Econometrics*, 157(2):236–247. (Cited on pages 54 and 108.)

Mikusheva, A. and Poi, B. 2006. Tests and confidence sets with correct size when instruments are potentially weak. *Stata Journal*, 6(3):335–347. (Cited on pages 54 and 108.)

Minelli, C., Thompson, J., Tobin, M., and Abrams, K. 2004. An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. *American Journal of Epidemiology*, 160(5):445–452. (Cited on pages 51, 52, and 140.)

Mogstad, M. and Wiswall, M. 2010. Linearity in instrumental variables estimation: Problems and solutions. Technical report, Forschungsinstitut zur Zukunft der Arbeit. Bonn, Germany. (Cited on pages 42 and 176.)

Moreira, M. 2003. A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048. (Cited on pages 54 and 108.)

Moreira, M., Porter, J., and Suarez, G. 2009. Bootstrap validity for the score test when instruments may be weak. *Journal of Econometrics*, 149(1):52–64. (Cited on page 54.)

Morris, A., Voight, B., Teslovich, T., et al. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9):981–990. (Cited on page 146.)

Mumby, H., Elks, C., Li, S., et al. 2011. Mendelian randomisation study of childhood BMI and early menarche. *Journal of Obesity*, available online before print. (Cited on page 11.)

Nagelkerke, N., Fidler, V., Bernsen, R., and Borgdorff, M. 2000. Estimating treatment effects in randomized clinical trials in the presence of noncompliance. *Statistics in Medicine*, 19(14):1849–1864. (Cited on page 60.)

Nelson, C. and Startz, R. 1990. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business*, 63(1):125–140. (Cited on pages 99 and 126.)

Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B., Whittaker, J., and Leon, D. 2006. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *American Journal of Epidemiology*, 163(5):397–403. (Cited on page 16.)

Norton, E. and Han, E. 2008. Genetic information, obesity, and labor market outcomes. *Health Economics*, 17(9):1089–1104. (Cited on page 11.)

Ogbuanu, I., Zhang, H., and Karmaus, W. 2009. Can we apply the Mendelian randomization methodology without considering epigenetic effects? *Emerging Themes in Epidemiology*, 6(1):3. (Cited on page 32.)

Palmer, T., Lawlor, D., Harbord, R., et al. 2011a. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research*, 21(3):223–242. (Cited on pages 114 and 123.)

Palmer, T., Sterne, J., Harbord, R., et al. 2011b. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. *American Journal of Epidemiology*, 173(12):1392–1403. (Cited on pages 64 and 71.)

Palmer, T., Thompson, J., Tobin, M., Sheehan, N., and Burton, P. 2008. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *International Journal of Epidemiology*, 37(5):1161–1168. (Cited on pages 37 and 60.)

Pauling, L., Itano, H., Singer, S., and Wells, I. 1949. Sickle cell anemia, a molecular disease. *Science*, 110(2865):543–548. (Cited on page 5.)

Pearl, J. 2000a. *Causality: models, reasoning, and inference*. Cambridge University Press. (Cited on page 25.)

Pearl, J. 2000b. *Causality: models, reasoning, and inference. Chapter 3, Section 3.1: The back-door criterion*. Cambridge University Press. (Cited on page 28.)

Pearl, J. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. (Cited on page 180.)

Pearl, J. 2010. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):1–60. (Cited on page 26.)

Peto, R., Doll, R., Buckley, J., and Sporn, M. 1981. Can dietary beta-carotene materially reduce human cancer rates? *Nature*, 290:201–208. (Cited on page 4.)

Pierce, B., Ahsan, H., and VanderWeele, T. 2011. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology*, 40(3):740–752. (Cited on pages 114 and 131.)

Pierce, B. and Burgess, S. 2013. Efficient design for Mendelian randomization studies: subsample and two-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184. (Cited on page 137.)

Pierce, B. and VanderWeele, T. 2012. The effect of non-differential measurement error on bias, precision and power in Mendelian randomization studies. *International Journal of Epidemiology*, 41(5):1383–1393. (Cited on page 19.)

Plenge, R., Scolnick, E., and Altshuler, D. 2013. Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery*, 12(8):581–594. (Cited on page 94.)

R Development Core Team 2011. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (Cited on page 69.)

Raal, F., Giugliano, R., Sabatine, M., et al. 2014. Reduction in lipoprotein(a) with PCSK9 monoclonal antibody evolocumab (AMG 145): a pooled analysis of more than 1,300 patients in 4 phase II trials. *Journal of the American College of Cardiology*, 63(13):1278–1288. (Cited on page 95.)

Rasbash, J., Steele, F., Browne, W., and Goldstein, H. 2009. *A user's guide to MLwiN, v2. 10*. Centre for Multilevel Modelling, University of Bristol. (Cited on page 141.)

Relton, C. and Davey Smith, G. 2010. Epigenetic epidemiology of common complex disease: Prospects for prediction, prevention, and treatment. *PLoS Medicine*, 7(10):e1000356. (Cited on page 180.)

Relton, C. and Davey Smith, G. 2012a. Is epidemiology ready for epigenetics? *International Journal of Epidemiology*, 41(1):5–9. (Cited on page 180.)

Relton, C. and Davey Smith, G. 2012b. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *International Journal of Epidemiology*, 41(1):161–176. (Cited on page 181.)

Riley, R., Abrams, K., Sutton, A., Lambert, P., and Thompson, J. 2007. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, 7(1):3. (Cited on page 141.)

Roberts, L., Davenport, R., Pennisi, E., and Marshall, E. 2001. A history of the Human Genome Project. *Science*, 291(5507):1195. (Cited on page 5.)

Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512. (Cited on page 65.)

Robins, J. 1992. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*, 79(2):311–334. (Cited on page 176.)

Robins, J. 1994. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics – Theory and Methods*, 23(8):2379–2412. (Cited on page 65.)

Robins, J. 1999. *Statistical models in epidemiology: the environment and clinical trials*, chapter Marginal structural models versus structural nested models as tools for causal inference, pages 95–134. Springer. (Cited on page 66.)

Robinson, J., Nedergaard, B., Rogers, W., et al. 2014. Effect of evolocumab or ezetimibe added to moderate-or high-intensity statin therapy on LDL-C lowering in patients with hypercholesterolemia: the LAPLACE-2 randomized clinical trial. *Journal of the American Medical Association*, 311(18):1870–1882. (Cited on page 95.)

Rossouw, J., et al. (Writing Group for the Women's Health Initiative Investigators) 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association*, 288(3):321–333. (Cited on page 4.)

Rothwell, P. 2010. Commentary: External validity of results of randomized trials: disentangling a complex concept. *International Journal of Epidemiology*, 39(1):94–96. (Cited on page 87.)

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701. (Cited on page 30.)

Sargan, J. 1958. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415. (Cited on page 68.)

Sarwar, N., Butterworth, A., Freitag, D., et al. (IL6R Genetics Consortium and Emerging Risk Factors Collaboration) 2012. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. *Lancet*, 379(9822):1205–1213. (Cited on page 20.)

SAS (SAS Institute and SAS Publishing Staff) 2004. *SAS/STAT 9.1 User's Guide*. SAS Institute Inc, Cary, NC. (Cited on page 69.)

Schatzkin, A., Abnet, C., Cross, A., et al. 2009. Mendelian randomization: how it can – and cannot – help confirm causal relations between nutrition and cancer. *Cancer Prevention Research*, 2(2):104–113. (Cited on pages 24 and 89.)

Scheet, P. and Stephens, M. 2006. 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*, 78(4):629–644. (Cited on page 134.)

Schunkert, H., König, I., Kathiresan, S., et al. 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, 43(4):333–338. (Cited on page 146.)

Sheehan, N., Didelez, V., Burton, P., and Tobin, M. 2008. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Medicine*, 5(8):e177. (Cited on page 20.)

Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145. (Cited on page 5.)

Small, D. 2014. *ivpack: Instrumental Variable Estimation*. R package version 1.1. (Cited on pages 54 and 72.)

Small, D. and Rosenbaum, P. 2008. War and wages: the strength of instrumental variables and their sensitivity to unobvserved biases. *Journal of the American Statistical Association*, 103(483):924–933. (Cited on page 121.)

Speliotes, E., Willer, C., Berndt, S., et al. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, 42(11):937–948. (Cited on page 176.)

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639. (Cited on page 152.)

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. 2003. WinBUGS version 1.4 user manual. Technical report, MRC Biostatistics Unit, Cambridge, UK. (Cited on pages 62, 69, and 141.)

Spirtes, P., Glymour, C., and Scheines, R. 2000. *Causation, prediction, and search*. MIT Press. (Cited on page 39.)

Staiger, D. and Stock, J. 1997. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586. (Cited on pages 67 and 106.)

StataCorp 2009. *Stata Statistical Software: Release 11*. College Station, TX. (Cited on page 69.)

Sterne, J. and Davey Smith, G. 2001. Sifting the evidence – What's wrong with significance tests? *British Medical Journal*, 322:226–231. (Cited on page 96.)

Stock, J., Wright, J., and Yogo, M. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*, 20(4):518–529. (Cited on page 67.)

Stock, J. and Yogo, M. 2002. Testing for weak instruments in linear IV regression. *SSRN eLibrary*, 11:T0284. (Cited on pages 70, 108, and 118.)

Stukel, T., Fisher, E., Wennberg, D., Alter, D., Gottlieb, D., and Vermeulen, M. 2007. Analysis of observational studies in the presence of treatment selection bias. *Journal of the American Medical Association*, 297(3):278–285. (Cited on page 59.)

Sussman, J. and Hayward, R. 2010. An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. *British Medical Journal*, 340:c2073. (Cited on page 14.)

Sutton, A., Kendrick, D., and Coupland, C. 2008. Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine*, 27(5):651–669. (Cited on page 150.)

Swanson, S. and Hernán, M. 2013. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology*, 24(3):370–374. (Cited on page 42.)

Swerdlow, D., Holmes, M., Kuchenbaecker, K., et al. (The Interleukin-6 Receptor Mendelian Randomisation Analysis Consortium) 2012. The interleukin-6 receptor as a target for prevention of coronary heart disease: a Mendelian randomisation analysis. *Lancet*, 379(9822):1214–1224. (Cited on pages 19, 36, and 183.)

Taubes, G. and Mann, C. 1995. Epidemiology faces its limits. *Science*, 269(5221):164–169. (Cited on page 4.)

Taylor, A., Davies, N., Ware, J., VanderWeele, T., Davey Smith, G., and Munafò, M. 2014. Mendelian randomization in health research: Using appropriate genetic variants and avoiding biased estimates. *Economics & Human Biology*, 13:99–106. (Cited on page 125.)

Taylor, F., Ward, K., Moore, T., et al. 2013. Statins for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews*, 2013:1. (Cited on page 91.)

Terza, J., Basu, A., and Rathouz, P. 2008. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3):531–543. (Cited on page 60.)

Thomas, D. and Conti, D. 2004. Commentary: the concept of 'Mendelian Randomization'. *International Journal of Epidemiology*, 33(1):21–25. (Cited on page 14.)

Thomas, D., Lawlor, D., and Thompson, J. 2007. Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. *Annals of Epidemiology*, 17(7):511–513. (Cited on pages 52 and 69.)

Thompson, J., Minelli, C., Abrams, K., Tobin, M., and Riley, R. 2005. Meta-analysis of genetic studies using Mendelian randomization – a multivariate approach. *Statistics in Medicine*, 24(14):2241–2254. (Cited on pages 11, 90, 116, 141, 144, 145, and 154.)

Timpson, N., Harbord, R., Davey Smith, G., Zacho, J., Tybjærg-Hansen, A., and Nordestgaard, B. 2009. Does greater adiposity increase blood pressure and hypertension risk? Mendelian randomization using the *FTO/MC4R* genotype. *Hypertension*, 54(1):84–90. (Cited on page 77.)

Timpson, N., Lawlor, D., Harbord, R., et al. 2005. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. *The Lancet*, 366(9501):1954–1959. (Cited on page 11.)

Timpson, N., Nordestgaard, B., Harbord, R., et al. 2011. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *International Journal of Obesity*, 35:300–308. (Cited on page 179.)

Tobin, M., Minelli, C., Burton, P., and Thompson, J. 2004. Commentary: Development of Mendelian randomization: from hypothesis test to 'Mendelian deconfounding'. *International Journal of Epidemiology*, 33(1):26–29. (Cited on pages 16 and 39.)

Trompet, S., Jukema, J., Katan, M., et al. 2009. Apolipoprotein E genotype, plasma cholesterol, and cancer: a Mendelian randomization study. *American Journal of Epidemiology*, 170(11):1415–1421. (Cited on page 11.)

VanderWeele, T. 2009. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883. (Cited on page 50.)

VanderWeele, T., Asomaning, K., Tchetgen Tchetgen, E., et al. 2012. Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. *American Journal of Epidemiology*, 175(10):1013–1020. (Cited on page 179.)

VanderWeele, T., Tchetgen Tchetgen, E., Cornelis, M., and Kraft, P. 2014. Methodological challenges in Mendelian randomization. *Epidemiology*, 25(3):427–435. (Cited on page 96.)

VanderWeele, T. and Vansteelandt, S. 2009. Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2(4):457–468. (Cited on page 180.)

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. 2011. On instrumental variables estimation of causal odds ratios. *Statistical Science*, 26(3):403–422. (Cited on pages 59 and 66.)

Vansteelandt, S. and Goetghebeur, E. 2003. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):817–835. (Cited on page 66.)

Voight, B., Peloso, G., Orho-Melander, M., et al. 2012. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*, 380(9841):572–580. (Cited on pages 11, 82, and 177.)

von Hinke Kessler Scholder, S., Davey Smith, G., Lawlor, D., Propper, C., and Windmeijer, F. 2010. Genetic markers as instrumental variables: An application to child fat mass and academic achievement. The Centre for

Market and Public Organisation 10/229, University of Bristol, UK. (Cited on page 11.)

Wald, A. 1940. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11(3):284–300. (Cited on page 45.)

Wan, E., Qiu, W., Baccarelli, A., et al. 2012. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Human Molecular Genetics*, 21(13):3073–3082. (Cited on page 180.)

Wardle, J., Carnell, S., Haworth, C., Farooqi, I., O'Rahilly, S., and Plomin, R. 2008. Obesity associated genetic variation in *FTO* is associated with diminished satiety. *Journal of Clinical Endocrinology & Metabolism*, 93(9):3640–3643. (Cited on pages 31, 78, and 89.)

Waterworth, D., Ricketts, S., Song, K., et al. 2010. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 30(11):2264–2276. (Cited on page 90.)

Watson, J. and Crick, F. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738. (Cited on page 5.)

Wehby, G., Ohsfeldt, R., and Murray, J. 2008. "Mendelian randomization" equals instrumental variable analysis with genetic instruments. *Statistics in Medicine*, 27(15):2745–2749. (Cited on pages 14 and 68.)

Welsh, P., Polisecki, E., Robertson, M., et al. 2010. Unraveling the directional link between adiposity and inflammation: a bidirectional Mendelian randomization approach. *Journal of Clinical Endocrinology & Metabolism*, 95(1):93–99. (Cited on page 179.)

Wooldridge, J. 2009. *Introductory econometrics: A modern approach. Chapter 5: Multiple regression analysis – OLS asymptotics*. South-Western, Nashville, TN. (Cited on pages 56, 67, and 127.)

Wright, P. 1928. *The tariff on animal and vegetable oils. Appendix B*. Macmillan, New York. (Cited on page 10.)

Yang, J., Lee, S., Goddard, M., and Visscher, P. 2011. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82. (Cited on page 182.)

Youngman, L., Keavney, B., Palmer, A., et al. 2000. Plasma fibrinogen and fibrinogen genotypes in 4685 cases of myocardial infarction and in 6002 controls: test of causality by 'Mendelian randomization'. *Circulation*, 102(Suppl II):31–32. (Cited on page 9.)

Zacho, J., Tybjaerg-Hansen, A., Jensen, J., Grande, P., Sillesen, H., and Nordestgaard, B. 2008. Genetically elevated C-reactive protein and ischemic vascular disease. *New England Journal of Medicine*, 359(18):1897–1908. (Cited on page 100.)

Zohoori, N. and Savitz, D. 1997. Econometric approaches to epidemiologic data: Relating endogeneity and unobserved heterogeneity to confounding. *Annals of Epidemiology*, 7(4):251–257. (Cited on pages 56 and 110.)

Chapman & Hall/CRC Interdisciplinary Statistics Series

Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation provides thorough coverage of the methods and practical elements of Mendelian randomization analysis. It brings together diverse aspects of Mendelian randomization spanning epidemiology, statistics, genetics, and econometrics.

Through several examples, the first part of the book shows how to perform simple applied Mendelian randomization analyses and interpret their results. The second part addresses specific methodological issues, such as weak instruments, multiple instruments, power calculations, and meta-analysis, relevant to practical applications of Mendelian randomization. In this part, the authors draw on data from the C-reactive protein Coronary heart disease Genetics Collaboration (CCGC) to illustrate the analyses. They present the mathematics in an easy-to-understand way by using nontechnical language and reinforcing key points at the end of each chapter. The last part of the book examines the potential of Mendelian randomization in the future, exploring both methodological and applied developments.

Features

- Offers first-hand, in-depth guidance on Mendelian randomization from leaders in the field
- Makes the diverse aspects of Mendelian randomization understandable to newcomers
- Illustrates the technical details using data from a large collaborative study
- Includes other real-world examples that show how Mendelian randomization is used in studies involving inflammation, heart disease, and more
- Discusses possible future directions for research involving Mendelian randomization

This book gives you the foundation to understand issues concerning the use of genetic variants as instrumental variables. It will get you up to speed in undertaking and interpreting Mendelian randomization analyses. Chapter summaries, paper summaries, web-based applications, and software code for implementing the statistical techniques are available on a supplementary website.

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K16638
ISBN: 978-1-4665-7317-8
90000
9 781466 573178
www.crcpress.com