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.