7

Mendelian randomisation: use of genetics to enable causal inference in observational studies

Marion Verduijn, Bob Siegerink, Kitty J Jager, Carmine Zoccali and Friedo W Dekker

Nephrol. Dial. Transplant. 2010; 25: 1394-98.

Abstract

The aim of aetiologic studies in epidemiology is to investigate whether factors are causally related to diseases and therefore become a potential target for therapeutic interventions. Mendelian randomisation enables estimation of causal relationships in observational studies with genetic variants as instrumental variables. An instrumental variable is a variable that can be considered to mimic the coin toss in a randomised study. Given the random assignment of alleles in gamete formation, the use of genetic variants is an alternative method to control for confounding. This chapter describes the approach of Mendelian randomisation, its underlying rationale and its assumptions.

Introduction

While a randomised controlled trial (RCT) is an unbeatable standard in intervention studies, RCTs might be inappropriate or even impossible to conduct when studying the effect of factors potentially involved in the aetiology of disease.[1] Observational study designs are the obvious means for studying these types of research questions, each with their own strengths and weaknesses.[2,3] Despite the valuable contribution of observational studies in understanding the aetiology of diseases, these studies are limited for making causal inference. Exposures that seemed to affect the risk of disease in observational studies turned out later to be non-causal (i.e. only a risk marker), as no or reverse effects of interventions on the presumed cause of the disease was shown in later RCTs. A likely explanation for the initially suggested relationships is 'residual confounding', caused by an incomplete or a lack of measurement of relevant identified or unidentified confounding factors. Moreover, the findings could be due to 'reverse causation', being the (subclinical) presence of disease influencing the presence of the exposure under investigation, rather than vice versa.[4] The subclinical presence of cancer, for instance, might cause leanness, instead of resulting from it, although leanness might be observed prior to cancer diagnosis (Figure 1). Observational studies, however, enable the estimation of causal relationships when an 'instrumental variable' is available that mimics the coin toss in a randomised study reducing confounding.[5] The use of 'genetic variants' as an instrumental variable in observational research is an example hereof. This approach is known as Mendelian randomisation.[6–8]

This chapter describes the approach of Mendelian randomisation used to make causal inference in observational studies, and its underlying rationale and assumptions. All concepts in italics are briefly defined in the glossary.

Example 1: Are low serum cholesterol levels a causal risk factor for cancer? The concept of Mendelian randomisation was originally suggested by Katan in 1986 in the debate on the hypothesis that low serum cholesterol levels (exposure) directly increase the risk of cancer (outcome).[9] In order to investigate whether the association between low serum cholesterol levels and cancer is causal, Katan suggested making use of the data of the apolipoprotein E (ApoE) gene. This gene is known to affect serum cholesterol levels, with the E2 variant being associated with lifelong lower serum cholesterol levels. Katan

hypothesised that, if the low serum cholesterol is a causal risk factor for cancer, an increased risk of cancer should be observed in individuals carrying the ApoE2 variant.

Rationale The main rationale in Mendelian randomisation is that, if an exposure is causally related to an outcome, a genetic variant, which is associated with the exposure, should have a similar relation to the outcome as the supposedly causal exposure itself. In contrast, if the genetic variant turns out to be not related to outcome, a causal role of the genetic product (i.e. the exposure) is less likely. In theory, the exposure of interest in Mendelian randomisation studies can be any property, but in general, protein levels as measured in blood are studied. In the study of Katan, the genetic variant ApoE2 was suggested for its association with the lower levels of serum cholesterol. Persons carrying the ApoE2 variant are lifelong exposed to lower cholesterol levels; if low cholesterol is indeed a causal risk factor for cancer, an increased risk of developing cancer would be expected in the carriers of the ApoE2 variant. A recent publication, however, clearly showed that carriers of the ApoE2 variant are not more susceptible to develop or die from cancer, ruling out low cholesterol as a causal risk factor for cancer.[10]

The use of a genetic variant as an instrumental variable for causal reasoning, whether it is a *single-nucleotide polymorphism* (SNP), a *haplotype* or a deletion, is directly based on the independent assortment of *alleles*. According to Mendel's second law (Gregor Mendel, 1822–1884), an individual's *genotype* is randomly assigned from his/her parental genotypes at gamete formation. In this respect, an observational study investigating the effect of a genetic variant has important similarities with an RCT studying a treatment effect. In Figure 2, both study designs are shown. The randomisation of treatment in an ideal RCT guarantees that all differences in patient characteristics are due to chance, so no confounding by indication is present. The differences observed during follow-up can be regarded as the likely sole effect of the treatment that was allocated in a randomised manner. Similarly, the random assortment of alleles guarantees that, when comparing patients according to a genetic variant, the differences in patient characteristics are due to chance except for the differences that result from the genetic variant. As such, confounding is eliminated. Moreover, since the genetic makeup is fixed at conception, a genetic variant cannot be influenced by the (subclinical) presence of disease, thereby excluding the possibility of reverse causation.

Figure 1. An association between an exposure and an outcome in observational studies might be due to (residual) confounding and/or reverse causation.

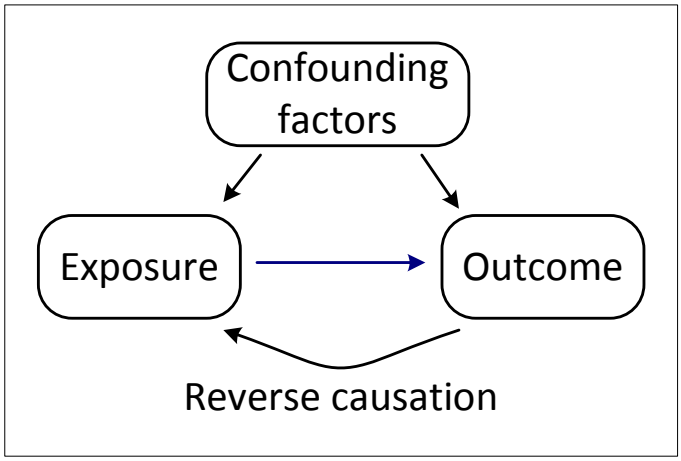
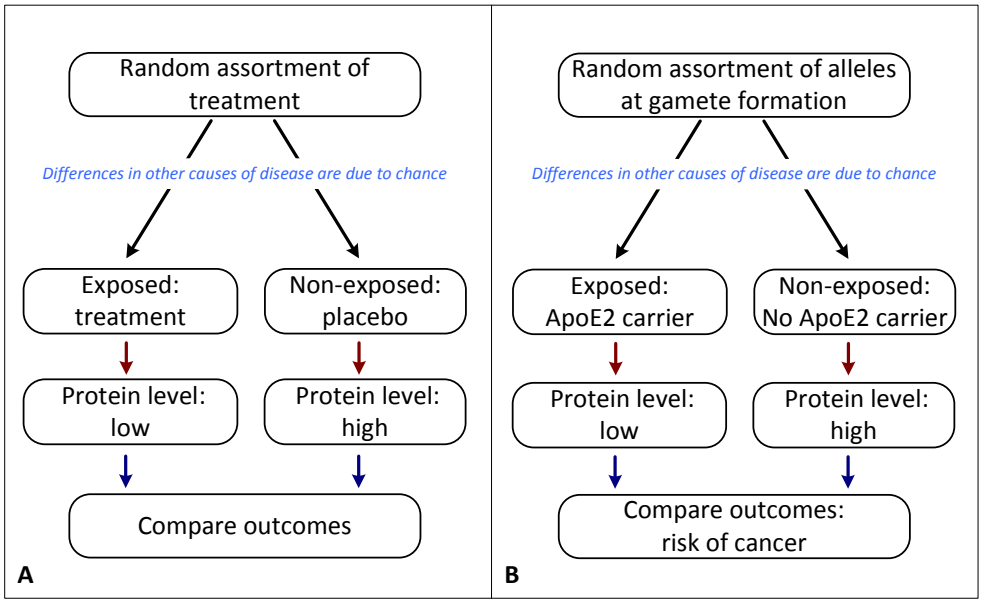


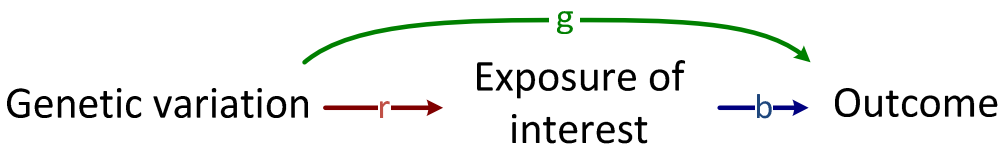
Figure 2. Parallel between RCT and Mendelian randomisation



Parallel between randomised control trials (panel A) and Mendelian randomisation (panel B), with in the diagram of Mendelian randomisation the original idea of Katan as an example.

Framework The framework of a Mendelian randomisation study is summarised in Figure 3 together with the text that describe the associations as examined in the studies used as examples in this chapter. Mendelian randomisation in its most basic form is to study the causal relationship between the exposure levels and the outcome (**blue arrow**) with a genetic variant known for influencing the exposure levels under investigation (**red arrow**) as an instrumental variable, by estimating the association between the genetic variant and the outcome (**green arrow**). In a more sophisticated approach, a quantitative analysis of the three associations is done. Given the observed association in the data between the genetic variant and the outcome and between the exposure level and the outcome, the expected associations can be calculated under the assumption that the exposure levels are causally related to the outcome. Under this assumption, the association between the genetic variant and the outcome (**green arrow**) is expected to be equal to the association between exposure levels and outcome (**blue arrow)** if a genetic variant would explain 100% of the variance in the exposure levels (**red arrow**). This relation between the exposure of interest and its proxy is also known as the instrument strength and can be

Figure 3. Framework of a Mendelian randomisation study; the table describes the associations as examined in the studies used as examples



Example	Association		
	blue arrow	red arrow	green arrow
Katan [9,10]	cholesterol levels – cancer	ApoE genotype – cholesterol levels	ApoE genotype – cancer
Zacho [15]	CRP levels – ischaemic heart disease Cholesterol levels – ischaemic heart disease	CRP genotype – CRP levels ApoE genotype – cholesterol levels	CRP genotype – ischaemic heart disease ApoE genotype – ischaemic heart disease

*The **blue arrow** (also denoted by a small letter b) represents the association between the exposure of interest and the outcome of the study. To determine whether this association indeed is a causal one, Mendelian randomisation uses the associations represented by the **red arrow**(r) and **green arrow**(g).*

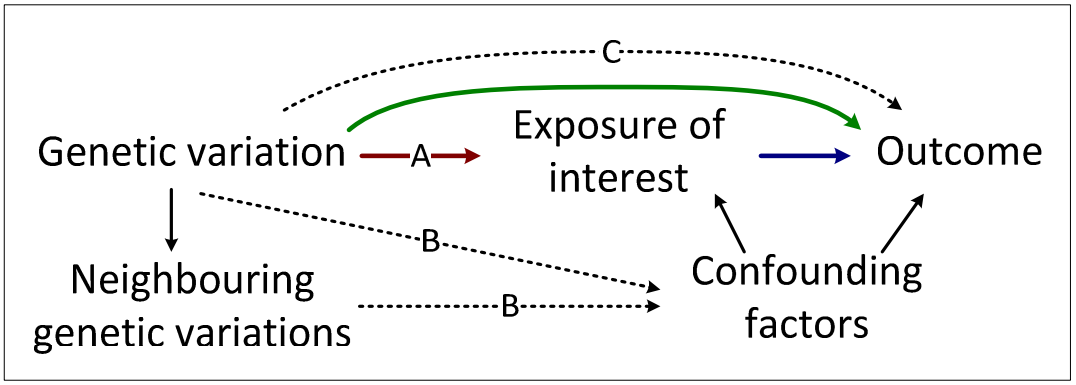
estimated for example with the F statistic.[11] A much more realistic scenario is that the genetic variant explains does not explain 100% of the variance in the exposure levels. In this case, the expected associations can be calculated in two stage regression analyses.[6] If the observed and expected associations between the genetic variant and the outcome and between the exposure level and the outcome are not similar (in direction and magnitude), this indicates that the observed association between the exposure levels and the outcome could be due to residual confounding and/or reverse causation. For more detailed methods for calculating the expected associations please refer to Lawler et al.[6] The use of this formal way of Mendelian randomisation analyses does require the presence of a strong instrument or, which is most often the case, a large study to circumvent the lack of a strong association between the exposure of interest and its instrument to prevent weak instrument bias.[11]

Assumptions Figure 4 depicts the three assumptions that are underlying a Mendelian randomisation study. The first assumption is that there is a robust association between the genetic variant and the exposure levels, i.e. the chosen instrument must not be weak. In order to ensure this, the association should preferably be demonstrated in independent (healthy) populations and confirmed in own data. More robust associations between genotype and exposure levels might be obtained with haplotypes or a genetic score as the instrumental variable instead of a single genetic variant (Figure 4, presence of association A). Also, additional variables which are not on the causal path can be used to minimise the variance of the exposure of interest to be explained by its instrument.[11,12] Where this first assumption can be evaluated, the other assumptions are about the absence of *pleiotropic* effects of the genetic variant which in theory cannot be verified empirically.

The second assumption is that the genetic variant is not associated with factors that confound the association between the exposure levels and outcome. The association between a genetic variant and a confounding factor can be direct, namely if the confounding factor itself is regulated by the genetic variant, or indirect when neighbouring variants (i.e. variants that are in linkage disequilibrium with the variant used as instrumental variable) regulate one or more confounding factors (Figure 4, absence of association B).

In the third assumption, the genetic variant is assumed to be independent of the outcome given the exposure levels and its confounding factors. This means that the genetic variant

Figure 4. Assumptions made in a Mendelian randomisation study



Three assumption may not be violated in order to yield an unbiased result from Mendelian randomisation analyses, denoted by a capital letter: ***A)*** *presence of a robust association between genetic variant and exposure (no weak instrument)* ***B)*** *absence of association, direct and indirect, between genetic variant and confounding factors (no pleiotropy / no confounding such as population stratification nor canalisation)* ***C)*** *absence of other pathways between genetic variants and outcome (no pleiotropy)*

is assumed to be exclusively related to the outcome via the exposure levels and that there are no other pathways to the outcome. Otherwise, the observed association between the genetic variant and the outcome also includes the impact of the variant on the outcome via other pathways. As such, this association does not prove that the particular exposure levels under investigation are causally related with the outcome (Figure 4, absence of association C). Knowledge of the function of the genetic variant and its neighbouring variants is needed to evaluate whether

the assumptions are likely to hold in a particular Mendelian randomisation study. When these assumptions hold, Mendelian randomisation can be used to test the null hypothesis that the exposure levels of interest are not associated with the outcome. The assumption that all associations are linear and unaffected by statistical interactions is needed when aiming at a precise estimate of the size of the causal effect. Although several approaches can be used to determine whether the assumptions are likely to hold, some of the assumptions can never be checked empirically, because they are based on the absence of a effect. Nonetheless, if thoroughly investigated and critically evaluated, Mendelian randomisation can be a valuable technique in causal inference.[13]

Some additional issues are relevant to mention in the context of Mendelian randomisation studies, as they may in part be underlying causes invalidating the assumptions. The first issue is ‘population stratification’, which is related to weak instrument bias and occurs

when the allele frequencies of the genetic variant and the distribution of the exposure levels (or outcome) vary substantially between the different subgroups in the study population.[6] In that case, an association is induced between the genetic variant and the exposure levels (or outcome) at the population level. This phenomenon can be regarded as confounding by ethnicity. Population stratification is unlikely to be a problem in practice, except in extreme situations,[14] and can be overcome by studying populations that are homogeneous with respect to ethnicity. A second issue is the possibility that a lifelong genetic (or environmental) exposure has induced developmental compensation via alternative pathways. This is known as '*canalisation*' and may invalidate the estimation of associations like those in Figure 3. In this perspective, it is worthy to note that Mendelian randomisation studies and RCTs are different in this respect: in RCTs, the random allocation treatment is normally done in adulthood, after the developmental period, while the random allocation of alleles takes place at conception, allowing the possibility of *canalisation*. It is, however, unclear how important this issue is in practice.[6]

Finally, as for all genetic association studies, relatively large sample sizes are required for Mendelian randomisation studies. The sample size highly depends on both the frequency of the genetic variation used as instrument, and the expected effect size of the association of interest between the genetic variant, the exposure and outcome. Conventional sample size calculations can be performed in order to get an approximation of the required sample size. Studies that are underpowered might be unable to observe small (but existing) effects, incorrectly suggesting that an association between an exposure and outcome is not causal. This encourages enlarged sample sizes through collaboration between research groups and by the setting up of well-defined cohorts. In addition, the evidence of the different studies can be combined in meta-analyses.

Example 2: Is CRP a causal risk factor for ischaemic vascular disease? The study of Zacho and colleagues shows an interesting application of the Mendelian randomisation approach. They aimed to test whether the observed association between the elevated levels of C-reactive protein (CRP) and increased risk of ischaemic vascular disease (i.e. ischaemic heart disease and ischaemic brain disease) is a causal association.[15] So, the exposure levels under investigation were CRP levels, and the outcome was ischaemic vascular disease. As the instrumental variable, they used four polymorphisms in the CRP gene that partially affect plasma CRP levels and combined them to one variable with nine genotype combinations. They observed that the risk of ischaemic heart disease was increased by a factor of 2.2

[95% confidence interval (CI) 1.6–2.9] and the risk of ischaemic brain disease was 60% increased [95%CI 1.1-2.5] in persons with CRP levels >3 mg/L as compared with persons with CRP levels <1 mg/L (adjusted for age, sex and statin use); this is the **blue arrow** in the framework of Figure 3 and 4. Moreover, the data confirmed that the CRP levels are regulated by the CRP genotype (**red arrow**): a difference of up to 64% in CRP levels among the genotype combinations. With these findings, an expected ischaemic heart disease relative risk of 1.32 (1.26–1.39) can be calculated for the genotype combination that was related to the highest CRP levels. However, no increased risk for ischaemic heart disease was observed for any of the CRP genotype combinations (**green arrow**). The results for ischaemic brain disease follow the same pattern. These findings suggests that the observed increased risk of ischaemic vascular disease associated to elevated CRP levels does not reflect a causal relationship, and the CRP is a mere risk marker for this outcome. In the same study, a proof of principle of the Mendelian randomisation approach was given, with the ApoE genotype as the instrumental variable for examining the causality of the association between cholesterol levels and ischaemic heart disease.[15] The increased risk for the outcome was expected across the ApoE genotypes [up to 1.12 (1.06– 1.17)] and was also observed [up to 1.35 (1.12–1.61)]. These results suggest that the association between cholesterol levels and ischaemic heart disease is indeed a causal one, although the difference in magnitude of effect could suggest the presence of residual confounding.

Conclusion Mendelian randomisation enables to study the hypothesis that an observed association between the exposure levels, as determined by the genetic variant used as the instrumental variable, and the outcome is causal. As such, it is a unique approach for providing more insights into potential causal relationships in aetiologic research using observational data, both as a formal analyses as well as a line of thought in causal inference.

Glossary

Alleles:	variant forms of a gene at a locus; a single allele is inherited from each parent.
Canalisation:	developmental compensation in response to disruptive influences on normal development from genetic and environmental forces.
Gene:	stretch of DNA which encodes for a particular protein.
Genetic variants:	variation in genes, including single-nucleotide polymorphisms, and insertion or deletions of stretches of DNA. Genotype: genetic makeup of an individual with regard to genetic variants.
Linkage:	tendency of DNA to be co-inherited due to their close physical proximity; linkage disequilibrium (LD) is a measure of linkage and indicates whether a non-random association between two alleles at different loci is present
Locus:	physical location of a gene or other genetic marker
Haplotypes:	combinations of SNPs; usually, a limited number of haplotypes can be used to cover most of the genetic variation within a population due to linkage
Pleiotropy:	phenomenon in which a genetic variant has multiple distinct phenotypic effects
Population stratification:	a form of confounding by race due to differences in allele frequencies in subgroups in a population
SNP:	or Single-Nucleotide Polymorphism. A genetic variant in which one specific nucleotide in the DNA is altered
Instrumental variable:	a variable in non-experimental data that can be considered to mimic the coin toss in a randomised study
Residual confounding:	confounding remaining after incomplete adjustment for confounders due to lack of measurement of relevant identified or unidentified confounding factors
Reverse causation:	(subclinical) presence of disease, or its effects, alters the exposure under investigation, rather than vice versa, i.e. cause and consequence are switched

1. Stel VS, Jager KJ, Zoccali C, Wanner C, Dekker FW. The randomized clinical trial: an unbeatable standard in clinical research? *Kidney Int*. 2007;72:539–42.

2. Jager KJ, Stel VS, Wanner C, Zoccali C, Dekker FW. The valuable contribution of observational studies to nephrology. *Kidney Int*. 2007;72:671–5.

3. Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet*. 2004;363:1728–31.

4. Sheehan N a, Didelez V, Burton PR, Tobin MD. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Med*. 2008;5:e177.

5. Myers J a, Rassen J a, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Glynn RJ. Myers et al. Respond to "Understanding Bias Amplification."*Am J Epidemiol*. 2011;174:1228–1229.

6. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Smith GD. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat Med*. 2008;27:1133–1163.

7. Davey Smith G, Ebrahim S. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32:1–22.

8. Zoccali C, Testa A, Spoto B, Tripepi G, Mallamaci F. Mendelian randomization: a new approach to studying epidemiology in ESRD. *Am J Kidney Dis*. 2006;47:332–41.

9. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*. 1986;1:507.

10. Trompet S, Jukema JW, Katan MB, Blauw GJ, Sattar N, Buckley B, Caslake M, Ford I, Shepherd J, Westendorp RGJ, de Craen AJM. Apolipoprotein e genotype, plasma cholesterol, and cancer: a Mendelian randomization study. *Am J Epidemiol*. 2009;170:1415–21.

11. Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol*. 2011;40:755–64.

12. Sheehan N a, Didelez V. Commentary: Can "many weak" instruments ever be "strong"? *Int J Epidemiol*. 2011;40:752–4.

13. Glymour MM, Tchetgen EJT, Robins JM. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol*. 2012;175:332–9.

14. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet*. 2003;361:598–604.

15. Zacho J, Tybjaerg-Hansen A, Jensen JS, Grande P, Sillesen H, Nordestgaard BG. Genetically elevated C-reactive protein and ischemic vascular disease. *New Eng J Med*. 2008;359:1897–908.