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

Marion Verduijn¹, Bob Siegerink¹, Kitty J. Jager², Carmine Zoccali³ and Friedo W. Dekker¹

¹Department of Clinical Epidemiology, Leiden University Medical Center, P.O. Box 9600 2300 RC Leiden, The Netherlands, ²ERA-EDTA Registry, Department of Medical Informatics, Academic Medical Center, University of Amsterdam, P.O. Box 22700 1100 DE Amsterdam, The Netherlands and ³CNR-IBIM Clinical Epidemiology and Pathophysiology of Renal Diseases and Hypertension, Renal and Transplantation Unit, Ospedali Riuniti, 89125 Reggio Calabria, Italy

Correspondence and offprint requests to: Marion Verduijn; E-mail: m.verduijn@lumc.nl

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 randomization enables estimation of causal relationships in observational studies using genetic variants as instrumental variables. An instrumental variable is a variable that can be considered to mimic the coin toss in a randomized study. Given the random assignment of alleles in gamete formation, the use of genetic variants is an alternative method to control for confounding. This educational article describes the approach of Mendelian randomization, its underlying rationale and its necessary assumptions.

Keyword: causality; epidemiology; nephrology

Introduction

While a randomized 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].

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 effect of interventions 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 (e.g. lifestyle 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 [3]. 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 randomized study [4], thus limiting confounding. The use of ‘genetic variants’ as an instrumental variable in observational research is an example hereof. This approach is known as Mendelian randomization [5–7].

This educational article describes the approach of Mendelian randomization used to make causal inference in observational studies, and its underlying rationale and assumptions. All concepts enclosed in single quotation marks are briefly defined in a glossary.

Example 1: Are low serum cholesterol levels a causal risk factor for cancer?

The concept of Mendelian randomization was originally suggested by Katan in 1986 in a debate on the hypothesis that low serum cholesterol levels (exposure) directly increase the risk of cancer (outcome). 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 hypothesized 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 [8].

Rationale

The main rationale in Mendelian randomization 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, this makes a causal role of the genetic product (i.e. the exposure) less likely. The exposure in Mendelian randomization studies can be any gene 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, thus ruling out low cholesterol as a causal risk factor for cancer [9].

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, 1822–84), 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 randomization 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 thus be regarded as the likely sole effect of the treatment that was allocated in a randomized 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.

Framework and assumptions

The framework of a Mendelian randomization study is summarized in Figure 3; Table 1 describes the associations as examined in the studies used as examples in this paper. Mendelian randomization in its most basic form is to study the causal relationship between the exposure levels and the outcome (Association 1) using a genetic variant known for influencing the exposure levels under investigation (Association 2) as an instrumental variable, by estimating the association between the genetic variant and the outcome (Association 3).

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 (Association 3) is expected to be equal to the association between exposure levels and outcome (Association 1), if a genetic variant would explain 100% of the variance in the exposure levels.
In the much more realistic scenario that the genetic variant explains <100% of the variance in the exposure levels, the expected associations can be calculated. 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 is due to residual confounding and/or reverse causation. In this case, the exposure levels as determined by the genetic variant seem to be a risk marker for the outcome, rather than a causal risk factor. We refer to Lawlor et al. for more detailed methods for calculating the expected associations [5].

Figure 4 depicts the three assumptions that are underlying a Mendelian randomization study. The first assumption is that there is a robust association between the genetic variant and the exposure levels. In order to ensure this, the association should be shown in independent (healthy) populations and confirmed in own data. More robust associations between genotype and exposure levels might be obtained by using haplotypes as the instrumental variable instead of a single genetic variant (Figure 4, presence of Association a).

The second and third assumptions are about the multiple—'pleiotropic'—effects of the genetic variant that is used as the instrumental variable. 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 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 verify whether the assumptions are satisfied in a particular Mendelian randomization study. These assumptions are sufficient to test the null hypothesis that the exposure levels 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.

Some additional issues are relevant to mention in the context of Mendelian randomization studies, as they may in part be underlying causes invalidating the assumptions. The first issue is 'population stratification', which 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 [5]. 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 [10], and can be overcome by studying populations that are homogeneous with respect to ethnicity.

A second issue is the possibility that a lifelong genetical (or environmental) exposure has induced developmental compensation via alternative pathways. This is known as 'canalization' and may invalidate the estimation of associa-

### Table 1. Associations as examined in the studies used as examples, given the framework of Mendelian randomization as presented in Figure 3

<table>
<thead>
<tr>
<th>Example</th>
<th>Association</th>
</tr>
</thead>
<tbody>
<tr>
<td>Katan [8]</td>
<td>Cholesterol levels–cancer</td>
</tr>
<tr>
<td></td>
<td>ApoE genotype–cholesterol levels</td>
</tr>
<tr>
<td></td>
<td>ApoE genotype–cancer</td>
</tr>
<tr>
<td>Zacho et al. [11]</td>
<td>CRP levels–ischaemic heart disease</td>
</tr>
<tr>
<td></td>
<td>CRP genotype–CRP levels</td>
</tr>
<tr>
<td></td>
<td>CRP genotype–ischaemic heart disease</td>
</tr>
<tr>
<td></td>
<td>ApoE genotype–cholesterol levels</td>
</tr>
<tr>
<td></td>
<td>ApoE genotype–ischaemic heart disease</td>
</tr>
</tbody>
</table>
tions like those in Figure 3. In this perspective, it is worthy to note that Mendelian randomization 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, thus allowing the possibility of canalization. It is, however, unclear how important this issue is in practice [5].

Finally, as for genetic association studies, relatively large sample sizes are required for Mendelian randomization studies. The sample size highly depends on the frequency of the minor allele of the genetic variant and the expected effect size of the association between the genetic variant 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 heart disease?

The study of Zacho and colleagues [11] shows an interesting application of the Mendelian randomization approach. They aimed to test whether the observed association between the elevated levels of C-reactive protein (CRP) and increased risk of ischaemic heart disease is a causal association. So, the exposure levels under investigation were CRP levels, and the outcome was ischaemic heart 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] 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 Association 1 in the framework of Figure 3. Moreover, the data confirmed that the CRP levels are regulated by the CRP genotype (Association 2): a difference of up to 64% in CRP levels among the genotype combinations. Using these findings, it could be calculated that an increased risk of 1.32 (1.26–1.39) was expected 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 (Association 3). This finding suggests that the observed increased risk of ischaemic heart 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 randomization approach was given, using the ApoE genotype as the instrumental variable for examining the causality of the association between cholesterol levels and ischaemic heart disease [10]. 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.

Conclusions

Mendelian randomization 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 in nephrology using observational data.

Acknowledgements. This work was supported by the applied GENomic stratEgies for Treatment and Prevention of Cardiovascular death in Uraemia and End stage REnal disease (GENECURE) project (www.geneecure.eu), a Specific Targeted Research or Innovation Project, funded by the European Commission under the Sixth Framework Programme as FP6-037696, and by a grant from the Leducq Foundation, Paris, France for the development of Transatlantic Networks of Excellence in Cardiovascular Research (grant 04 CVD 02).

Conflict of interest statement. None declared.
Glossary

Genetic terms

Alleles: variant forms of a gene at a locus; a single allele is inherited from each parent.
Canalization: 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.
Single-nucleotide polymorphism (SNP): a genetic variant in which one specific nucleotide in the DNA is altered.

Epidemiologic terms

Instrumental variable: a variable in non-experimental data that can be considered to mimic the coin toss in a randomized 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 alters the exposure under investigation, rather than vice versa, i.e. cause and consequence are switched.

References


Received for publication: 20.10.09; Accepted in revised form: 8.2.10