# Mendelian randomization: genetic anchors for causal inference in epidemiological studies

George Davey Smith

Gib Hemani

MRC Integrative Epidemiology Unit (IEU) at the University of Bristol, School of Social and Community Medicine, Bristol, UK

Observational epidemiological studies have contributed to identifying important modifiable causes of disease, but also have a poor record of separating causal from non-causal associations. Many examples exist of apparently robust epidemiological associations between behavioural, pharmacological or physiological measures and disease risk which, when subjected to randomised controlled trials, do not deliver the anticipated health benefits. These include many nutritional factors (e.g. several vitamins) pharmacological agents (e.g. hormone replacement therapy) and circulating biomarkers (e.g. HDL cholesterol). In these situations it appears likely that confounding and various biases generate the associations, and that these cannot be adequately accounted for by apparently robust study design and statistical adjustment.[[1]](#endnote-1) [[2]](#endnote-2) The recognition of these problematic aspects of epidemiological investigation have led to the application of a series of methods aimed at improving causal inference in observational epidemiological studies[[3]](#endnote-3) [[4]](#endnote-4), including the use of genetic variants as indicators of exposure that are not subject to the influences that vitiate conventional study designs, an approach known as Mendelian randomization (MR)[[5]](#endnote-5) [[6]](#endnote-6). We will not repeat the many detailed reviews that now exist of MR5 [[7]](#endnote-7) [[8]](#endnote-8) [[9]](#endnote-9) [[10]](#endnote-10) [[11]](#endnote-11) [[12]](#endnote-12) [[13]](#endnote-13) , nor summarise the hundreds of empirical studies applying the technique to a wide range of exposures and disease outcomes, rather, after a brief summary of the foundational principles, we will outline recent developments and potential future directions of the field.

## Mendelian randomization: basic principles

In practice, MR can be used to achieve two slightly different objectives. Its first role is concerned with making unbiased estimates in ordinary least squares regression (OLS). Supposing that the independent variable has an effect on some outcome, but it is also correlated with other (unobserved) variables that too have an effect on the outcome, the estimate of the effect will be biased. MR offers a way to overcome this problem. The second function is perhaps of more fundamental importance in modern biology: making inference about the causal relationship between correlated variables.

Suppose that trait A and trait B are correlated, it follows that if this correlation arises because A is causing B, then any variable that causes trait A should also cause trait B. So the key to inferring a causal relationship between A and B is to identify an ‘instrument’ which is known to cause A. Biologists are in a privileged position in this regard because it is no exaggeration to say that virtually all traits of interest are at least partially influenced by genetic effects, and genetic effects can serve as excellent instruments for a number of reasons. First, and most crucially, it is universally acknowledged that in a genetic association the direction of causation is from the random genetic polymorphism inherited at conception leading to the subsequent trait of interest, and not vice versa. Second, using environmental exposures as instruments for inferring causality (common in conventional observational studies in epidemiology), but this can be problematic because environmental exposures are often associated with a wide range of behavioural, social, and physiological factors that confound associations[[14]](#endnote-14). Functional genetic variants on the other hand can be precise indicators for a particular trait. Third, genetic variants and their effects are subject to relatively little measurement error or bias. Finally, in the era of genome wide association studies (GWAS) and high throughput genomic technologies they are routinely obtainable for huge sample sizes for a vast number of traits.

## Mendelian randomization and instrumental variables analysis

Conventional instrumental variable (IV) analysis requires that the instruments are valid, and in order to be valid they must meet three conditions. Assuming trait A causes B, an instrument

1. must be associated with the causal trait;
2. must be associated with the consequential trait only through the causal trait; and
3. must not be related to unobserved confounders that influence traits A and B.

In MR condition (1) is inherently straightforward to test, but empirically proving (2) and (3) is often problematic. For example, if the variant is functionally pleiotropic, or if it is in linkage disequilibrium (LD) with another variable that causes the consequential trait directly, this can lead to erroneous causal estimation. We discuss ways in which to overcome this issue, but assuming all conditions are met, then the unbiased estimate of the effect of the causal trait on the consequential trait can be made using two-stage least squares (2SLS) regression. In stage 1 the fitted values are obtained from , where is the instrument for A. In stage 2 the regression coefficient is obtained from . thus represents an asymptotically unbiased estimate of the effect of A on B that occurs due to alone, so in order to translate this into the full effect of A on B we calculate , where the potentially biased is obtained from . There exist several implementations of 2SLS in open source software.

## Analogy between Mendelian randomization and randomised controlled trials

14

Include MR and drug target validation

## Examples of Mendelian randomization studies

In table 1 we provide examples of MR studies that cover the range of applications from biomarkers (CRP, LDL cholesterol), physiological measures (BMI, blood pressure), behaviours (alcohol, smoking), and maternal influences (alcohol).

## Mendelian randomization without an instrumental variables estimate of the causal effect

(this is from another paper – Gib, needs rewriting)

Interleukin-6 (IL6) is a second cytokine that has been associated with CHD risk in observational studies[[15]](#endnote-18). IL6 is strongly associated with CRP and with the clotting factor fibrinogen (Figure 3 top panel) and all three biomarkers predict CHD risk (Figure 3, bottom panel). Conventional analyses fail in their ability to separate the effects of these. As shown above, Mendelian randomization studies suggest that CRP is not causally associated with CHD risk, and the same is true for fibrinogen[[16]](#endnote-19) [[17]](#endnote-20) [[18]](#endnote-21). Genetic variants strongly and specifically related to circulating IL-6 levels have not been well characterised, however there is a variant that relates to IL-6 receptor function that can serve as an indicator of IL-6 activity[[19]](#endnote-22) [[20]](#endnote-23). The variant that relates to reduced IL-6 signalling is associated with higher circulating levels of IL-6, but indicates lower activity, and relates quantitatively in the expected way with factors thought to be regulated by IL-6. Furthermore, associations with the variant are concordant with effects produced by tocilizumab, a monoclonal antibody that blocks membrane-bound and circulating IL-6R. The variant associated with lower IL-6 activity is very robustly related to lower risk of CHD[[21]](#endnote-24) [[22]](#endnote-25). Whilst in this situation it is not possible to generate an instrumental variables estimate of the casual effect of higher circulating IL-6 levels on CHD the evidence strongly supports a causal interpretation, and a quantitative analysis of the predicted effects of the variant on CHD with the effect of the variant on IL-6 regulated biomarkers demonstrates that the relative magnitude of effects are as would be predicted with a causal interpretation. However it is possible that the signalling effects of the receptor variant may differ from the effects of differential levels of circulating IL-6 in particular tissues (for discussion see supplementary material in22). Tocilizumab is currently being evaluated as a secondary prevention treatment for CHD.

## Limitations of Mendelian randomization

* Table from 10 years; add (1) if genetic variant causes the disease and the disease influences a marker (e.g. atheroma and BMI influences CRP) then apparent misleading causal effects can be generated if genetic variant is mistaken as being a variant for the exposure (ie an atheroma or obesity variant taken to be one for CRP;[[23]](#endnote-30) (Bowden Stats in Medicine)

## Use of multiple variants to increase power and test assumptions

## Two sample Mendelian randomization

## Bidirectional Mendelian randomization

## Network Mendelian randomization

## Factorial Mendelian randomization

The manner by which causes of disease act together to increase disease risk can have important public health implications, as above additive effects lead to the clustering of risk factors generating a greater burden of disease in the population. For example it has been suggested that the risk of liver disease associated with the combination of obesity and heavy alcohol consumption is greater than multiplicative [[24]](#endnote-31), and adverse trends for these two risk factors would be predicted to generate considerable increases in liver disease. It is difficult to estimate such effects, however, as confounding can be magnified when examining two already confounded risk factors. By analogy with factorial randomised controlled trials, where separate randomisation of different treatments allows characterization of interactions between them[[25]](#endnote-32), Mendelian randomization studies can investigate disease risk associated with combinations of genetic variants and through this obtain unconfounded estimates of the effect of co-occurrence of the two risk factors that the genetic variants are taken to be indicators of.

## Multiphenotype Mendelian randomization

In some situations genetic variants tend to be associated with more than one intermediate phenotype, and estimating the causal effect of one particular intermediate phenotype is rendered problematic. A case in point are categories of blood lipids that associated with coronary heart disease, in particular HDL cholesterol and triglycerides, which are highly inversely correlated and cannot be reliably separated using purely statistical approaches in observational studies[[26]](#endnote-33). The large number of genetic variants related to HDL-C and triglycerides generally associate with both, but to a greater or lesser degrees[[27]](#endnote-34) and thus genetic instruments created with multiple variants do not purely relate to one of the two lipids. In this context regression methods can be applied to attempt to separate the effects; two independent studies utilizing this approach suggested that the causal influence of triglycerides was robust, whereas the apparent protective effect of HDL-C was not[[28]](#endnote-35) [[29]](#endnote-36). Unlike in the situation with factorial Mendelian randomization there is a dependence on attempting to statistically separate effects, which reintroduces problems in conventional observational studies. Both the appropriateness of different statistical approaches and whether reliable answers can be obtained in the multiphenotype context remain areas of active investigation.

## Gene by environment interaction interpreted within the Mendelian randomization framework

## Two step Mendelian randomization

## GWAS and Mendelian randomization

## Hypothesis free Mendelian randomization

## Conclusions

## Acknowledgements

Thanks to Professor Sheila Bird who (in 2002) suggested the term “factorial Mendelian randomization” and to Dr Tom Palmer who suggested the term “multipheotype Mendelian randomization”

**Table 1** - Limitations of Mendelian randomization

|  |  |  |
| --- | --- | --- |
| **Limitation** | **Role in Mendelian randomization (MR) studies** | **Approaches to evaluating or avoiding the limitation** |
| Low statistical power | MR studies are often of low power and effect estimates are imprecise because of this | Increase sample size and/or combine genetic variants so they explain more of the variance of the intermediate phenotype |
| Population stratification | A potential problem but can be avoided by applying standard genetic epidemiology methodologies | Restrict analyses to ethnically homogeneous groups, families and/or apply correction methods using ancestrally informative markers or principal components from genome wide data |
| Reintroduced confounding though pleiotropy | A genetic variant may directly influence more than one post-transcriptional process. Known to be the case for some genetic variants | When possible utilise cis variants with respect to the intermediate phenotype under study, as these may be less likely to have pleiotropic effects. Apply multiple instrument approaches with more than one independent genetic variant as unlikely pleiotropy will generate the same associations for different instruments |
| Linkage disequilibrium (LD) induced confounding | LD is crucial in genetic association studies as it allows marker SNPs to proxy for un-genotyped causal SNPs. However this can reintroduce confounding if LD leads to the association of SNPs related to more than one post-transcriptional process. This case will be similar to the pleiotropy situation | Studies can be carried out in populations with different LD structures. Approaches to avoiding distortion by pleiotropy will also counter problems due to LD |
| Canalization / developmental compensation | During development compensatory processes may be generated that counter the phenotypic perturbation consequent on the genetic variant utilized as an instrument | No general approach developed, although context–specific biological knowledge can be appealed to. Period of lifecourse when influence of genetic variation on IPs emerge can indicate whether canalization could, in principle, be an issue |
| Lack of genetic variants to proxy for modifiable exposure of interest | No reliable genetic variant associations for many intermediate phenotypes of interest, although an increasing number of these now identified | Continued genome wide and sequencing based studies |
| Complexity of associations | Without adequate biological knowledge misleading inferences regarding intermediate phenotypes and disease may be drawn | Increased biological understanding of genotype – phenotype links |

1. Davey Smith G, Ebrahim S. Epidemiology - is it time to call it a day? Int J Epidemiol 2001;30:1-11. [↑](#endnote-ref-1)
2. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiological studies; a simulation study. Am J Epidemiol 2007;166:646-655. [↑](#endnote-ref-2)
3. Lipsitch M, Tchetgen ET, Cohen T. [Negative Control Exposures in Epidemiologic Studies](http://journals.lww.com/epidem/Citation/2012/03000/Negative_Control_Exposures_in_Epidemiologic.28.aspx" \o "Negative Control Exposures in Epidemiologic Studies). Epidemiology. 2012;23:351-352 [↑](#endnote-ref-3)
4. Davey Smith G. Assessing intrauterine influences on offspring health outcomes: can epidemiological findings yield robust results? Basic and Clinical Pharmacology and Toxicology 2008;102:245-256. [↑](#endnote-ref-4)
5. Davey Smith G, Ebrahim S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiology 2003;32:1-22. [↑](#endnote-ref-5)
6. Podolsky SH, Davey Smith G. Park’s story and Winters’ tale: alternate allocation clinical trials in turn of the Century America. JLL Bulletin: Commentaries on the history of treatment evaluation. Journal of Royal Society of Medicine 2011; 104:262-8 (updated version on http://www.jameslindlibrary.org/illustrating/articles/parks-story-and-winters-tale-alternate-allocation-clinical-tr). [↑](#endnote-ref-6)
7. Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol 2004;33:30-42. [↑](#endnote-ref-7)
8. Sheehan NA, Didelez V, Burton PR, Tobin MD. Mendelian Randomisation and Causal Inference in Observational Epidemiology. PLoS Med 2008;5: e177 [↑](#endnote-ref-8)
9. Lawlor DA, Harbord RM, Sterne JAC, Timpson NJ, Davey Smith G. Mendelian randomization : Using genes as instruments for making causal inferences in epidemiology. Statistics in Medicine 2008; 27:1133-1163 [↑](#endnote-ref-9)
10. Bochud M, Rousson V. Usefulness of Mendelian Randomization in Observational Epidemiology. International Journal of Environmental Research and Public Health. 2010; 7(3):711-728. [↑](#endnote-ref-10)
11. VanderWeele TJ.; Tchetgen Tchetgen, EJ.; Cornelis, M. [Methodological Challenges in Mendelian Randomization](http://journals.lww.com/epidem/Abstract/2014/05000/Methodological_Challenges_in_Mendelian.14.aspx" \o "Methodological Challenges in Mendelian Randomization). Epidemiology. 2014;25:427-435 [↑](#endnote-ref-11)
12. Richmond RC, Davey Smith G, Ness AR, den Hoed M, McMahon G, Timpson NJ. Assessing causality in the association between child adiposity and physical activity levels: A Mendelian randomization analysis. PLoS Medicine 2014; doi:10.1371/journal.pmed.1001618 [↑](#endnote-ref-12)
13. Davey Smith G. Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. Genes & Nutrition 2011;6:27–43. [↑](#endnote-ref-13)
14. Davey Smith G, Lawlor DA, Harbord R, Timpson NJ, Day I, Ebrahim S. Clustered Environments and Randomized Genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Medicine 2007;4:1985-1992. PMCID: PMC2121108 [↑](#endnote-ref-14)
15. IL6R Genetics Consortium and Emerging Risk Factors Collaboration. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. [Lancet.](http://www.ncbi.nlm.nih.gov/pubmed/22421339" \o "Lancet.) 2012 Mar 31;379(9822):1205-13. [↑](#endnote-ref-18)
16. Davey Smith G, Harbord R, Milton J, Ebrahim S, Sterne JAC. Does Elevated Plasma Fibrinogen Increase the Risk of Coronary Heart Disease?: Evidence from a Meta-Analysis of Genetic Association Studies. Arterioscler Thromb Vasc Biol 2005; 25: 2228-2233. [↑](#endnote-ref-19)
17. [Sabater-Lleal M](http://www.ncbi.nlm.nih.gov/pubmed?term=Sabater-Lleal%20M%5BAuthor%5D&cauthor=true&cauthor_uid=23969696), [Huang J](http://www.ncbi.nlm.nih.gov/pubmed?term=Huang%20J%5BAuthor%5D&cauthor=true&cauthor_uid=23969696), [Chasman D](http://www.ncbi.nlm.nih.gov/pubmed?term=Chasman%20D%5BAuthor%5D&cauthor=true&cauthor_uid=23969696), [Naitza S](http://www.ncbi.nlm.nih.gov/pubmed?term=Naitza%20S%5BAuthor%5D&cauthor=true&cauthor_uid=23969696), [Dehghan A](http://www.ncbi.nlm.nih.gov/pubmed?term=Dehghan%20A%5BAuthor%5D&cauthor=true&cauthor_uid=23969696) et al. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. [Circulation.](http://www.ncbi.nlm.nih.gov/pubmed/23969696" \o "Circulation.) 2013 128:1310-24 [↑](#endnote-ref-20)
18. Ames BN. Cancer prevention and diet: Help from single nucleotide polymorphisms. Proceedings of the National Academy of Science 1999; 96:12216-12218. [↑](#endnote-ref-21)
19. IL6R Genetics Consortium and Emerging Risk Factors Collaboration. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. [Lancet.](http://www.ncbi.nlm.nih.gov/pubmed/22421339" \o "Lancet.) 2012 Mar 31;379(9822):1205-13. [↑](#endnote-ref-22)
20. The Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. The Lancet, [379](http://www.thelancet.com/journals/lancet/issue/vol379no9822/PIIS0140-6736(12)X6013-6): 1214 – 1224. [↑](#endnote-ref-23)
21. IL6R Genetics Consortium and Emerging Risk Factors Collaboration. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. [Lancet.](http://www.ncbi.nlm.nih.gov/pubmed/22421339" \o "Lancet.) 2012 Mar 31;379(9822):1205-13. [↑](#endnote-ref-24)
22. The Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. The Lancet, [379](http://www.thelancet.com/journals/lancet/issue/vol379no9822/PIIS0140-6736(12)X6013-6): 1214 – 1224. [↑](#endnote-ref-25)
23. Bowden J, Vansteelandt S. [Mendelian randomization analysis of case-control data using structural mean models](http://onlinelibrary.wiley.com/doi/10.1002/sim.4138/abstract). STATISTICS IN MEDICINE 2011;30:678-694 [↑](#endnote-ref-30)
24. Hart C, Morrison DS, Batty GD, Mitchell RJ, Davey Smith G. Effect of body mass index and alcohol consumption on liver disease: analysis of data from two prospective cohort studies. BMJ 2010;34:c1240. PMCID: PMC2837144 [↑](#endnote-ref-31)
25. Montgomery 2003 [↑](#endnote-ref-32)
26. Phillips A, Davey Smith G. How independent are "independent" effects? Relative risk estimation when correlated exposures are measured imprecisely. J Clin Epidemiol 1991;44:1223-31. [↑](#endnote-ref-33)
27. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nat. Genet.2013;45:1274–1283 [↑](#endnote-ref-34)
28. Do R, Willer CJ, Schmidy EM et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. Nature Genetics 2013;45, 1345-1352 doi:10.1038/ng.2795 [↑](#endnote-ref-35)
29. Holmes MV, Asselbergs FW, Palmer TM, et al. Mendelian randomization of blood lipids for coronary heart disease. Eur Heart J 2014, doi:10.1093/eurheartj/eht571 [↑](#endnote-ref-36)