Mendelian randomization at 20: how can it avoid hubris, whilst achieving more?

Professor George Davey Smith, MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK

Professor Shah Ebrahim, London School of Hygiene and Tropical Medicine, Keppel Street, London, UK

Corresponding author:

Professor George Davey Smith

Bristol Medical School

University of Bristol

Bristol

UK

Email pa-ieudirector@bristol.ac.uk

Mendelian randomisation, the initial extended formulation of which appeared 20 years ago,[1] uses germline genetic variation to strengthen causal inference regarding modifiable risk factors for disease. In the original formulation, a genetic variant of known function, often a single-nucleotide polymorphism (SNP), was taken to proxy for the exposure. Mendelian randomisation was introduced at a time when it was becoming clear that conventional observational epidemiological studies could produce misleading findings, with randomised controlled trials of the exposures they identified as likely causes of disease yielding null results.[1, 2] Confounding and reverse causation (in which disease processes influence the apparent exposure, rather than vice versa) were able to produce statistically robust and replicable findings, that were nevertheless spurious. Proxying an exposure through a SNP transmitted from a parent to their offspring incorporates variation that is fixed at conception and cannot be influenced by reverse causation. Furthermore, conventional confounding of SNPs from behavioural and socioeconomic factors is, in most settings, unlikely.[3]

In within-family studies, the laws of mendelian genetics imply that parent-to-offspring transmission of genetic variants is random, and thus they are unrelated to potential confounders and to other genetic variants, except those in close physical proximity to the variant in question on the chromosome it is on.[2, 3] The original exposition of mendelian randomisation[1] presented data on a SNP in the gene methylenetetrahydrofolate reductase (*MTHFR*) that relates to lower folate levels and to higher homocysteine, to the degree expected from randomised trials of folate supplementation. A meta-analysis of case-control studies of congenital neural tube defects with the *MTHFR* SNP data available for mothers, fathers, and the child found that if the mothers carried the variant related to lower folate and higher homocysteine then their offspring had a two-fold higher risk of being born with a neural tube defect. However, there was no association of the SNP in the fathers and neural tube defects among offspring. This provided evidence (congruent with randomised trial results) that better folate status in mothers substantially reduced risk of neural tube defects, presumably through an antenatal influence. The risk associated with the SNP carried by the offspring themselves was lower than that with the SNP in the mothers, but above the null association with the SNP in the fathers, as anticipated for a variant acting through the maternally provided intrauterine environment.

There were no large-scale genotyped family-based studies available in 2003 that would have allowed a true—ie, within-family—mendelian randomisation approach to how exposures in middle-aged adults influence their risk of disease.[1] Approximate mendelian randomisation, in which population rather than family data were used, can be biased by population stratification,[1, 3] but genetic variants will probably give less biased assessment than that obtained from measured exposures. Mendelian randomisation analyses confirmed a causal effect of LDL cholesterol on coronary heart disease,[1] but estimates centred on the null for the effects of fibrinogen[4] and C-reactive protein on coronary heart disease.[5] Mendelian randomisation findings suggesting that developing pharmaceutical agents for some potential treatment targets—such as C-reactive protein—would not yield clinical benefit led to cessation of attempts to develop such agents.

To interpret mendelian randomisation findings as indicating expected effects of interventions requires the gene–environment equivalence assumption,[3] such that modification of the exposure by the intervention and modification by genetic variation transmitted from parents would produce the same effect on the outcome. In the case of LDL cholesterol and coronary heart disease, it is clear that genetic variation produces differences in circulating levels from birth onwards;[6] thus, the arterial walls are exposed across the lifetime. Randomised trials of LDL-cholesterol-lowering drugs versus placebo generally continued for around 5 years. There is clear equivalence of mechanism as many LDL-cholesterol-lowering agents (eg, PCSK9 inhibitors, statins) target genes (eg, *PCSK9, HMGCR*), variation in which associates with LDL cholesterol concentrations. The difference in exposure duration translates into roughly 40% of the lifetime effect being seen within the 5-year trial period, with early mendelian randomisation[1]

closely matching contemporary comparisons using the vastly greater amount of genetic and trial data now available.[7]

The relative simplicity of mendelian randomisation when introduced[1] has been transformed by the availability of massive genome-wide association study (GWAS) resources that could not have been envisaged in 2003. This has provided many opportunities for the rapid production and publication of mendelian randomisation papers produced from publicly available summary data.[3] The ease of this production has led to the domain of mendelian randomisation being shrunk, from representing a broad range of approaches for using the special properties of germline genetic variation to infer how one phenotype influences another phenotype[1] to simply being instrumental variable effect estimation. The exponential explosion of mendelian randomisation papers (figure) contains many that are in our view highly implausible, claiming to be able to identify the effects of exposures that could not realistically be proxied by genetic variation. Around the same time that the rate of appearance of summary data mendelian randomisation studies overtook that of individual data-based studies, the risk of serious errors being introduced was recognised,[8] leading to the retraction of an erroneous paper.[9] Many such flawed papers are now appearing, the vast majority of which will go uncorrected. The advent of artificial-intelligence-assisted writing of papers will probably contribute to this avalanche of spurious papers. Already the number of submissions of two-sample mendelian randomisation papers are overwhelming journals, but the simple requirement for authors to comply with STROBE–mendelian randomisation reporting would reduce submission number and improve the average quality of papers considered for review.[10]

A second explosion has been in complex mendelian randomisation methods that are difficult for most readers and co-authors to scrutinise. These methods require assumptions beyond those of conventional instrumental variable analyses,[3] and even when these are violated to the extent that they produce impossible findings they can go unnoticed and are published.[11] Indeed, this was the case for the paper that has been retracted and republished in this issue of *The Lancet Diabetes & Endocrinology*.[12] The original paper suggested that non-linear mendelian randomisation predicted a beneficial effect of supplementary vitamin D for those with below-average vitamin D levels. Mendelian randomisation here simply recapitulated confounded observational findings, that readers might have wanted to believe. If, like many previous mendelian randomisation studies, it had suggested no benefit of vitamin D it is unlikely to have received the attention it did. The aforementioned impossible findings[11] are now accepted by the authors to be erroneous[12] and the reasons why discussed, with a modification of the non-linear mendelian randomisation method developed and applied; the reanalysis suggests no non-linearity and no overall effect.[12] Unfortunately, many papers have been published using the same approach as the retracted paper, which are probably equally misleading in terms of the results they present,[13] but are protected by not having reported such obviously spurious findings as the retracted paper. Considerable circumspection should be applied to interpretation of all published non-linear mendelian randomisation papers, and further stress-testing of the new non-linear mendelian randomisation method is required before its findings are accepted.[13]

We believe that the original notion of mendelian randomisation as exploiting genetic variation to understand phenotypic causation has much to offer, but it should not be constrained to instrumental variable estimation and should always be situated within a triangulation of evidence framework.[14] The power of triangulation—ie, of combining different causal inference approaches applied to the same question—would have been shown through combining non-linear mendelian randomisation with a negative control approach. Doing so results in the non-sensical conclusion that vitamin D influences biological sex and age, making the already evident problems with the published analyses yet clearer.[13]

To conclude on an optimistic note, as long as considered within a triangulation framework[14] mendelian randomisation remains an advance over naive observational studies, that were so often misleading with regard to disease prevention in the past.[1, 2] The optimistic scenario is underpinned by realising there is more to mendelian randomisation than simply being mechanical instrumental variable estimation, including: first, expansion of family-based mendelian randomisation; second, development of methods that allow estimation of treatment effects rather than disease prevention, the latter being the focus of almost all current mendelian randomisation studies that use summary data on disease occurrence as their outcome; third, application of mendelian randomisation to understand evolutionary processes and within non-human organism contexts; fourth, use of mendelian randomisation principles to identify novel environmental causes of disease; and finally, development of methods that allow qualitative evidence from mendelian randomisation approaches to be included in evidence synthesis. Developing these approaches would produce genuine advances in understanding. Conversely, the publication of two-sample mendelian randomisation studies based on available data without triangulation, and implementing methods that have not been subjected to appropriate scrutiny and produce spurious but headline-attracting findings is not positively contributing to scientific knowledge.

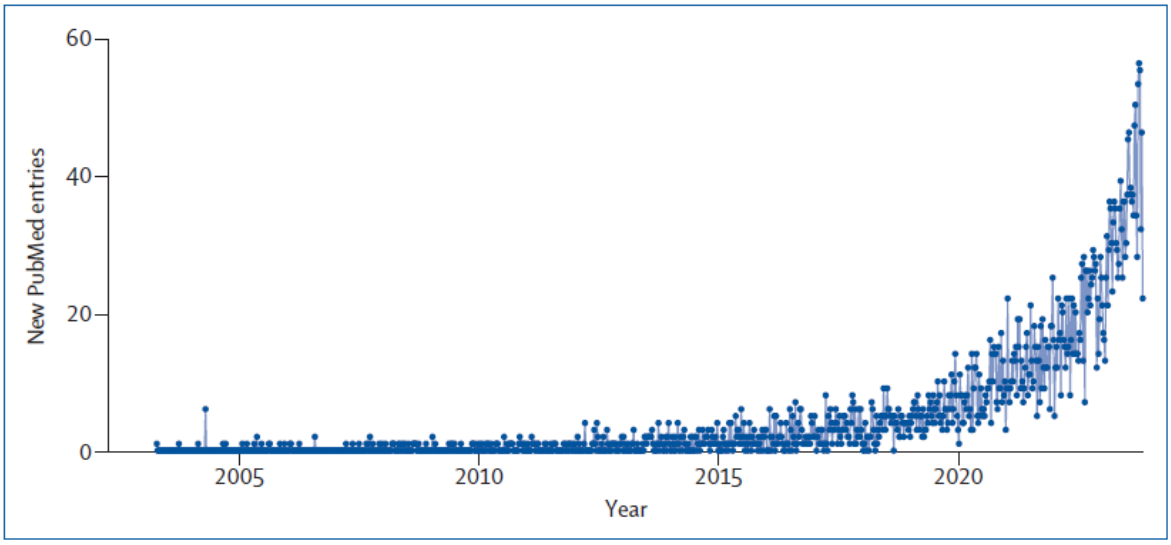

Figure: PubMed search for "Mendelian randomisation" or "Mendelian randomization" in the title of papers reported per week
Data as of Nov 1, 2023.

1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32:1-22.
2. Davey Smith G, Holmes MV, Davies NM, Ebrahim S. Mendel's laws, Mendelian randomization and causal inference in observational data: substantive and nomenclatural issues. Eur J Epidemiol. 2020;35:99-111.
3. Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafò MR, et al. Mendelian randomization. Nature Reviews Methods Primers. 2022;2:6.
4. Davey Smith G, Harbord R, Milton J, Ebrahim S, Sterne JA. 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;25:2228-33.

5. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC), Wensley F, Gao P, et al. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ. 2011;342:d548.
6. Latsuzbaia A, Jaddoe VWV, Hofman A, Franco OH, Felix JF. Associations of genetic variants for adult lipid levels with lipid levels in children. The Generation R Study. J Lipid Res. 2016;57:2185-92.
7. Holmes MV, Davey Smith G. Dyslipidaemia: Revealing the effect of CETP inhibition in cardiovascular disease. Nature Reviews Cardiology. 2017;14:635-6.
8. Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. International Journal of Epidemiology. 2017;45:1717-26.
9. Inoshita M, Numata S, Tajima A, Kinoshita M, Umehara H, Nakataki M, et al. Retraction Note: A significant causal association between C-reactive protein levels and schizophrenia. Scientific Reports. 2018;8:46947.
10. Kjaergaard AD, Davey Smith G, Stewart P. Mendelian Randomization Studies in Endocrinology: Raising the Quality Bar for Submissions and Publications in The Journal of Clinical Endocrinology & Metabolism. J Clin Endocrinol Metab 2023; published online Oct 5 https://doiorg/101210/clinem/dgad569.
11. Davey Smith G. Mendelian randomisation and vitamin D: the importance of model assumptions. Lancet Diabetes Endocrinol. 2023;11:14.
12. Emerging Risk Factors Collaboration/EPIC-CVD/Vitamin D Studies Collaboration. Estimating dose-response relationships for vitamin D with coronary heart disease, stroke, and all-cause mortality: observational and Mendelian randomisation analyses. Lancet Diabetes Endocrinol. 2023 published online Dec 1. https://doi.org/10.1016/S2213-8587(23)00287-5.;12.
13. Hamilton F, Hughes DA, Spiller W, Tilling K, Davey Smith G. Non-linear mendelian randomization: evaluation of biases using negative controls with a focus on BMI and Vitamin D. medRxiv. 2023 published online Aug 23 https://doi.org/10.1101/2023.08.21.23293658 (preprint).
14. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. International Journal of Epidemiology. 2017;45:1866-86.