Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org

Mendelian Randomization

Ewan Birney

Deputy Director General of the European Molecular Biology Laboratory and Director of EMBL's European Bioinformatics Institute (EMBL-EBI), Cambridge CB10 1SD, United Kingdom

Correspondence: birney@ebi.ac.uk

Mendelian randomization borrows statistical techniques from economics to allow researchers to analyze the effects of the environment, drug treatments, and other factors on human biology and disease. Taking advantage of the fact that genetic variation is randomized among children from the same parents, it allows genetic variants known to influence factors like alcohol consumption or low-density lipoprotein (LDL) levels to be used as instrumental variables that can disentangle the effects of these factors on outcomes such as pregnancy or cardiovascular disease, respectively. There are caveats to analyses using Mendelian randomization and related techniques that researchers should be aware of, but they are increasingly powerful tools for solving problems in epidemiology and human biology.

Understanding human biology has fascinated scientists since—or perhaps before—the renaissance, and this understanding is a bedrock of knowledge for the practice of medicine. However, the scientific method of creating and recording the outcomes of carefully controlled perturbations is impossible in a whole human setting, with the exception of the highly regulated and specific randomized controlled trials used most widely in medical therapeutics. Much of our understanding of humans, in particular as whole organisms, therefore, comes from observational studies (i.e., the long-standing disciplines of epidemiology, physiology, and the complex behavioral worlds of psychology and sociology), where one can observe many things, but cannot directly influence the actions or settings of a human in a controlled manner.

A combination of key insights from epidemiologists, bringing in new tools and theories from genetics and borrowing and extending statistical theory from economics, has provided a new powerful method to understand humans: Mendelian randomization. The core of this method relies on two key features. The first is the simple observation that all individuals in a natural population have a series of unique genetic variation, and some of this genetic variation perturbs the biology of the individuals. The second is that this genetic variation is nearly perfectly randomized when present on different chromosomes between children from the same parents. This randomization extends to the majority of variation due to recombination within chromosomes, and, with some care in the modeling, the randomization is also a large compo-

Editors: George Davey Smith, Rebecca Richmond, and Jean-Baptiste Pingault
Additional Perspectives on Combining Human Genetics and Causal Inference to Understand Human Disease and Development available at www.perspectivesinmedicine.org

 doi: 10.1101/cshperspect.a041302
Cite this article as *Cold Spring Harb Perspect Med* 2022;12:a041302

nent of the large variation in genetic differences in individuals in a population. As such, each individual can be considered part of a massive simultaneous natural experiment, with at least tens of thousands of different perturbations being randomized compared to their siblings and, largely, other individuals in a population. Borrowing methods and terminology from econometrics, which explored the statistical use of natural experiments in their field, the genetic variables can be considered a type of "instrumental variable," which is known to be reliably associated with an explanatory variable (usually some part of human biology) via a causal pathway but is unrelated to other confounding variables that might change dependent variables. By tracing the impact of the instrumental variable on the explanatory variable to a dependent variable, one can infer the usually otherwise complex relationship between the explanatory variable and dependent variable. Although the statistical methods are borrowed from econometrics, the analogy with clinical trials is probably easier for biologists and clinicians to digest. Here the genetic perturbations can be considered equivalent to a (usually weak) dose of a drug, and the randomization between siblings or individuals in a population the equivalent of the assignment to the drug or placebo arm.

To illustrate with real examples: some individuals carry genetic variants that change the consumption of alcohol due to metabolic processing pathways. When these variants are used as instrumental variables in understanding the impact of alcohol consumption on pregnancy, they show that alcohol consumption carries some risk of harm on the later life of the child. This is at odds with a "weak protective effect" from observational studies of moderate alcohol intake, and the discrepancy can be traced to differential reporting of alcohol drinking during pregnancy between socioeconomic classes. Another example is the numerous genetic variants that impact the circulating levels of low-density lipoprotein (LDL); in nearly all cases the genetic variant that lowers LDL levels also lowers the risk of heart attacks, and the relationship between the impact on LDL levels and on heart attack risk is reasonably linear. This linear relationship is consistent with the outcome of clinical trials on therapeutic changes to LDL levels by different drugs and drug dosages. In contrast, genetic variants that change the levels of high-density lipoprotein (HDL) (in the best case without impacting other lipid levels) have no effect on heart attack risk, a fact that a number of very expensive (in total more than a billion U.S. dollars) clinical trials recapitulated with proposed therapeutics.

Mendelian randomization is a deceptively simple and yet powerful framework. It uses genetics to probe aspects of human biology, which are otherwise some tangled Gordian knot of correlations. Most of these correlations are between the attributes of humans, which clearly have plenty of nongenetic causal components running through the same pathways that the genetics can probe. Importantly, the genetic effects do not have to be large (which is fortunate; many genetic effects on complex human physiology are not large). Furthermore, modern Mendelian randomization techniques have leveraged the likely presence of multiple genetic variants (such as those in the LDL example) for a particular hypothesis, making the process more robust. However, like all statistical techniques, Mendelian randomization has many pitfalls and complications, mostly whether the assumptions the epidemiologist or statistician makes by applying these techniques hold. There are assumptions about the action of the explanatory variable and the route of how the genetics works; there are assumptions of the relationships between individuals in a population; there are assumptions of the statistical independence of the genetic variables. More important than these assumptions is the experience of where this technique is best suited, and which questions it can more easily answer now, and perhaps which types of questions will be more tractable in the future.

Looking ahead, it is likely that Mendelian randomization and other genetically informed approaches will continue to grow in both utility and popularity in human biology. The growth in utility is because of the increasing number of physiological phenotypes for which there are genetic instruments. These physiological phenotypes will provide a richer description of potential causal mechanisms, which, as long as the practitioner is careful of his or her statistics

and multiple testing, should give rise to more real results. The growth in popularity is because it is one of the few ways to convert observational studies into more supportive evidence for causality. Broadly, this growth in usage will be a good thing but it comes with a number of risks as well as opportunities.

The risks will be around its broader use, meaning that it is likely that many people will apply the methods with less appreciation to assumptions or caveats of the analysis. In my own work I often apply Mendelian randomization, but I have learned the hard way that it does not always work; it does not work sometimes simply because one's starting hypothesis is wrong (in which case, this is more like a negative result, but the method itself is working). It does not work because there are simply no good instruments—a paucity of variation that impacts the exposure one wishes to study; or it does not work because the assumptions of independence or the causal pathway is inappropriate for the test. The latter case is more common when the "exposure" is a molecular phenotype (such as RNA-seq or methylation), and often the variation is nearby in the genome with complex local pleiotropy effects confounded by variants in linkage disequilibrium. Coupled with the risks is the constant habit for all scientists to often "want" a particular result, as the Richard Feynman quote suggests "The first principle is that you must not fool yourself—and you are the easiest person to fool."

Mitigating these risks requires a mixture of good scientific practice and a certain amount of experience. Good scientific practice includes preparing data sets carefully and independently of analysis, knowing the hypotheses you will test, and being comfortable with the "wrong" (often null) result. The most explicit way to show that you follow this best practice is to have a published analysis plan up front, which one can always deviate from for good reason, but at least you have to be honest with yourself (and your reader) on the deviations. There are, for example, some often-repeated processing errors and more subtle aspects of which covariates to use and when. Here, experience—with the battle scars of apparently interesting results that have dissolved on further probing—is invaluable. If at all possible for a newcomer into this field, I would recommend searching out someone with experience to discuss analysis plans and show intermediate results.

The opportunity though is impressive. Human health and disease are fundamentally about normal or dysfunctional physiological processes, and Mendelian randomization allows one to start to untangle the mass of correlation present in all biological systems, humans being no exception. There is nothing mystical about only using disease as outcomes; all processes can be studied, and this brings in the more complex worlds of behavior, sociology, and even economics and policy. I am both excited about this stretch "beyond health" and also wary of the complexity. As one moves into outcomes that are societal—from aspects of mental health through educational performance to behavior—one needs to both bring in the disciplines that understand these areas, and explain the technique while learning the new field. A particular complexity is explaining that the genetic instruments do not have to be that strongly predictive of the exposure; most other fields think of genetics solely in a predictive manner. In addition, I will be concerned that many of the techniques used in genetics to create a pseudorandom modeling of the environment (for example, genetic principal components in GWAS) will not work for more complex societal and behavioral exposures or outcomes; we live in very stratified societies, and recent genetic ancestry is often correlated with this stratification in both obvious and nonobvious ways. Here the extension of the Mendelian randomization technique to within-family studies will be particularly useful.

These opportunities and risks though should not in any way deter the interested practitioner from learning Mendelian randomization and applying it to appropriate problems in human epidemiology and physiology. There is plenty of low hanging fruit—and more coming—in understanding human biology, and I look forward to an increasingly broad community making discoveries with this technique.