Causal inference with genetic data: past, present and future.

Jean-Baptiste Pingault[1]

Rebecca Richmond[2]

George Davey Smith[2]

[1] Division of Psychology & Language Sciences, University College London, London, UK

[2] MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

Identifying causal risk and protective factors for human disease and development is a critical endeavour across social and biomedical sciences. Examples of causal questions that can be interrogated using the methods discussed in this collection include the following: Does Vitamin D protect against multiple sclerosis[1]? Does Selenium supplementation protect against cancer[2]? Do elevated inflammation biomarkers such as C-Reactive Protein increase the risk of depression[3]? Do high levels of circulating testosterone increase bone mineral density and decrease body fat[4]? Does maternal smoking during pregnancy lower birth weight or increase the risk of child Attention-Deficit Hyperactivity Disorder[5]? Does higher education worsen myopia[6]? Does victimisation worsen adolescent mental health[7]? Does a tuberculosis infection increase the risk of lung adenocarcinoma[8]? Or, as an example of particular interest at the time of writing, can inflammatory biomarkers such as IL6 be targeted to decrease the risk of severe outcomes following SARs-CoV-2 infection[9]?

These questions only begin to capture the wide array of modifiable risk factors that can be investigated, including dietary supplements, biomarkers, lifestyles, social environments, or infections. More elaborate causal questions can also be asked by jointly modelling several risk factors. For example, is high-density lipoprotein cholesterol (HDL-C) really 'good cholesterol' (i.e. does it independently protect against CHD, even after accounting for the effect of other lipoproteins)[10]? Or, what is the role of epigenetic markers (such as DNA methylation) in mediating the effect of obesity on cardiometabolic diseases[11]?

Answering those questions not only provides insights into the aetiology of human disease and development but can also directly inform interventions. Conversely, inaccurate answers can lead to costly research dead ends and adverse public health consequences, such as the widespread consumption of inefficient and potentially iatrogenic supplements. Causal inference methods can be implemented to help answer those questions, by first providing evidence for or against the existence of a causal effect and by identifying its direction and estimating its magnitude.

In this collection, we focus on a subset of such methods that use genetic data to strengthen causal inference in observational studies. We note that philosophers and scientists have grappled with the notion and definition of cause over centuries. We do not aim to contribute to this debate. Instead, in this collection, we focus on causal inference methods as tools to identify modifiable factors that, when changed, should lead to a change in the outcome of interest.

This collection has emerged from the convergence of two scientific fields – genetics and causal inference. We will discuss each in turn before demonstrating how they have converged to feed into genetically informed causal inference methods.

From Mendelian to molecular genes

Modern genetics can be traced back to Gregor Mendel's experimental work, published in 1866, and the suggestion that discrete entities randomly transmitted across generations can explain the inheritance of discrete phenotypic (i.e. nongenetic) features, such as the colour of peas[12]. The focus on complex continuous traits such as human height came with the Biometricians, Francis Galton[13] and Karl Pearson[14], from the latter decades of the 19th to the start of the 20th century. Rather than from breeding experiments, within the context of human traits, biometricians used observed trait associations between family members to derive the role of genetic influences. Genetic influences were quantified using the concept of correlation introduced by Galton and formalized by Pearson.

The apparent paradox of discrete entities – the genes – having to account not only for discrete characteristics but also continuous traits initially divided Mendel's disciples and biometricians. Several contributions early in the 20th century contributed to the resolution of this debate (see [15,16]). In 1911, Brownlee stated explicitly that "there is nothing necessarily antagonistic between the evidence advanced by the biometricians and the Mendelian theory". He showed that discrete elements consistent with the Mendelian theory could result in a normal distribution, concluding: "If the inheritance of stature depends upon a Mendelian mechanism, then the distribution of the population as regards height will be that which is actually found, namely, a distribution closely represented by the normal curve"[17]. In 1918, Ronald Fisher proposed an extended model including environmental effects in addition to many discrete genetic variants, or "cumulative Mendelian factors"[18]. Fisher's extended model showed how the resulting phenotypic variance of complex traits could be partitioned into genetic and nongenetic components, laying the foundations of the field of quantitative genetics. Quantitative genetics has developed considerably since by relying on the known genetic relatedness between relatives (e.g. twins) to better understand the respective importance of genetic and environmental factors and their interplay in explaining individual differences. Such aetiological studies partition the variance of a single trait or the covariance between traits into genetic, shared and non-shared environmental components (the non-shared component conflates external environmental influences but also measurement error and a likely substantial component due to variation explained by the intrinsic stochasticity of molecular processes[19–22]).

Following the discovery of DNA as the substrate for genes, followed by the uncovering of the structure of DNA, the notion of 'genes' has evolved from 'Mendelian genes' – abstract statistical entities explaining inheritance – to incarnate sequences of DNA or 'molecular genes'. These discoveries sparked the quest to identify molecular genes associated with diseases and traits. Among methods used to this end, Genome-Wide Association Studies (GWAS) have played a key role in the past 15 years. GWAS test the association of millions of genetic variants, typically Single Nucleotide Polymorphisms (SNPs), with a given trait. A vast array of downstream analyses can then be implemented to identify the genetic variants causing the disease. Analyses to identify causal variants in humans have typically relied on statistical methods, for example testing which genetic variant stays the most associated to the trait after accounting for neighbouring genetic variants (so-called “conditional analysis”)[23]. The advent of genome editing methods so precise that they can edit out and/or replace a few genetic variants in vivo[24] offers exciting opportunities to identify causal variants experimentally. The downstream effects of changing a given variant can be directly tested, consistent with the aforementioned notion that a risk factor is causal for an outcome when changing the risk factor also changes the outcome. Recent methods in the area hold considerable promise in uncovering the true causal variants and genes leading to diseases[25].

Although the focus of the molecular era has mainly been to identify causal genetic variants, it has profoundly changed quantitative genetics and our ability to study the genetic architecture of traits. Genetic relatedness between distantly related individuals can also be calculated based on genome-wide markers. Similarly to family-based studies, this can then be exploited to derive the role of genetics in the variance and covariance of traits. For example, SNP-heritability is the proportion of variance in a trait explained by the additive effects of all measured SNPs[26]. Based on those measured SNPs, additional methods can be implemented to estimate the genetic correlation – e.g. how correlated are the genetic factors underlying schizophrenia and bipolar disorder[27]? In turn, models based on genetic correlation matrices for many traits can help us to better understand the genetic architecture of families of traits (e.g. how psychiatric traits cluster into subsets that are closely genetically related)[28,29]. Such advances that jointly model all measured SNPs largely mimic what was possible with family-based studies, i.e. estimating population-level statistics like heritabilities or genetic correlations. However, a decisive advantage of the molecular era is that the information available at the individual level is considerably richer. Instead of knowing the place of an individual in a particular pedigree (e.g. as a member of a twin pair) we have access to millions of genetic variants for that individual. The cumulative effect of genetic variants can be thus captured by a polygenic score for any given trait, i.e. an individual-level score computed by summing risk variants weighted by effect sizes derived from GWAS[30,31]. Polygenic scores can be computed based on genome-wide data or a subset of variants (e.g. genome-wide significant variants). Current polygenic scores based on GWAS for height and education predict 24% and 11% of the variance in their respective phenotype[32,33]. Such scores can then be used in (multivariate) models to examine genetic influences on an array of traits[34,35]. As individual-level variables, they can improve predictive models of disease (e.g. cardiovascular disease[36]) and may lead to clinical applications[37,38].

Of note is that, despite often being labelled ‘aetiological’, investigations decomposing the variance of traits into genetic and environmental components have little to do with the identification of causal risk factors. This is because, in any particular study, the variance explained by genetics may largely depend on the distribution of environmental factors; conversely, the variance explained by environmental factors may depend on allele frequency in the study population. The decomposition of variance is therefore local to a study population and can be entirely different from the true role of genetics and the environment in explaining trait variation for a given trait. Only with an assumption of strict additivity – i.e. genetic effects are the same across all environments and vice versa – can study estimates reflect the respective aetiological role of genetic and environmental influences. As Lewontin put it: “In view of the terrible mischief that has been done by confusing spatiotemporally local analysis of variance with the global analysis of cause, I suggest that we stop the endless search for better methods of estimating useless quantities[39].” As noted later, however, ‘the local objection’ is not really an objection regarding causal interpretation, but rather regarding “the generalisability of particular research findings”, which is not only a problem for the analysis of variance but for any causal analysis[40]. More fundamental, however, is the realization that high heritability estimates may simply reflect a restricted range of observed environmental conditions in a given study. Changing environmental factors, for example by intervention, can therefore still shift the distribution of a trait despite low estimates of environmental influences. Thus, heritability estimates say little about the malleability of traits to change; they reflect what *is,* rather than what *could be*. In addition, the concepts of heritable and environmental factors remain abstract in the sense that they do not identify specific modifiable factors that can be targeted for intervention, as is essential in useful causal analysis. This is true not only of classic methods of analysis of variance – e.g. twin heritability – referred to by Lewontin but also of newer methods such as SNP-heritability, which estimate the variance explained by all common SNPs, rather than identify specific genetic variants.

Despite their importance, this collection does not focus on methods aiming to elucidate the genetic architecture of traits or identify causal genetic variants. Instead, we focus on the use of genetics as a powerful tool for establishing causal relationships at the phenotypic level. We aim to delineate how specific phenotypic risk factors cause phenotypic outcomes. That said, methods aiming to elucidate the genetic architecture of traits or identify causal genetic variants provide an essential background to the methods presented in this collection.

Causal inference in observational data

Randomised experiments and their implementation in clinical medicine as Randomised Controlled Trials (RCTs) have come to be considered the gold standard for causal inference. The fundamental intuition is that if a treatment is allocated randomly to different units (e.g. human participants) then the treatment and control group will only differ due to the treatment. Comparing treatment and control groups on any outcome of interest (e.g. disease) thus allows us to establish the causal effect of the treatment and estimate its magnitude. Establishing causation has become so intertwined with experimentation and randomization that mentioning the 'C-word' within observational research has been largely taboo in some fields[41,42]. That said, as discussed below, this taboo has been far from absolute with key contributions to causal reflection and modelling in observational settings in the second half of the 20th century. Still, authors are regularly compelled by journal policies to change wording from 'effect' or 'impact' to ‘association’ or ‘link’, thwarting the need for explicit and transparent reporting of the aim and methods of causal inference studies[41]. Such a rigid application of the mantra “correlation is not causation” has long contributed to stalling the debate on causal inference in observational settings. Reducing causal inference to experimentation is not tenable. First, RCTs have their own limitations, for example that randomization may not balance confounders in any single trial or that it may be difficult to generalize their findings[43,44]. In addition, RCTs are often not feasible or ethical. Yet, public health may require that a pragmatic, even if imperfect, consensus is reached on the causal status of a given risk factor. In such cases, simply computing observed correlations between variables is often unhelpful and investigations within an explicit causal inference framework are required.

For example, no one now contests that smoking cigarettes is a causal risk factor for lung cancer. The causal status of smoking was hotly debated and contested by some including RA Fisher on the basis that confounding (BOX 1) – including genetic confounding – prevented causal inference in observational data[45,46]. But such objections were discarded based on converging observational evidence. In particular, Jerome Cornfield argued that the strength of confounding from genetic factors or other confounders would need to be implausibly high to account for all of the observed effect of smoking and cancer[47], laying the foundations of what is now known as sensitivity analysis. Concluding the debate, the report of the Royal College of Physicians in 1962 and the Surgeon General’s Report in 1964 reached a consensus on the causal status of smoking, paving the way for large-scale prevention efforts[48,49]. This conclusion was attained without randomly allocating human participants to smoking, which would have been unethical.

Since then, and despite the aforementioned resistance, causal inference in observational settings has been continuously refined and formalized with inputs from both statistics and epidemiology. In 1965, Bill Cochran reflected on how to plan observational studies when experimentation is not possible “to elucidate cause-and-effect relationships, or at least to investigate the relationships between one set of specified variables x_i and a second set y_i in a way that suggests or appraises hypotheses about causation”[50]. The choice of words is enlightening here: inference in observational settings may not provide definitive answers but can shift the cursor on a continuum from correlation to virtual causal certainty. Also in 1965, Bradford Hill set out a list of viewpoints to consider when appraising empirical evidence in favour or against a causal hypothesis, including temporal relationships, dose-response relationships and plausibility[51]. Importantly, Hill's list was later misconstrued as a set of 'criteria' to establish causality, a mechanical terminology he neither endorsed nor – it is clear – advocated[52].

In his words: “Here, then, are nine different viewpoints from all of which we should study association before we cry causation. What I do not believe – and this has been suggested – is that we can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more likely than cause and effect?” In the same text, Hill also criticised the overreliance on tests of significance, suggesting that in some cases, descriptive tables are so clear that such tests do not add any value; or that “the glitter” of “magic formulae” can divert our attention from substantial study shortcomings. He concluded: “Like fire, the χ^2 test is an excellent servant and a bad master”.

Hill’s informal approach to causal inference was later criticised (see [53]) as none of his viewpoints is sufficient or necessary to infer causality, which he himself recognized (on the tension between Hill's criteria and statistical formalization, see [53,54]). The most influential formal causal inference frameworks to date are arguably Donald Rubin’s counterfactual or potential outcomes framework and Judea Pearl’s structural models[55,56]. Both frameworks are very general and most causal inference designs or statistical methods in observational settings (and even randomised trials) can be subsumed under these frameworks. Both formalize assumptions under which causal estimates can be attainable in observational data. Exchangeability is a fundamental notion in both frameworks and is achieved when exposed and nonexposed groups are balanced on all confounders, as occurs in an appropriately randomized trial. Causal models within those frameworks can be conveniently represented in formal diagrams, or “Directed Acyclic Graphs” (DAGs) (BOX 1). Both frameworks have a dedicated statistical notation that is considerably more sophisticated than statistics like the t-test and chi-squared test referred to by Hill.

The divide between empiricists like Hill and advocates of the primacy of formal statistical frameworks is unwarranted. First, the focus of both sides is somewhat different. Hill’s address was aimed at practitioners of occupational medicine, with a strong focus on pragmatic decisions. He concludes his address by a case for action: although scientific knowledge is by nature incomplete, acting on such knowledge should not be postponed when required. Conversely, formalists focus on methodological advances aiming to provide the best answer to a causal question and estimates of causal effects under a given set of assumptions. However, even the most sophisticated causal models in observational data can only yield the right causal estimates when the specified model is mostly correct (e.g. a sufficient set of confounders, see BOX 1). Substantive prior knowledge is required to specify appropriate models, assess their assumptions, the plausibility of their findings or even to formulate relevant causal questions in the first place. That is, causal inference cannot be reduced to algorithms. In turn, however, formalised tools remove some unwarranted arbitrariness in the decision-making process regarding the causal status of risk factors[54,57]. As in all empirical sciences, a constant dialogue must be maintained between theoretical frameworks, statistical methods and empirical evidence.

Genetics and phenotypic causal inference

Genetics and causal inference have developed largely in parallel but have converged along two lines of inquiry. First, family-based designs used in quantitative genetics to understand the genetic and environmental architecture of traits have also been used explicitly for causal inference. The idea of using twins that are genetically identical to identify environmental causes of diseases has been present since the 1950s, with, for example, an analysis of smoking habits in twins which concluded that a sufficient number of discordant twins would help in establishing the injurious effect of tobacco smoking[58], which was confirmed much later[59,60]. The approach has then been systematized on large twin samples, based on the principle that identical twins exposed to a risk factor can be matched with their non-exposed co-twins[61,62], enhancing exchangeability. Other family-based designs such as the in-vitro fertilization design can be used to account for genetic confounding by comparing genetically related and genetically unrelated parent-child pairs[5]. Many such methods and examples of applications are proposed in this collection.

Second, and more recently, measured genetic variants associated with an exposure (e.g. cholesterol) have been used as instruments (BOX 1) to estimate the causal effect of that exposure on relevant outcomes (e.g. cardiovascular diseases). This approach was named Mendelian randomization, as it capitalises on the randomisation of genetic material occurring at conception in order to approximate exchangeability and strengthen inference. Interestingly, the idea that randomisation at conception can help for controlled comparison had been grasped by

Fisher. Indeed, Fisher established himself the filiation between his central contribution to the statistics of randomized experiments and his early work on the transmission of Mendelian factors[63]. In his words:

“And here I may mention a connection between our two subjects which seem not to be altogether accidental, namely that the factorial method of experimentation, now of lively concern so far afield as the psychologists, or the industrial chemists, derives its structure and its name, from the simultaneous inheritance of Mendelian factors (...) Genetics is indeed in a peculiarly favoured condition in that Providence has shielded the geneticist from many of the difficulties of a reliably controlled comparison. The different genotypes possible from the same mating have been beautifully randomised by the meiotic process. A more perfect control of conditions is scarcely possible, than that of different genotypes appearing in the same litter."[64].

As such, Mendelian randomisation can be construed as a return to the roots of causal inference; it has developed considerably over the past decade with a flurry of methods and applications reviewed in this collection.

Contributions

Lynch’s contribution discusses the specific meaning of “cause” in genetics from a philosophical perspective, building on the distinction between Mendelian and molecular genes. The rest of this collection focuses more pragmatically on describing genetically informed methods for causal inference and their applications. *Thapar & Rice* present a range of family-based designs for causal inference while *McAdams et al.* focus on the twin design and its extensions to larger pedigrees. *Richmond & Davey Smith* turn to explaining the fundamentals of Mendelian Randomization and how genetic variants and in particular SNPs can be used as instruments for causal inference. *Dudbridge* follows up on the many extensions of Mendelian randomization jointly modelling many SNPs as instruments. *Sanderson* focuses on a special case of polygenic Mendelian randomization using instruments associated with several exposures to identify the independent or mediating effects of such exposures on outcomes of interest. Although the two major lines of research using genetics for causal inference – i.e. family-based and Mendelian randomization – have emerged and evolved independently, *Hwang et al.* outline the major opportunities arising from their recent integration. To some extent, family-based MR returns to Fisher's insight of randomisation of genotypes *in the same litter*.

Note that we define phenotypes broadly as including any individual characteristic other than genotypes, which includes all omics other than genomics. Methods covered in *Kutalik et al.* aim to query the potential of genetically informed methods in elucidating the role of metabolomics as modifiable risk factors for diseases. Such research questions build on large-scale and growing datasets and *Richardson et al.* cover much needed computational tools for causal inference.

Experimental and observational causal inference methods have often been artificially opposed. However, formal causal inference languages subsume both under the same theoretical frameworks and notations and, in practice, they can and should be complementary. *Ference et al.* show how Mendelian randomization can be used to improve the design of randomised control trials and *Schmidt et al.* how genetics can be used to prioritize drug targets for trials.

Naturally, each of the methods covered in this collection has its own challenges and limitations. *Munafò et al.* conclude this collection by reflecting on how triangulation of evidence from multiple genetically informed and non-genetic methods can help in further strengthening causal inference.

Trends and future developments

Following a more succinct attempt[65], this collection is the first to comprehensively cover genetically informed designs for causal inference. New trends are already apparent and should further develop in the near future. We expect that the development of new methods or the refinement of existing ones will continue at a fast pace. At the theoretical level, classical models such as the discordant twin models should be rewritten using more formal causal inference language to better understand their underlying assumption and the meaning of the resulting causal estimates[66]. Ever more robust Mendelian randomization estimators are continuously being developed. In particular, new methods leveraging genome-wide data for causal inference are emerging and should become a powerful viable complement to current approaches that use a few dozen or hundreds of genetic variants as instruments[67,68]. Emerging methods further discussed by *Hwang et al.* in this collection like Mendelian randomization within families are promising to address some of the shortcomings of Mendelian randomization[69]. In addition, these approaches offer new opportunities to further examine old questions such as what underlying processes explain the transmission of risk across generations. Intergenerational causal inference can elucidate whether parental risk factors have causal effects on offspring outcomes or whether intergenerational associations are better explained by genetic and environmental confounding[70–72].

So far, quantitative genetic methods have largely relied on controlling for confounding to strengthen inference whereas methods using molecular genetic data like Mendelian randomization have relied on instrumental variable approaches (BOX 1). New methods can arise from crossing these boundaries. For example, genetic scores can be used as instruments within the twin design[73]. Genetically informed methods can also be combined with more classical methods for causal inference. For example, Mendelian randomization can be combined with negative control analyses[74]. Genome-wide polygenic scores can be used to implement genetically informed sensitivity analyses[65,75], building on the concept of sensitivity analysis that emerged from the work of Jerome Cornfield during the smoking-lung cancer controversy.

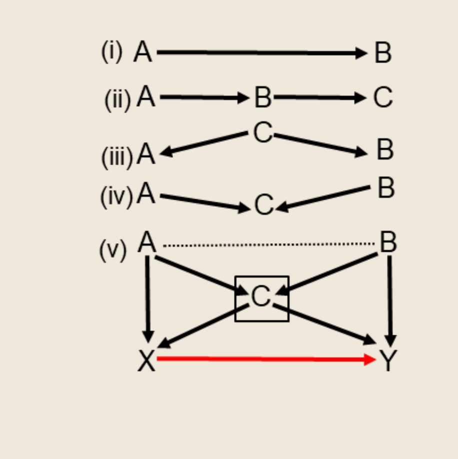
The scope of application of methods presented in this collection becomes wider as new datasets are made available. The emerging literature on the genetic architecture of COVID-19, made possible by the data collected by the COVID-19 Host Genetics Initiative offers a good example of how methods presented in this collection can make decisive contributions to emerging questions. A genetic instrument for IL6R was found to be associated with a lower risk of hospitalisation for COVID-19, suggesting the relevance of therapeutic inhibition of the IL-6 receptor, which has now been confirmed in clinical trials[9,76]. Mendelian randomization has also been used to systematically scan hundreds of druggable proteins to priorise targets for drug trials, for example highlighting 0AS1 as a candidate for drug development[77,78]. Additional studies point towards host antiviral defence mechanisms and mediators of inflammatory organ damage as mechanisms underlying critical illness in COVID-19[79].

Although genetically informed causal inference is relatively recent, we expect that we will start reaping rewards in the near future, i.e. not only in terms of our understanding of human disease and development but in terms of tangible translational applications such as drug development.

We thank Barbara Acosta and colleagues for their patience and help in putting this collection together. We thank contributors who took the time to write valuable contributions to this collection and to the field, and to reviewers who read and provided important feedback on the contributions. We hope the reader will learn from these contributions as much as we did.

//

Box 1. Causal diagrams

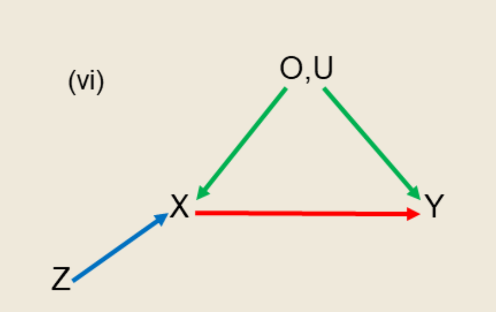


Directed acyclic graphs (DAGs) can be used to encode causal models and assumptions. In (i) the directed arrow represents a causal effect of A on B. In (ii) the directed path goes from A to C via B. B is here a mediator, in the sense that the causal effect of A on C is happening indirectly via B. In DAGs, absent arrows are as important as represented arrows. In (ii) for example, we assume that all the effect of A on C is mediated by B, i.e. there is no additional arrow directly from A to C or via another variable than B. In (iii) C independently causes A and B. C is therefore a confounder of the association between A and B. The path between A and B via C is called a 'backdoor' path. Such a backdoor path creates an observed association between A and B even in the absence of a causal effect, that is represented by the absence of a directed arrow between A and B. This constitutes a fundamental challenge in epidemiology as observed associations between two variables cannot be assumed to stem from a causal relationship between those two variables. If C has been observed and perfectly measured, then statistically adjusting for C will remove confounding and enable the estimation of the causal effect between A and B, here a null effect.

Directions of arrows matter in DAGs. For example, if the arrow between A and C is reversed in (iii) then C becomes a mediator rather than a confounder. In (iv) C is a collider as both arrows from A and B ‘collide’ in C. In this situation, the path is blocked in C. As such, and contrary to the confounder situation, there is no observed association between A and B. However, if C is adjusted for, this creates a spurious association between A and B. This collider bias is another key challenge in epidemiology. If C is a collider but mistakenly identified as a confounder, the adjusted association will be further from the causal effect than the unadjusted association. Collider bias can also generate bias in many study settings. For example, if two independent factors (A and B) cause hospitalization (C), then, in a study restricted to hospitalised patients, A and B will be associated. This is because the stratification (i.e. focusing only on hospitalised rather than hospitalised and non-hospitalised people) is a form of adjustment.

In (V) C is a confounder of X and Y and should therefore be adjusted for in order to retrieve the causal effect of X on Y. However, C is also a collider of A and B. Adjusting for C thus creates a spurious association between A and B, which introduces a backdoor path from X to Y via A and B. The induced association upon confounder adjustment in this context has been referred to as “M-bias”. In addition to adjusting for C, it is thus necessary to adjust for A or/and B to further block the newly created path. Importantly, in theory, if the model DAG corresponds to the true model, finding a sufficient set of confounders, here C and A (or B) for example, is sufficient to retrieve the causal effect of X and Y. In practice, however, we do not know the underlying causal model and the variables are not measured without error. This is a major impediment for causal inference based on statistical adjustment only given the nature of epidemiological data where unclear underlying models, unmeasured confounders and measurement error are the norm. In this collection, we present a number of methods which (partly) adjust by design for unobserved confounders (e.g. the twin design).

.



The DAG (vi) encodes the instrumental variable design. Z is the instrument, which is used in an 'instrumental' fashion to estimate the causal effect of X on Y. Note that the DAG encodes three assumptions of the instrumental variable approach which are necessary for Z to be a valid instrument, enabling the inference that X causes Y. First, Z needs to be (robustly) associated with X (blue arrow), which is called the relevance assumption, i.e. Z needs to be relevant to assess the effect of X. The second assumption is exchangeability and is encoded by the absence of a common cause of Z and Y. Exchangeability is key to understand why X enables us to make causal inference regarding X to Y. To illustrate, if Z is binary and positively predicts X, then participants in group Z_1 will have higher levels of X than participants Z_0. Although they differ on the level of X, participants Z_0 and Z_1 do not differ on any other variables. Participants Z_0 and Z_1 are thus exchangeable and only differ on the exposure X. If Z_0 and Z_1 have different outcomes, i.e. different levels of Y, we can conclude that X is causally related to Y. This is similar to a RCT, in the sense that Z plays the role of the random assignment which creates two groups with a different level of the variable influenced by the treatment X but balanced on all other confounders. Third, Z needs to be associated to Y only via its effect on X, which is called the restriction exclusion assumption. In other words, similarly to DAG ii, X fully mediates the association between Z and Y. When using a genetic instrument, this is often called mediated pleiotropy (or vertical pleiotropy) as opposed to unmediated (or horizontal) pleiotropy, which would be represented by a direct arrow from Z to Y. Importantly, even when the DAG in (vi) fully holds, we do expect an observed association between Z and Y, which is equal to the path from Z to Y via X.

//

References

1. Mokry, L. E. *et al.* Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. *PLOS Med.* **12**, e1001866 (2015).
2. Kho, P. F., Glubb, D. M., Thompson, D. J., Spurdle, A. B. & O'Mara, T. A. Assessing the Role of Selenium in Endometrial Cancer Risk: A Mendelian Randomization Study. *Front. Oncol.* **9**, (2019).
3. Prins, B. P. *et al.* Investigating the Causal Relationship of C-Reactive Protein with 32 Complex Somatic and Psychiatric Outcomes: A Large-Scale Cross-Consortium Mendelian Randomization Study. *PLoS Med.* **13**, (2016).
4. Mohammadi-Shemirani, P. *et al.* Effects of lifelong testosterone exposure on health and disease using Mendelian randomization. *eLife* **9**, e58914 (2020).
5. Thapar, A. *et al.* Prenatal smoking might not cause attention-deficit/hyperactivity disorder: evidence from a novel design. *Biol. Psychiatry* **66**, 722–727 (2009).
6. Mountjoy, E. *et al.* Education and myopia: assessing the direction of causality by mendelian randomisation. *The BMJ* **361**, (2018).
7. Singham, T. *et al.* Concurrent and Longitudinal Contribution of Exposure to Bullying in Childhood to Mental Health: The Role of Vulnerability and Resilience. *JAMA Psychiatry* **74**, 1112–1119 (2017).
8. Wong, J. Y. Y. *et al.* Tuberculosis infection and lung adenocarcinoma: Mendelian randomization and pathway analysis of genome-wide association study data from never-smoking Asian women. *Genomics* **112**, 1223–1232 (2020).
9. Bovijn, J., Lindgren, C. M. & Holmes, M. V. Genetic variants mimicking therapeutic inhibition of IL-6 receptor signaling and risk of COVID-19. *Lancet Rheumatol.* **2**, e658–e659 (2020).
10. Davey Smith, G. & Phillips, A. N. Correlation without a cause: an epidemiological odyssey. *Int. J. Epidemiol.* **49**, 4–14 (2020).

11. Mendelson, M. M. *et al.* Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS Med.* **14**, (2017).

12. Mendel G. Experiments on plant hybrids. in *Gregor Mendel's Experiments on Plant Hybrids: A Guided Study* (Rutgers University Press, 1993).

13. Galton, F. *Natural Inheritance*. (Macmillan, 1889).

14. Pearson, K. & Henrici, O. M. F. E. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character* **187**, 253–318 (1896).

15. Visscher, P. M. Commentary: Height and Mendel's theory: the long and the short of it. *Int. J. Epidemiol.* **42**, 944–945 (2013).

16. Yule, G. U. Mendel's Laws and Their Probable Relations to Intra-Racial Heredity. *New Phytol.* **1**, 222–238 (1902).

17. Brownlee, J. The Inheritance of Complex Growth Forms, such as Stature, on Mendel's Theory*. *Int. J. Epidemiol.* **42**, 932–934 (2013).

18. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **52**, 399–433 (1918).

19. Plomin, R., DeFries, J. C., Knopik, V. S. & Neiderhiser, J. M. *Behavioral genetics*. (Worth Publishers, 2013).

20. Jonsson, H. *et al.* Differences between germline genomes of monozygotic twins. *Nat. Genet.* **53**, 27–34 (2021).

21. Tikhodeyev, O. N. & Shcherbakova, O. V. The Problem of Non-Shared Environment in Behavioral Genetics. *Behav. Genet.* **49**, 259–269 (2019).

22. Smith, G. D. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *Int. J. Epidemiol.* **40**, 537–562 (2011).

23. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375, S1-3 (2012).

24. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).

25. Broekema, R. V., Bakker, O. B. & Jonkers, I. H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* **10**, 190221 (2020).

26. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

27. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

28. Peyre, H. *et al.* Combining multivariate genomic approaches to elucidate the comorbidity between autism spectrum disorder and attention deficit hyperactivity disorder. *J. Child Psychol. Psychiatry* **n/a**,.

29. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).

30. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

31. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genet.* **9**, e1003348 (2013).

32. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

33. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

34. Krapohl, E. *et al.* Phenome-wide analysis of genome-wide polygenic scores. *Mol. Psychiatry* **21**, 1188–1193 (2016).

35. Krapohl, E. *et al.* Widespread covariation of early environmental exposures and trait-associated polygenic variation. *Proc. Natl. Acad. Sci.* **114**, 11727–11732 (2017).

36. Sun, L. *et al.* Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLoS Med.* **18**, e1003498 (2021).

37. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).

38. Wray, N. R. *et al.* From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA Psychiatry* **78**, 101–109 (2021).

39. Lewontin, R. C. The analysis of variance and the analysis of causes. 1974. *Int. J. Epidemiol.* **35**, 520–525 (2006).

40. Vreeke, G.-J. Commentary: The attainability of causal knowledge of genetic effects in complex human traits. *Int. J. Epidemiol.* **35**, 531–534 (2006).

41. Hernán, M. A. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *Am. J. Public Health* **108**, 616–619 (2018).

42. Grosz, M. P., Rohrer, J. M. & Thoemmes, F. The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. *Perspect. Psychol. Sci.* **15**, 1243–1255 (2020).

43. Imai, K., King, G. & Stuart, E. A. Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Stat. Soc. Ser. A Stat. Soc.* **171**, 481–502 (2008).

44. Deaton, A. & Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* (2017) doi:10.1016/j.socscimed.2017.12.005.

45. Fisher, R. A. Alleged dangers of cigarette-smoking. *Br. Med. J.* **2**, 4 & 297–298 (1957).

46. Fisher, R. Cigarettes, Cancer, and Statistics. *Centen. Rev. Arts Sci.* **2**, 151–166 (1958).

47. Cornfield, J. *et al.* Smoking and Lung Cancer: Recent Evidence and a Discussion of

Some Questions. *JNCI J. Natl. Cancer Inst.* **22**, 173–203 (1959).

48. *Smoking and Health. Summary of a Report of the Royal College of Physicians of London on SMoking in relation to Cancer of the Lung and Other Diseases*. (Pitman Medical Publishing Co. Ltd., 1962).

49. *Smoking and Health. Report of the advisory committee to the surgeon general of the public health service.* 386 (1964).

50. Cochran, W. G. & Chambers, S. P. The Planning of Observational Studies of Human Populations. *J. R. Stat. Soc. Ser. Gen.* **128**, 234–266 (1965).

51. Hill, A. B. The Environment and Disease: Association or Causation? *Proc. R. Soc. Med.* **58**, 295–300 (1965).

52. Davey Smith, G. Post–Modern Epidemiology: When Methods Meet Matter. *Am. J. Epidemiol.* **188**, 1410–1419 (2019).

53. Rothman, K. J. The Wrong Message from the Wrong Talk. *Obs. Stud.* **6**, 30–32 (2020).

54. VanderWeele, T. J. Hill's Causal Considerations and the Potential Outcomes Framework. *Obs. Stud.* 47–54 (2020).

55. Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences*. (Cambridge University Press, 2015).

56. Pearl, J. *Causality*. (Cambridge University Press, 2009).

57. Baiocchi, M. Following Bradford Hill, in Reprint of Hill's 'The Enviroment and Disease: Association or Causation?' and Comments. *Obs. Stud.* **6**, 1–9 (2020).

58. Friberg, L., Kaij, L., Dencker, S. J. & Jonsson, E. Smoking Habits of Monozygotic and Dizygotic Twins. *Br. Med. J.* **1**, 1090–1092 (1959).

59. Hjelmborg, J. *et al.* Lung cancer, genetic predisposition and smoking: the Nordic Twin Study of Cancer. *Thorax* **72**, 1021–1027 (2017).

60. Kaprio, J. & Koskenvuo, M. Twins, smoking and mortality: a 12-year prospective study of smoking-discordant twin pairs. *Soc. Sci. Med. 1982* **29**, 1083–1089 (1989).

61. McGue, M., Osler, M. & Christensen, K. Causal inference and observational research: the utility of twins. *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* **5**, 546–556 (2010).

62. Carlin, J. B., Gurrin, L. C., Sterne, J. A., Morley, R. & Dwyer, T. Regression models for twin studies: a critical review. *Int. J. Epidemiol.* **34**, 1089–1099 (2005).

63. Smith, G. D. Commentary: Random Allocation in Observational Data: How Small But Robust Effects Could Facilitate Hypothesis-free Causal Inference. *Epidemiology* **22**, 460–463 (2011).

64. Fisher, R. Statistical methods in genetics1. *Int. J. Epidemiol.* **39**, 329–335 (2010).

65. Pingault, J.-B. *et al.* Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.* **19**, 566–580 (2018).

66. Petersen, A. H. & Lange, T. What Is the Causal Interpretation of Sibling Comparison Designs? *Epidemiology* **31**, 75–81 (2020).

67. Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M. & He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.* **52**, 740–747 (2020).

68. Darrous, L., Mounier, N. & Kutalik, Z. Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. *medRxiv* 2020.01.27.20018929 (2020) doi:10.1101/2020.01.27.20018929.

69. Howe, L. J. *et al.* Within-sibship GWAS improve estimates of direct genetic effects. *bioRxiv* 2021.03.05.433935 (2021) doi:10.1101/2021.03.05.433935.

70. Balbona, J., Kim, Y. & Keller, M. C. Estimation of parental effects using polygenic scores. *bioRxiv* 2020.08.11.247049 (2020) doi:10.1101/2020.08.11.247049.

71. Lawlor, D. *et al.* Using Mendelian randomization to determine causal effects of maternal pregnancy (intrauterine) exposures on offspring outcomes: Sources of bias and methods for assessing them. *Wellcome Open Res.* **2**, 11 (2017).

72. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–

428 (2018).

73. Minică, C. C., Dolan, C. V., Boomsma, D. I., de Geus, E. & Neale, M. C. Extending Causality Tests with Genetic Instruments: An Integration of Mendelian Randomization with the Classical Twin Design. *Behav. Genet.* **48**, 337–349 (2018).

74. Sanderson, E., Richardson, T. G., Hemani, G. & Davey Smith, G. The use of negative control outcomes in Mendelian randomization to detect potential population stratification. *Int. J. Epidemiol.* (2021) doi:10.1093/ije/dyaa288.

75. Pingault, J.-B. *et al.* Genetic sensitivity analysis: Adjusting for genetic confounding in epidemiological associations. *PLOS Genet.* **17**, e1009590 (2021).

76. The WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group *et al.* Association Between Administration of IL-6 Antagonists and Mortality Among Patients Hospitalized for COVID-19: A Meta-analysis. *JAMA* (2021) doi:10.1001/jama.2021.11330.

77. Zhou, S. *et al.* A Neanderthal OAS1 isoform Protects Against COVID-19 Susceptibility and Severity: Results from Mendelian Randomization and Case-Control Studies. *medRxiv* 2020.10.13.20212092 (2020) doi:10.1101/2020.10.13.20212092.

78. Gaziano, L. *et al.* Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. *medRxiv* 2020.11.19.20234120 (2020) doi:10.1101/2020.11.19.20234120.

79. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in Covid-19. *Nature* 1–1 (2020) doi:10.1038/s41586-020-03065-y.