The Wright Stuff: Genes in the Interrogation of Correlation and Causation

GEORGE DAVEY SMITH

MRC Centre for Causal Analyses in Translational Epidemiology, University of Bristol

KZ.Davey-Smith@bristol.ac.uk

Abstract: The contemporary use of what are now called causal diagrams can be traced back to Sewall Wright's introduction of path coefficients in the early 1920s. Wright was explicit that causal evidence was required to formulate such diagrams and that these schema could not alone provide evidence for or against causality. In population sciences, germline genetic variation can provide required anchors for the separation of causal from (mere) correlational associations. Advances in biological and other material sciences offer more for improved causal understanding than new ways of conceptualising and representing associations. Copyright © 2012 John Wiley & Sons, Ltd.

For most branches of science, the distinction between (mere) correlation and causation is a central issue. My discipline, epidemiology, is one prone to over interpretation—by the media, by researchers or both—of associations observed in data sets that are most plausibly explained by chance, bias or confounding (Davey Smith & Ebrahim, 2002). James Lee muses on these issues in the context of behavioural traits within the psychological literature and promotes the graphical approach (in particular, directed acyclic graphs) now beloved of many working within the epidemiological tradition. His clear presentation merits a close reading and raises issues of general relevance. I will focus on the opportunities offered by his statement that 'the soundness of any causal conclusion depends on both conforming data and the correctness of the requisite assumptions. Our substantial prior knowledge of genetics justifies many powerful assumptions which lead to correspondingly powerful results.' Indeed, leveraging the power of germline genetic variation transforms our ability to elucidate the causal chains within the networks of associations within the biological realm (Zhu et al., 2007), and whereas graphical presentations may help, it is the biological realities, rather than new ways to draw these on paper, that contain the most promise. These are only now beginning to yield findings but will transform how we approach causality in the population sciences.

Lee invokes the evolutionary biologist and population geneticist Sewall Wright, the progenitor of path analysis (and, through that, structural equation modelling, favored more in the psychological than epidemiological literature) in the prehistory of the now triumphant directed acyclic graph. I must admit to being pleased that structural equation models largely failed to penetrate epidemiology; their (sometimes) manner of presentation as a form of alchemy that can isolate causal pathways in an intercorrelated morass of data being scarcely credible. In the epidemiological setting, underlying social and biological processes, combined with reverse causation (outcome influencing apparent exposure, rather than *vice versa*), leads to association being the norm rather than the exception (Davey Smith et al., 2008). Levels of measurement error that exist in most domains simply cannot be disciplined, and the confident production of coefficients that apparently have meaning seems chimeral. Thus, coming across Wright, authoring a paper in 1921 with the exact same title as Lee, setting out his stall for his form of path analysis was enlightening:

> The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated by well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.
>
> The present paper is an attempt to present a method of measuring the direct influence along each separate path

DOI: 10.1002/per

> in such a system and thus of finding the degree of which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain the method can be used to find the logical consequences of any particular hypothesis in regard to them. (Wright, 1921, p. 557)

Wright's famous path analyses (Figure; Wright, 1920) required prior causal knowledge to make sense. With this, they introduced important new understanding, not the least of which was the identification of what Wright termed 'intangible variance'—induced by what we may call stochastic or chance events—that lead to group level, rather than individual trajectory, understanding. This is the best that can ever be hoped for in the population sciences (Davey Smith, 2011b).

To an extent, known biological relationships in quantitative genetic analyses in the behavioural genetics field provide a form of reliable prior information on the presence and direction of causation. However, in the molecular genetics era, the most powerful source of prior causal knowledge that can, and is, now being leveraged comes from germline genetic variants that have established associations with particular traits. R.A. Fisher explicitly referred to the essentially randomised nature of genetic perturbations (Fisher, 1952), as Lee mentions and as others directly associated with Fisher have written about (Bodmer, 2003; Box, 2010), although the possibility that Mendelian randomisation came before experimental randomisation in Fisher's intellectual biography has been little recognised (Davey Smith, 2006). That genetic variation inducing a group-level difference in a potentially modifiable phenotype can provide evidence of the downstream causal effect of this phenotype, free of the influence of confounding or reverse causation, is now widely recognised and implemented in epidemiological studies (Timpson, Wade, & Davey Smith, 2012). To give just one example of relevance to the study of behavioural traits—the topic of Lee's paper—such 'Mendelian randomization' (as the method is generally termed; Davey Smith & Ebrahim, 2003) has been applied to the effects of smoking. As proof of principle, such studies have demonstrated that a genetic variant robustly associated with smoking behaviour relates to lung cancer risk to the degree expected by the association of the variant with appropriately ascertained smoking behaviour (Munafo et al., 2012; Wang, Broderick, Matakidou, Eisen, & Houlston, 2011), and associations with several other smoking-related diseases have been made. Such studies have also shown that smoking lowers body mass index (Freathy et al., 2011); despite naive observational associations sometimes being in the opposite direction, given confounding by socioeconomic position and various other socially patterned exposures.

The various assumptions of such Mendelian randomisation studies have been reviewed (Davey Smith, 2010; Lee, this issue; Sheehan, Didelez, Burton, & Tobin, 2008) and are reflected in Lee's discussion of the distinction between Fisher's notion of the as-observed 'average excess' associated with a genetic difference and the 'average effect' that would be seen with a gene substitution. That confounding can exist in genetic association studies is of course widely recognised, with ancestral population differences in both gene frequency and disease risk ('population stratification') being the most likely culprit. There are well-established methods of accounting for this using genome-wide data as indices of such population stratification, and with established genetic variant-phenotype links, it is remarkable how homogeneous the associations seen within different populations generally are, despite allele frequency often varying between populations (Hindorff et al., 2012). Empirical data also demonstrate that confounding of genetic variants with social, behavioural and physiological factors that plague conventional observational studies are conspicuous by their absence (Davey Smith et al., 2008).

Lee considers at length the possibility that selection bias related to participation in studies could bias findings. Thus, if a genetically influenced trait was related to willingness to participate in a study, and this was differential for cases and controls, a spurious association could be generated. This is in principle true, but common control groups have been used for various diseases (e.g. the Wellcome Trust Case Control Consortium, 2007), and unless the participation effect was condition specific, such bias would generate similar associations for all the diseases, which were not seen. Even if such a participation effect was disease specific, it would only influence case-control studies, not prospective studies, and generally, genetic associations have been similar across study designs (Hindorff et al., 2012). More complex hypotheses could be advanced involving interactions of genetic variants influencing participation and condition-specific disease risk, but plausibility decreases with increasing elaboration of the hypothesis in this regard. Again, the fact that similar effects for established variants tend to be seen in designs with widely differing participation rates, from high response rate general population cohorts to what are essentially volunteer studies, is reassuring in this regard.

Graphical approaches to causal inference are certainly of value in forcing investigators to be explicit about their assumptions and can help in the identification of unrecognised potential biases. There are also often unrecognised drawbacks to formulaic or mechanical imposition of such approaches (Dawid, 2008). In epidemiological circles, it is now not uncommon to receive peer review comments that focus on "the possible adjustment for a collider in model 3 of Supplementary Table 4", the reviewer clearly considering this more important than having an informed overview of the totality of evidence presented. 'Inference to the best explanation' (Lipton, 2004), which is surely what any attempt at causal reasoning is aiming at, can go out of the window as the d-connected nodes, rather than how the world actually is, become the focus of attention.

DOI: 10.1002/per

Wright opined that 'great refinement in statistical treatment is often a waste of effort' (Wright, 1917). William Provine (1986), in his unsurpassable intellectual biography of Wright, discusses the development of path analysis and how, working with methods that tried to hold other factors constant through statistical manipulation, 'Wright was still dissatisfied. He saw clearly that by itself the partial correlation coefficient, like the correlation coefficient, was a mathematical quantity not tied or leading by itself to any causal interpretation of the relations under examination. Wright wanted to minimise correlational statistics and maximise the quantitative causal interpretation of the variables' (p. 127). This can only be carried out when causal anchors—that come from how the material world is, not how we draw diagrams on paper—are introduced into the mix. Germline genetic variants provide precisely such anchors and open up vast new vistas of possible causal understanding generated from observational data (Davey Smith, 2011a).

ACKNOWLEDGEMENTS

Thanks to Dave Evans for discussion of selection bias in genome-wide association study.

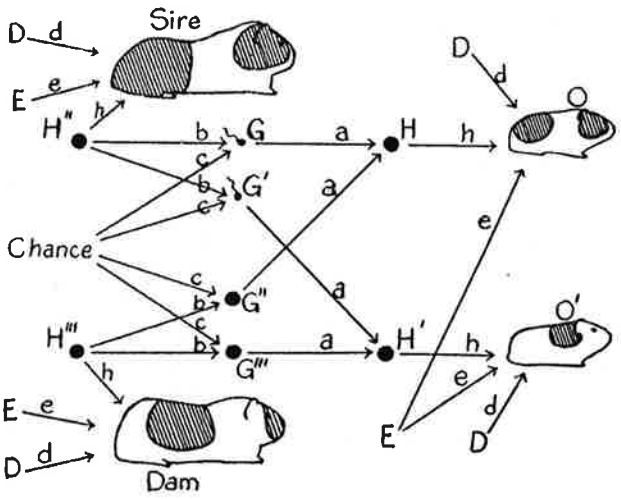

Diagram illustrating the casual relations between litter mates (O, O') and between each of them and their parents. H, H', H" and H"' represent the genetic constitutions of the four individuals; G, G', G" and G"' represent that of four germ cells. E represents such environmental factors as are common to litter mates. D represents other factors, largely ontogenetic irregularity. The small letters stand for the various path coefficients.