Causal inference from observational data using Mendelian randomisation

Sha Meng [1], Vanessa Didelez [2] and Nuala A. Sheehan[1]

[1] Department of Health Sciences, School of Medicine, University of Leicester, UK
[2] Department of Mathematics, University of Bristol, UK

Detection and assessment of the effect of some modifiable risk factor on a disease with view to informing public health intervention policies are of fundamental concern in aetiological epidemiology. In order to have solid evidence that such a public health intervention will have an effect, it is necessary to ascertain that an observed association or correlation between a risk factor and a disease means that the risk factor is *causal* for the disease. Inferring causality from observational data is difficult as it is not always clear which of two associated variables is the cause, which the effect, or whether both are common effects of a third unobserved variable, or confounder. Confounding in these applications is usually due to social, behavioural or physiological factors which are difficult to control for and particularly difficult to measure accurately.

A possible approach to testing for or estimating causal effects when confounding is believed to be present but not fully understood is based on the method of instrumental variables and is known under the name of Mendelian randomisation if the instrument is a genetic predisposition. It exploits the idea that a well understood genotype, known to affect levels of a modifiable exposure or phenotype, only affects the disease status indirectly via its effect on the phenotype and is assigned randomly at meiosis (given the parental types). The implication is that the population genotype distribution is largely unrelated to the usual confounders that have distorted interpretations of findings from observational epidemiology and the genotype thus satisfies the requirements for an instrument.

While testing for the presence of a causal effect is generally straightforward, point estimates of such an effect are only obtainable under additional and fairly strong assumptions. An average causal effect can be estimated in the case where all relationships between the variables are linear with no interactions, for instance. In observational epidemiology, the outcome variable is often a binary indicator of disease status. While it may be reasonable to model the relationship between genotype and phenotype as a linear regression, that between phenotype and disease will be non-linear e.g. logistic regression. There are several situations which permit estimation of specific causal parameters for binary outcomes but what can be estimated, and under what assumptions, will depend upon the scenario considered.

We will discuss the main issues with binary outcome data for Mendelian randomisation applications. In particular, we will consider alternative methods of estimation for the non-linear case and some simulation results on sensitivity to various assumptions will be presented.