Causal Effect Estimation in Mendelian Randomisation Studies - Evaluating a Modern Bayesian Approach to Genetic Pleiotropy Versus Established Weighted Median Methodology

B233241

September 2024 - July 2025

Contents

Acknowledgements

I would like to thank Dr Athina Spiliopoulou for her expertise, kindness, and infinite patience in supervising this dissertation. I thank Professor Paul McKeigue for his input regarding the background and methodology of the project. Finally, my deepest gratitude goes to my wife, for keeping all our plates spinning while I've been busy shouting at my laptop.

Statement of Originality & Contributions

I confirm that all work is my own except where indicated, and that all sources are clearly referenced. Parts of this dissertation were informed by my participation in a related project with my supervisor's research group, which included comparison of Mendelian randomisation causal estimation methods using similar methodology, i.e. data simulation and citation searching. I confirm that all simulation code, literature searches and analyses contained within this dissertation are solely my own work, produced under the appropriate guidance of my supervisor.

Word Count

Word count: 9998

1 Abstract

1.1 Background

Mendelian randomisation (MR) uses data from observational genetic studies to support causal inference between exposures and outcomes of interest. Pleiotropy - where genetic variants influence outcomes through multiple pathways - can bias MR causal estimates. Inadequate controls for pleiotropy likely contribute to high false-positive causal report rates across MR literature. MR-Hevo is a recently proposed MR methodology claiming superior handling of pleiotropy and fewer false-positive reports of causality versus the established weighted median estimator (WME) method.

1.2 Aims

To evaluate differences in causal effect estimates between WME versus MR-Hevo methods, and to establish whether these may alter conclusions drawn in real-world studies.

1.3 Methods

Outputs from each method were compared through parallel analysis of simulated data, following the published approach used to validate WME. Simulations represented plausible combinations of population parameters and assumption violations. To investigate differences between methods using real-world data, both were applied to a sample of ten highly-cited MR studies reporting WME causal effect estimates alongside sufficient data to allow replication.

1.4 Results

Using simulated data with null causal effect, MR-Hevo demonstrated lower false-positive report rates versus WME across all 24 combinations of parameters/assumption violations considered (mean false-positive report rate 0.41% versus 5.1%). Using simulated data with true causal effect, MR-Hevo demonstrated higher power to detect this, both on average (mean true-positive report rate 31% versus 28%) and in most cases considered (14 of 24). Cases with higher true- and false-positive report rates for WME versus MR-Hevo correlated with conditions biasing away from the null, suggesting MR-Hevo estimates may be more robust to assumption violations. Re-analysis of highly-cited MR studies found poor reproducibility of published WME estimates in 4 of 10 studies included. Causal effect estimates were similar in magnitude between MR-Hevo and WME; conclusions regarding presence of causality were consistent between both methods across all 10 studies.

1.5 Conclusions

Compared to WME, MR-Hevo exhibited lower false-positive report rates and less perturbation by assumption violations. Across published MR literature reporting a causal effect, re-analysis using MR-Hevo may change conclusions in a minority of cases. Future work should investigate the non-reproducibility of MR results observed.

Word count: 349

2 Introduction and Background

2.1 Introduction to Mendelian Randomisation (MR)

Epidemiology is the study of determinants and distribution of disease across populations; a common epidemiological study aim is therefore to seek evidence as to whether a given exposure (e.g. cigarette smoking) may cause a given outcome (e.g. lung cancer)[?]. Logistics limit experimental interventions across large groups, so insights into associations between exposures and outcomes are gleaned from observational data of people in the population of interest. Comparing health outcomes between individuals with different levels of a particular exposure may highlight potential links, e.g. higher cancer incidence in those who smoke more is consistent with a causal role for cigarettes in carcinogenesis[?].

However, correlation does not prove causation. A key epidemiological challenge is accounting for so-called "confounding" factors; these are other variables, associated with both the exposure and the outcome of interest, which represent an alternative causal explanation for any exposure-outcome links observed[?]. If smokers also drink more alcohol than non-smokers, then an observed link between smoking and increased cancer risk could plausibly be caused by increased alcohol exposure, either partially or entirely. Another potential issue with observational data is "reverse causation", where the presumed outcome is in fact a cause of the exposure; this might be the case if a cancer diagnosis drove individuals to drink and smoke more, and data were collected without respect to exposure timings.

Mendelian randomisation (MR) is a methodology intended to support causal inference from observational data. It applies the principles of instrumental variable (IV) analysis to genetic data, performing a type of natural experiment often likened to a randomised-controlled trial (RCT)[?].

In a properly conducted RCT, causality can be inferred due to a randomisation process being used as an "instrument" to allocate different levels of exposures to different experimental groups. If groups are randomly allocated, any confounding variables which might otherwise influence exposure-outcome relationships should be evenly distributed between groups, whether these confounders are known or not. As such, there should be no systematic differences between individuals from different groups in the exposure of interest - that is, there should be no bias[?]. Statistical methods can quantify the probability that any observed outcome differences could have occurred by chance, and thereafter any outcome differences can be interpreted as caused by exposure differences. As allocation and receipt of exposures is known to precede outcome measurements, reverse causality is impossible.

In MR, naturally occurring genetic variants - "genetic instruments" – are chosen based on their known association to an exposure of interest. Random assignment of alleles (i.e. variants of a given gene) from parents to offspring during meiosis creates randomisation analagous to that performed for an RCT – both measured and unmeasured confounders should be distributed evenly between the groups created. Such genetic randomisation should therefore enable valid causal inference, provided that assumptions of IV analysis are met[?].

2.2 Causal Effect Estimation in MR

At its simplest, the relationship between two continuous variables - an exposure X and outcome Y - can be represented as a linear model:

$$Y = \alpha + \beta X + \epsilon \tag{1}$$

where α represents all non-X determinants of Y, β is the causal effect of X on Y and ϵ is an error term. The β term is a numerical measure of strength of causal exposure-outcome association, where:

- $\beta = 0$ implies no causal link between exposure and outcome
- $\beta > 0$ implies X causes Y