9. Appendices

B233241

2025-01-07

Appendix A: List of Abbreviations

Appendix B: Simulation Code

Generating Data and Models

The data generating model used was from Appendix 3 of Bowden et al (ref); the relevant section describing their model is reproduced below:

_“…

$$U_i = \sum_{j=1}^{J} \phi_j G_{ij} + \epsilon_i^U \tag{1}$$

$$X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \epsilon_i^X \tag{2}$$

$$Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \beta X_i + U_i + \epsilon_i^Y \tag{3}$$

for participants indexed by $i = 1, ..., N$, and genetic instruments indexed by $j = 1, ..., J$. The error terms $\epsilon_i^U, \epsilon_i^X$ and ϵ_i^Y were each drawn independently from standard normal distributions. The genetic effects on the exposure j are drawn from a uniform distribution between 0.03 and 0.1. Pleiotropic effects α_j and ϕ_j were set to zero if the genetic instrument was a valid instrumental variable. Otherwise (with probability 0.1, 0.2, or 0.3):

1. In Scenario 1 (balanced pleiotropy, InSIDE satisfied), the α_j parameter was drawn from a uniform distribution between -0.2 and 0.2.
2. In Scenario 2 (directional pleiotropy, InSIDE satisfied), the α_j parameter was drawn from a uniform distribution between 0 and 0.2.
3. In Scenario 3 (directional pleiotropy, InSIDE not satisfied), the ϕ_j parameter was drawn from a uniform distribution between -0.2 and 0.2.

The causal effect of the exposure on the outcome was either $\beta X = 0$ (null causal effect) or $\beta X = 0.1$ (positive causal effect). A total of 10 000 simulated datasets were generated for sample sizes of N = 10 000 and 20 [sic] participants. Only the summary data, that is genetic associations with the exposure and with the outcome and their standard errors as estimated by univariate regression on the genetic instruments in turn, were used by the analysis methods. In the two-sample setting, data were generated on 2N participants, and genetic

associations with the exposure were estimated in the first N participants, and genetic associations with the outcome in the second N participants."_ (ref)

To reproduce this model, code was written in R to generate the relevant participant level data. First, a function ('simulate_MR_data) was written which included parameters specified by Bowden et al, and also to allow testing of data simulation:

This initial simulation function generated data in the following format:

```
## List of 10
##  $ U           :List of 2
##   ..$ : num [1:2000, 1] 0 0 0 0 0 0 0 0 0 0 ...
##   ..$ : num [1:2000, 1] 0 0 0 0 0 0 0 0 0 0 ...
##  $ X           :List of 2
##   ..$ : num [1:1000, 1] 1.12 1.59 1.76 1.49 1.56 ...
##   ..$ : num [1:1000, 1] 1.84 1.7 1.6 1.66 1.5 ...
##  $ Y           :List of 2
##   ..$ : num [1:1000, 1] -0.24 -0.311 -0.393 -0.227 -0.1 ...
##   ..$ : num [1:1000, 1] -0.872 -0.901 -0.772 -0.999 -0.477 ...
##  $ G_X         :List of 2
##   ..$ : int [1:1000, 1:25] 0 1 1 1 1 0 0 0 0 0 ...
##   ..$ : int [1:1000, 1:25] 1 2 1 2 2 2 2 2 2 2 ...
##  $ G_Y         :List of 2
##   ..$ : int [1:1000, 1:25] 0 1 1 0 1 0 0 0 0 0 ...
##   ..$ : int [1:1000, 1:25] 2 2 2 2 1 2 1 1 2 1 ...
##  $ alpha       :List of 2
##   ..$ : num [1:25] -0.106 0 -0.121 0 0 ...
##   ..$ : num [1:25] 0 0 -0.0786 0 0 ...
##  $ gamma       :List of 2
##   ..$ : num [1:25] 0.0902 0.0878 0.08 0.0832 0.084 ...
##   ..$ : num [1:25] 0.0374 0.0721 0.0975 0.085 0.0322 ...
##  $ phi         :List of 2
##   ..$ : num [1:25] 0 0 0 0 0 0 0 0 0 0 ...
##   ..$ : num [1:25] 0 0 0 0 0 0 0 0 0 0 ...
##  $ beta        :List of 2
##   ..$ : num 0.1
##   ..$ : num 0.1
##  $ prop_invalid:List of 2
##   ..$ : num 0.3
##   ..$ : num 0.3
```

A function (`extract_models`) was then written to create linear models from each dataset generated as per Bowden et al:

This model generated estimates of the coefficient of gene:exposure association (`coeff_G_X`), coefficient of gene:outcome association (`coeff_G_Y`), and the relevant standard errors of these estimates. The values of parameters inputted were also returned to aid in further testing of data/model generation, i.e. actual gene:exposure associations (`gamma`), pleiotropic effects of invalid instruments (`alpha`), additional pleiotropic effects when InSIDE assumption not satified (`phi`), causal effect of exposure on outcome (`beta`) and the proportion of invalid genetic instruments with pleiotropic effects on the outcome (`prop_invalid`).

```
##     dataset    Instrument   coeff_G_X        coeff_G_X_SE
##  Min.   :1   Min.   : 1   Min.   :0.03006   Min.   :1.591e-16
##  1st Qu.:1   1st Qu.: 7   1st Qu.:0.03791   1st Qu.:1.702e-16
##  Median :1   Median :13   Median :0.05578   Median :1.847e-16
```