Software Application Profile

MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data

Olena O. Yavorska[1] and Stephen Burgess[2]*

[1]Newnham College, University of Cambridge and [2]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

*Corresponding author. Strangeways Research Laboratory, 2 Worts Causeway, Cambridge CB1 8RN, UK. E-mail. sb452@medschl.cam.ac.uk

Abstract

MendelianRandomization is a software package for the R open-source software environment that performs Mendelian randomization analyses using summarized data. The core functionality is to implement the inverse-variance weighted, MR-Egger and weighted median methods for multiple genetic variants. Several options are available to the user, such as the use of robust regression, fixed- or random-effects models and the penalization of weights for genetic variants with heterogeneous causal estimates. Extensions to these methods, such as allowing for variants to be correlated, can be chosen if appropriate. Graphical commands allow summarized data to be displayed in an interactive graph, or the plotting of causal estimates from multiple methods, for comparison. Although the main method of data entry is directly by the user, there is also an option for allowing summarized data to be incorporated from the PhenoScanner database of genotype—phenotype associations. We hope to develop this feature in future versions of the package. The R software environment is available for download from [https://www.r-project.org/]. The MendelianRandomization package can be downloaded from the Comprehensive R Archive Network (CRAN) within R, or directly from [https://cran.r-project.org/web/packages/MendelianRandomization/]. Both R and the MendelianRandomization package are released under GNU General Public Licenses (GPL-2|GPL-3).

Key words: Mendelian randomization, instrumental variable, causal inference, summarized data, two-sample, data parasite

---

Key messages

- The MendelianRandomization software package enables Mendelian randomization analyses to be undertaken using summarized data on genetic associations for a variety of previously published methods.
- The use of this package should lead to more reproducible and more credible causal inferences, due to the range of robust methods included in the package.
- However, a critical and principled approach to Mendelian randomization investigations is required for causal claims to be justified; robust methods can reveal inconsistencies in the analyses, but conclusions that rely on statistical approaches in the absence of biological understanding of the variants included in the analysis will always be speculative.

Introduction

Mendelian randomization is the use of genetic variants as instrumental variables for assessing causal relationships from observational data.[1,2] Increasingly, Mendelian randomization analyses are being performed using summarized data, in particular the associations (beta-coefficients and standard errors) of each genetic variant with the exposure and outcome variables.[3] Such associations are often made publicly available by studies or consortia.[4,5] The previously proposed inverse-variance weighted method gives the same estimates using these summarized data as the well-established two-stage least squares method that uses individual-level data.[6,7]

The inverse-variance weighted (IVW) method provides a consistent estimate of the causal effect of the exposure on the outcome when each of the genetic variants satisfies the assumptions of an instrumental variable.[6] Two further methods have been proposed for providing consistent causal estimates from summarized data for multiple genetic variants under weaker assumptions. These are the MR-Egger method[8] and the weighted median (or median-based) method.[9] Several publications have used some or all of these methods,[10–12] and the use of all three methods is recommended when there are multiple genetic variants to assess robustness of any causal finding to different sets of assumptions.[13] Additionally, several variations on these methods have been proposed, such as the use of robust regression instead of standard linear regression in the IVW or MR-Egger methods, or the penalization of weights from genetic variants with heterogeneous causal estimates.[14]

Software code is available for the implementation of each of these methods. However, this code is currently spread among the appendices of various manuscripts. The MendelianRandomization package brings the code together in a single location and makes implementation of these methods simpler and easier to reproduce, meaning that applied researchers can focus on the most important aspects of a Mendelian randomization investigation, namely the choice of genetic variants and assessment of the instrumental variable assumptions. Code for implementing all of the methods in the paper using the Mendelian Randomization package is provided in the Online Appendix.

Implementation and usage

The data entry function in the MendelianRandomization package is the mr_input() function. The required inputs for this function are the beta-coefficients for associations with the exposure and their standard errors, and the beta-coefficients for associations with the outcome and their standard errors. A minimal example is:

```
MRdata_HDL_CHD    <- mr_input(bx = hdlc, bxse = hdlcse, by = chdlodds, byse = chdloddsse)
```

The variables hdlc, hdlcse, chdlodds and chdloddsse are provided as part of the MendelianRandomization package as example data, and represent the associations of 28 variants with high-density lipoprotein cholesterol (HDL-c) and their standard errors (hdlc and hdlcse), and the associations of the same 28 variants with coronary heart disease (CHD) risk and their standard errors (chdlodds and chdloddsse), taken from Waterworth et al.[15] The mr_input() function ensures that the data are formatted correctly as an MRInput object for processing by one of the estimation or graphical functions. We note that the mr_input() function makes no attempt to align the genetic associations to the same effect allele; this step is critically important, but is left to the user. Genetic associations obtained from PhenoScanner are automatically aligned to the same effect allele.[16] Special care should be taken when it is not clear whether alleles are given for the forward or backward strand, particularly for palindromic variants (C/G or A/T polymorphisms).

The IVW, MR-Egger and weighted median methods are then implemented by taking the MRInput object as an input, and outputting causal estimates and related statistics. For the IVW method, an example is:

```
mr_ivw(MRdata_HDL_CHD)
```

where MRdata_HDL_CHD is as defined above. By default, the mr_ivw() function reports the causal estimate and its standard error, a 95% confidence interval based on a normal approximation, and the residual standard error and associated heterogeneity test statistic, indicating whether the causal estimates from the individual genetic variants are similar to each other or not.[17]

Similar output is provided by the mr_egger() and mr_median() functions, except that the MR-Egger method additionally reports an intercept parameter, representing the average pleiotropic effect of a genetic variant.[8] Several options are available to the user in the ability: to use robust regression instead of standard linear regression in the IVW and MR-Egger methods;[14] to penalize the contribution to the analysis of genetic variants with heterogeneous causal estimates;[14] to use an unweighted rather than a weighted analysis in the median-based method;[9] to make inferences based on a t-distribution rather than a normal distribution; and to report a confidence interval at a different alpha level (for example, a 99% confidence interval). A maximum likelihood method mr_maxlik() that fits a likelihood model to the summarized data, allowing for uncertainty in genetic associations with both the exposure and the outcome,[6] is also implemented in the package.

A further option is the possibility of accounting for correlation between genetic variants in either the IVW or the MR-Egger method. This illustrates one benefit of the package–although the use of generalized weighted linear regression has been proposed for performing the IVW method with correlated variants,[18] it has not been formally proposed for the MR-Egger method,despite the extension being relatively straightforward. The package means that if a methodological advance is made for one method, it can be implemented for multiple methods, and if two methodological advances are made that can interact, their interaction can be explored by setting both options in the package.

The mr_allmethods() function can be used to implement several of the methods offered in this package at once. An example is:

mr_allmethods(MRdata_HDL_CHD,
method = "main")

This implements the IVW, MR-Egger, weighted and simple median methods, and puts the estimates, standard errors and confidence intervals into a table for easy comparison (Figure 1A).

There are two options for providing graphical output from this package. Both of these are implemented by the mr_plot() function. If the argument for the mr_plot() function is an MRAll object created using the mr_allmethods() function:

mr_plot(mr_allmethods(MRdata_HDL_CHD,
method = "main"))

then the result is a static graph illustrating the various causal estimates (Figure 1B). As can be seen in this example, the MR-Egger estimate differs substantially from the IVW and median-based estimates, suggesting some inconsistencies in the instrumental variable assumptions.

Although the IVW method suggests that HDL-c is a causally protective risk factor for CHD, the MR-Egger estimate is compatible with the null, which is more in line with experimental findings.[19]

If the argument for the mr_plot() function is an MRInput object, for example:

mr_plot(MRdata_HDL_CHD)

then the result is an interactive graph of the genetic associations with the outcome against associations with the exposure. By default, error bars for the associations and the IVW estimate are also plotted. By mousing over a data-point, an infobox appears with the name of the genetic variant and its associations with the exposure and with the outcome. This can be used for identifying genetic variants having heterogeneous associations with the outcome compared with other variants, and may be helpful in detecting variants to check for pleiotropy and possible violation of the instrumental variable assumptions.

As an example, the output from the mr_egger() function applied to the HDL-c and CHD data provided as part of the package is presented in Figure 2A. The estimates of the slope and intercept parameters from the regression model are displayed, as well as their standard errors, 95% confidence intervals and associated P-values. The residual standard error from the weighted regression is substantially greater than 1, and the null hypothesis of the heterogeneity test is rejected, indicating that the genetic variants are not all estimating the same causal effect and suggesting the presence of pleiotropy for one or more variants. The I^2 statistic quoted is a measure of weak instrument bias, and is described elsewhere.[20] The mr_plot() command is then applied to the same data, and the MR-Egger estimate is illustrated using the code:

mr_plot(MRdata_HDL_CHD, orientate = TRUE, line
= "egger")

A screenshot of the resulting interactive graph shows the possibility of identifying individual variants from the graph. The orientate = TRUE option ensures that all variants are plotted after orientation to the exposure-increasing allele, as this re-orientation is undertaken in the MR-Egger method.

Finally, although the primary method for inputting data into the package is directly by the user, it is possible to incorporate data from PhenoScanner, a database of genotype—phenotype associations that can be queried using a web browser.[16] Two .csv files from PhenoScanner (v1.1, Little Miss Sunshine) are distributed as part of this package, representing the output from PhenoScanner for four genetic variants with proxies turned on and off. The user can then choose the exposure and outcome from those

A

```
> MRdata_HDL_CHD <- mr_input(hdlc, hdlcse, chdlodds, chdloddsse)
> mr_allmethods(MRdata_HDL_CHD, method="main")

=========================================================
Method            Estimate Std error 95% CI        P-value
---------------------------------------------------------
Simple median     -2.952    0.834    -4.586 -1.318  0.000
Weighted median   -1.822    0.593    -2.985 -0.659  0.002
IVW               -2.375    0.678    -3.704 -1.047  0.002
MR-Egger          -0.492    1.057    -2.563 1.580   0.642
(intercept)       -0.043    0.019    -0.081 -0.005  0.026
---------------------------------------------------------

> mr_plot(mr_allmethods(MRdata_HDL_CHD, method="main"))
```
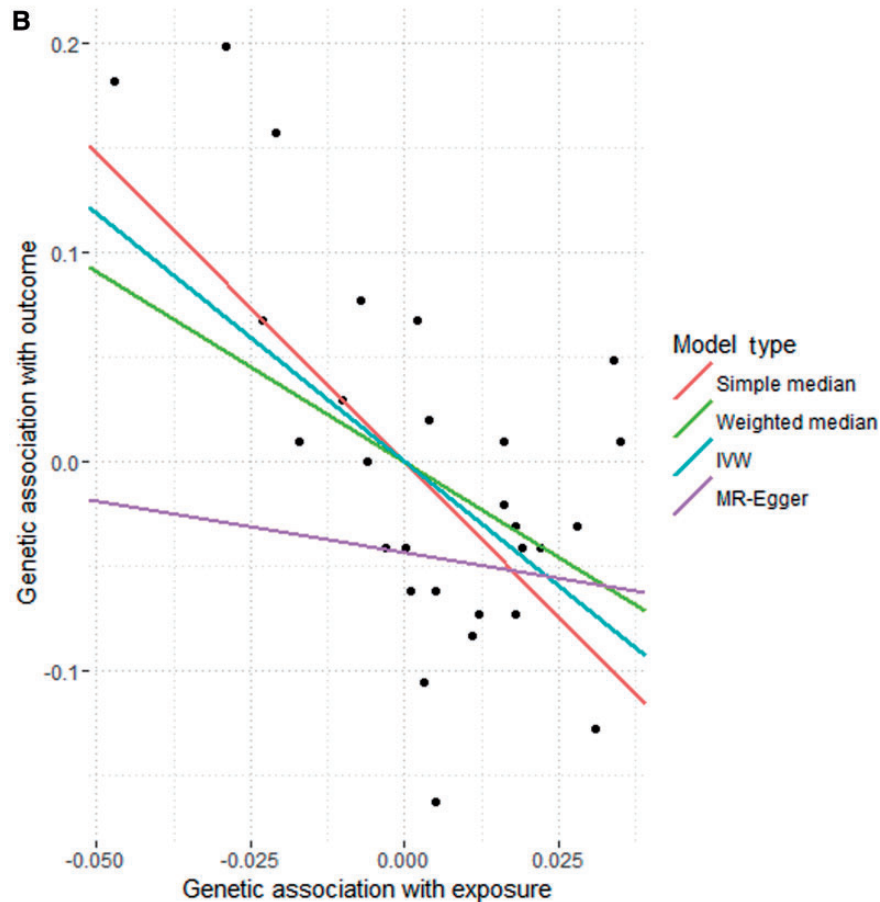
B

Figure 1. A. Code and output from implementation of various Mendelian randomization analysis methods using mr_allmethods() function for analysis of causal effect of high-density lipoprotein-cholesterol (HDL-c) on coronary heart disease (CHD) risk. B. Output from mr_plot() function applied to mr_allmethods() object. Static graph illustrating genetic associations with CHD risk (log odds ratios) against genetic associations with HDL-c (in standard deviation units). Lines represent causal estimates from the different methods. STD, standard.

reported by PhenoScanner, and an MRInput object is created. For example:

path.proxies   <- system.file("extdata", "vitD_snps_PhenoScanner_proxies.csv",
package = "MendelianRandomization")
# this file has been downloaded from PhenoScanner (v1.1)
extract.pheno.csv(

exposure = "log(eGFR creatinine)", pmidE = 26831199
, ancestryE = "European",
outcome = "Asthma", pmidO = 20860503, ancestryO = "European",
rsq.proxy = 0.6, file = path.proxies, snps = "all")

This code takes the file provided by PhenoScanner, and outputs an MRInput object consisting of the genetic associations with the exposure "log(eGFR creatinine)" estimated
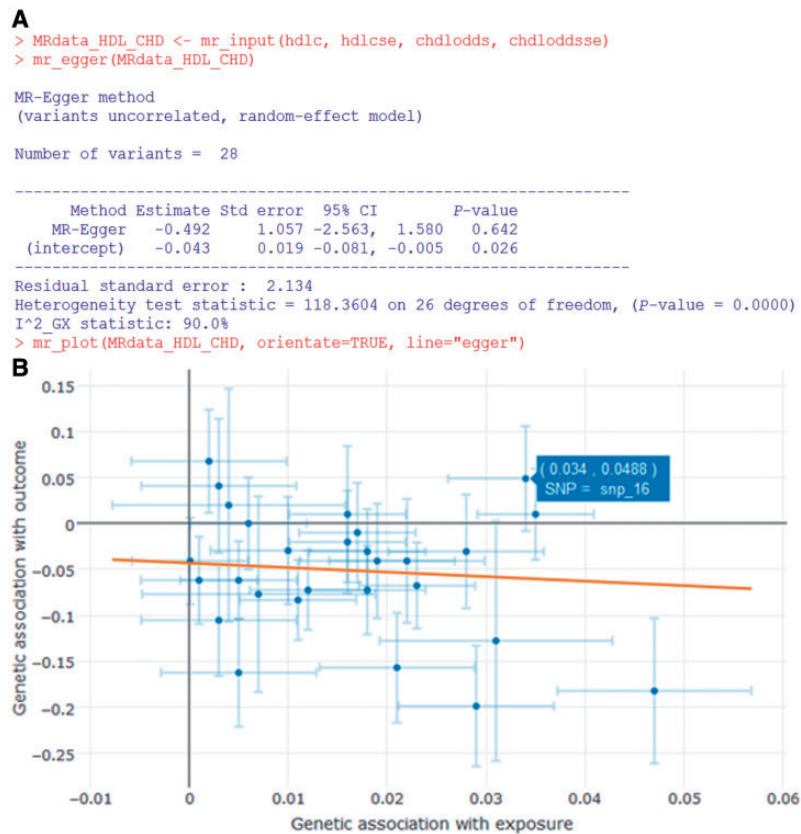
Figure 2. A. Code and output from implementation of MR-Egger method using mr_egger() function for analysis of causal effect of high-density lipo-protein-cholesterol (HDL-c) on coronary heart disease (CHD) risk. B. Output from mr_plot() function applied to mr_input() object. Screenshot of inter-active graph illustrating genetic associations with CHD risk (log odds ratios) against genetic associations with HDL-c (in standard deviation units) with error bars representing 95% confidence intervals for the associations. The variants are all orientated to the HDL-c-increasing allele. The line repre-sents the MR-Egger causal estimate. One of the genetic variants is highlighted by mousing over. The infobox gives the name of the variant (snp_16), and its associations with the exposure and with the outcome. STD, standard.

in Europeans and reported by the study with PubMed ID 26831199,[21] and genetic associations with the outcome "Asthma" estimated in Europeans and reported by the study with PubMed ID 20860503.[22] If genetic associations for one or the other of the exposure and outcome are not reported for a particular variant, the function searches for a proxy variant with correlation measure r^2 greater than the threshold of 0.6 provided. Correlation estimates are provided by PhenoScanner, and can be calculated either for 1000 Genomes or HapMap participants of European descent. In this example, associations with asthma are not available for two variants; they are replaced by associations with suitable proxies. The output from the extract.pheno.csv() function can then be used in the mr_ivw() or similar functions for estimating a causal effect.

The eventual aim is to make this functionality seamless, so that PhenoScanner can be called directly from R as part of the MendelianRandomization package, or as a dependency to the package. Both this package and the PhenoScanner tool are still under development, and we hope that this aim will be achieved in the near future.

Discussion

The MendelianRandomization package makes performing Mendelian randomization analyses using summarized data relatively straightforward. The package is still under development: future additions will include greater integration with the PhenoScanner and MR-Base [www.mr-base.org] platforms to enable increased automation of Mendelian randomization analyses, as well as a wider range of methods such as methods for multivariable Mendelian randomization.[23] The package will also be updated as additional robust methods for performing Mendelian randomization using summarized data are developed.

As a final note of caution, the increasing automation of Mendelian randomization analyses is welcome for reducing the burden on investigators and the capacity for human errors to influence results, but there is a danger that the tools provided may facilitate large numbers of speculative Mendelian randomization analyses to be performed in an unprincipled way [see references 24–26 for some critical comments on Mendelian randomization]. It is important that Mendelian randomization is not performed in a way

that avoids critical thought. In releasing this package, the hope is that it will lead to more comprehensive and more reproducible causal inferences from Mendelian randomization, and not simply add more noise to the literature.

Supplementary Data

The appendix is available as Supplementary data at *IJE* online.

Acknowledgement

Funding

References

1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32**:1–22.
2. Burgess S, Thompson SG. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation.* Boca Raton FL: Chapman & Hall/CRC Press., 2015.
3. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG; EPIC-InterAct Consortium. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol* 2015;**30**:543–52.
4. Burgess S, Timpson NJ, Ebrahim S, Davey Smith G. Mendelian randomization: where are we now and where are we going? *Int J Epidemiol* 2015;**44**:379–88.
5. Thompson JR, Minelli C, Del Greco MF. Mendelian Randomization using Public Data from Genetic Consortia. *Int J Biostat* 2016;**12**. doi: 10.1515/ijb-2015–0074.
6. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013;**37**:658–65.
7. Burgess S, Dudbridge F, Thompson SG. Re: "Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects." *Am J Epidemiol* 2015;**181**:290–91.
8. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015;**44**:512–25.
9. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* 2016;**40**:304–14.
10. Mokry LE, Ross S, Timpson NJ, Sawcer S, Davey Smith G, Richards JB. Obesity and multiple sclerosis: a mendelian randomization study. *PLoS Med* 2016;**13**:e1002053.
11. Tyrrell J, Jones SE, Beaumont R *et al*. Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *BMJ* 2016;**352**:i582.
12. Zhao JV, Schooling CM. Endogenous androgen exposures and ischemic heart disease, a separate sample Mendelian randomization study. *Int J Cardiol* 2016;**222**:940–45.
13. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology* 2017;**28**:30–42.
14. Burgess S, Bowden J, Dudbridge F, Thompson SG. Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. *arXiv* 2016;1606.03729.
15. Waterworth DM, Ricketts SL, Song K *et al*. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol* 2010;**30**:2264–76.
16. Staley JR, Blackshaw J, Kamat MA *et al*. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* 2016. doi:10.1093/bioinformatics/btw373.
17. Del Greco MF, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat Med* 2015;**34**:2926–40.
18. Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods: combining instrumental variables in Mendelian randomization. *Stat Med* 2016;**35**:1880–906.
19. Schwartz GG, Olsson AG, Abt M *et al*. Effects of dalcetrapib in patients with a recent acute coronary syndrome. *N Engl J Med* 2012;**367**:2089–99.
20. Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic. *Int J Epidemiol* 2016, Sep 11. doi:10.1093/ije/dyw220. [Epub ahead of print.]
21. Pattaro C, Teumer A, Gorski M *et al*. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat Commun* 2016;**7**:10023.
22. Moffatt MF, Gut IG, Demenais F *et al*. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med* 2010;**363**:1211–21.
23. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol* 2015;**181**:251–60.
24. Glymour MM, Tchetgen Tchetgen EJ, Robins JM. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol* 2012;**175**:332–39.
25. VanderWeele T, Tchetgen Tchetgen E, Cornelis M, Kraft P. Methodological challenges in Mendelian randomization. *Epidemiology* 2014;**25**:427–35.
26. Burgess S, Butterworth AS, Thompson JR. Beyond Mendelian randomization: how to interpret evidence of shared genetic predictors. *J Clin Epidemiol* 2016;**69**:208–16.