MendelianRandomization v0.5.0: updates to an R package for performing Mendelian randomization analyses using summarized data [version 1; peer review: 3 approved with reservations]

Jim R. Broadbent, Christopher N. Foley, Andrew J. Grant, Amy M. Mason, James R. Staley, Stephen Burgess

1Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge, CB2 0SR, UK
2MRC Biostatistics Unit, University of Cambridge, Cambridge, CB1 8RN, UK

Abstract
The MendelianRandomization package is a software package written for the R software environment that implements methods for Mendelian randomization based on summarized data. In this manuscript, we describe functions that have been added to the package or updated in recent years. These features can be divided into four categories: robust methods for Mendelian randomization, methods for multivariable Mendelian randomization, functions for data visualization, and the ability to load data into the package seamlessly from the PhenoScanner web-resource. We provide examples of the graphical output produced by the data visualization commands, as well as syntax for obtaining suitable data and performing a Mendelian randomization analysis in a single line of code.

Keywords
Mendelian randomization, instrumental variable, summarized data, genetic epidemiology, post-GWAS analysis, causal inference, genetic associations.

Open Peer Review
Reviewer Status
Invited Reviewers
1
2
3

version 2
(revision)
10 Dec 2020
version 1
22 Oct 2020

1. Karl Smith Byrne, The International Agency for Research on Cancer, Lyon, France
2. Wes Spiller, University of Bristol, Bristol, UK
3. Marilyn C. Cornelis, Northwestern University, Chicago, USA

Any reports and responses or comments on the article can be found at the end of the article.
Corresponding author: Stephen Burgess (sb452@medschl.cam.ac.uk)


Competing interests: No competing interests were disclosed.

Grant information: This work is supported by the Wellcome Trust through a Sir Henry Dale Fellowship jointly funded by the Royal Society to SB [204623]. This work was also supported by funding from the National Institute for Health Research (Cambridge Biomedical Research Centre at the Cambridge University Hospitals National Health Service Foundation Trust). The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Broadbent JR et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Broadbent JR, Foley CN, Grant AJ et al. MendelianRandomization v0.5.0: updates to an R package for performing Mendelian randomization analyses using summarized data [version 1; peer review: 3 approved with reservations]
Wellcome Open Research 2020, 5:252 https://doi.org/10.12688/wellcomeopenres.16374.1
Introduction
Mendelian randomization is an epidemiological technique that uses genetic variants to link modifiable risk factors to clinical outcomes\(^1\). The MendelianRandomization package is a software package written for the R software environment\(^3\) that implements methods for Mendelian randomization based on summarized data\(^4\). While the basic functionality and initial features of the package have been discussed previously\(^5\), several functions have been added to the package in recent years. These features can be divided into four categories: robust methods for Mendelian randomization, methods for multivariable Mendelian randomization, functions for data visualization, and the ability to load data into the package seamlessly from the PhenoScanner web-resource. We discuss each of these categories in turn. A list of functions in the package is provided as Table 1.

Methods
Implementation
The initial release of the MendelianRandomization package included four functions for the estimation of causal effects based on summarized genetic data in a univariable (that is, one risk factor) Mendelian randomization framework. These were mr_ivw (inverse-variance weighted method, IVW)\(^6\), mr_egger (MR-Egger method)\(^7\), and mr_median (simple and weighted method methods)\(^8\), and mr_maxlik (maximum likelihood method)\(^6\). Each of these estimation functions takes an MRInput object as input, created using the mr_input command. The syntax is:

\[
\text{mr_ivw(mr_input(ldlc, ldlcse, chdlodds, chdloddsse))}
\]

where ldlc and ldlcse are genetic associations with low-density lipoprotein (LDL) cholesterol and their standard errors for 28 genetic variants as previously reported by Waterworth \textit{et al.}\(^9\), and chdlodds and chdloddsse are genetic associations with coronary heart disease risk for the same variants. These data variables are provided

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mr_input</td>
<td>Data entry for univariable analysis</td>
</tr>
<tr>
<td>mr_mvinput</td>
<td>Data entry for multivariable analysis</td>
</tr>
<tr>
<td>mr_ivw</td>
<td>Inverse-variance weighted (IVW) method</td>
</tr>
<tr>
<td>mr_median</td>
<td>Median method</td>
</tr>
<tr>
<td>mr_egger</td>
<td>MR-Egger method</td>
</tr>
<tr>
<td>mr_maxlik</td>
<td>Maximum likelihood method</td>
</tr>
<tr>
<td>mr_mbe</td>
<td>Mode-based estimation method</td>
</tr>
<tr>
<td>mr_hetpen</td>
<td>Heterogeneity penalized method</td>
</tr>
<tr>
<td>mr_conmix</td>
<td>Contamination mixture method</td>
</tr>
<tr>
<td>mr_lasso</td>
<td>Lasso method</td>
</tr>
<tr>
<td>mr_allmethods</td>
<td>Runs several methods</td>
</tr>
<tr>
<td>mr_mvivw</td>
<td>Multivariable IVW method</td>
</tr>
<tr>
<td>mr_mvmedian</td>
<td>Multivariable median-based method</td>
</tr>
<tr>
<td>mr_mvegger</td>
<td>Multivariable MR-Egger method</td>
</tr>
<tr>
<td>mr_mvlasso</td>
<td>Multivariable lasso method</td>
</tr>
<tr>
<td>mr_plot</td>
<td>Scatter plot</td>
</tr>
<tr>
<td>mr_forest</td>
<td>Forest plot</td>
</tr>
<tr>
<td>mr_funnel</td>
<td>Funnel plot</td>
</tr>
<tr>
<td>mr_loo</td>
<td>Leave-one-out plot</td>
</tr>
<tr>
<td>extract.pheno.csv</td>
<td>Data entry from PhenoScanner .csv file (legacy)</td>
</tr>
<tr>
<td>pheno_input</td>
<td>Data entry from web-based PhenoScanner</td>
</tr>
</tbody>
</table>
with the package. Syntax for the default operation of the mr_egger and mr_median commands (and all the other estimation commands discussed in this section) is identical, although user-options and the output from each method is different. Some methods rely on all variants being uncorrelated; others allow correlated variants using the correl option. All methods allow confidence intervals to be calculated using a t-distribution rather than a normal distribution (distribution = "t-dist") or based on a different significance level (alpha = 0.05 corresponds to a 95% confidence interval). Other options are specific to particular methods.

Operation
The R software environment runs on a wide variety of UNIX platforms, Windows, and MacOS, and requires minimal computer resources (256 kilobytes of RAM is recommended). The package requires R version 3.0.1 or higher.

Use cases
Robust methods for Mendelian randomization
There are two options in the mr_ivw method that represent different robust methods. The robust option performs the IVW method method using robust regression (referred to as MR-Robust). The penalized option performs the IVW method with penalized weights. The syntax is:

```
mr_ivw(mr_input(ldlc, ldlcse, chdlodds, chdloddsse), robust=TRUE)
mr_ivw(mr_input(ldlc, ldlcse, chdlodds, chdloddsse), penalized=TRUE)
```

Other methods implemented in the package are the mode-based method (mr_mbe), the heterogeneity penalized method (mr_hetpen), the lasso method (mr_lasso), and the contamination mixture method (mr_conmix). As for the mr_ivw command, the syntax is:

```
mr_mbe(mr_input(ldlc, ldlcse, chdlodds, chdloddsse))
```

and similarly for the other methods.

The mr_mbe method has options weighting = "weighted" or weighting = "simple", corresponding to weighted and unweighted versions of the method. It also has options stderr = "simple" or stderr = "delta" corresponding to first- and second-order standard errors.

The mr_hetpen method has options prior to set the prior probability of a genetic variant being a valid instrument (default is 0.5), and CIMin, CIMax, and CIStep to allow feasible and efficient calculation of confidence intervals.

The mr_conmix method has options psi to set the value of the standard deviation of the distribution of invalid estimands (that is, how variable are the quantities targeted by genetic variants that are invalid instrumental variables), and CIMin, CIMax, and CIStep as above.

The mr_lasso method has the option lambda to set the tuning parameter in the penalized (lasso) regression model.

These methods were discussed in greater detail and compared in a review of robust methods for Mendelian randomization.

Methods for multivariable Mendelian randomization
Multivariable Mendelian randomization is an extension of the standard Mendelian randomization paradigm to include multiple risk factors in a single analysis model. Methods for multivariable Mendelian randomization take an MRMVInput object as an input, created using the mr_mvinput command. Four functions are included for the estimation of causal effects based on summarized genetic data in a multivariable Mendelian randomization framework. These are mr_miviw (multivariable IVW method), mr_mvegger (multivariable MR-Egger method), and mr_mvmedian (multivariable median-based method), and mr_mvlasso (multivariable lasso method). The syntax is:

```
mr_miviw(mr_mvinput(bx = cbind(ldlc, hdlc, trig),
             bxse = cbind(ldlcse, hdlcse, trigse),
             by = chdlodds, byse = chdloddsse))
```
where hdlc and hdlcse are genetic associations with high-density lipoprotein (HDL) cholesterol and their standard errors, and trig and trigse are genetic associations with triglycerides and their standard errors for the same 28 variants. Again, these data variables are provided with the package. Syntax for the mr_mvegger, mr_mvmedian, and mr_mvlasso commands is identical. As in the univariable case, the mr_mviw command can be implemented using robust regression with the robust = TRUE option\(^{19}\). The mr_mviw and mr_mvegger methods have a correl option to allow for correlated variants. All methods have distribution and alpha options as discussed above. The mr_mvlasso method has the lambda option to set the penalization parameter as in the univariable case.

Functions for data visualization
The initial release of the MendelianRandomization package included two options for data visualization, both implemented using the mr_plot function. Application of the mr_plot function to an MRInput object gave an interactive scatter plot of the genetic associations together with a line representing the IVW estimate. Genetic associations with the risk factor are plotted on the horizontal axis, and genetic associations with the outcome on the vertical axis. Application of the mr_plot function to an MRAll object plotted a similar (although non-interactive) scatter plot of the genetic associations together with lines representing the estimates from various methods. An MRAll object is created by the mr_allmethods function, which returns estimates from various estimation methods.

We have added functionality so that the mr_plot function can now be applied to an MRMVInput object. In this case, we still plot the estimated genetic associations with the outcome on the vertical axis. On the horizontal axis, we plot predicted genetic associations with the outcome. These are fitted values from the multivariable IVW method, which regresses the genetic associations with the outcome on the genetic associations with the risk factors. Horizontal error bars represent confidence intervals for these fitted values. These reflect uncertainty in the multivariable IVW estimates, but not in the genetic associations with the risk factors, which are assumed to be known without error. A diagonal line is plotted with gradient 1 to help the detection of outliers, which may be pleiotropic variants. The syntax is:

```
mr_plot(mr_mvinput(bx = cbind(ldlc, hdlc, trig), # Figure 1
bxse = cbind(ldlcse, hdlcse, trigse),
by = chdlodds, byse = chdloddsse))
```

In the example of Figure 1, we additionally set the option interactive = FALSE to produce a non-interactive version of this plot. In updating the package, we have added several additional functions for data visualization.

![Figure 1. Scatter plot created by mr_plot command applied to a MRMVInput object.](image)

Estimated genetic associations with the outcome (vertical axis) are plotted against predicted associations with the outcome from the multivariable inverse-variance weighted method (horizontal axis). Error bars are 95% confidence intervals, and the diagonal line has gradient 1.
The default implementation of the `mr_forest` function plots the variant-specific estimates in a forest plot, with the pooled estimate from the IVW method at the bottom (Figure 2a). The variant-specific estimates are the ratio estimates from each genetic variant in turn. This plot allows the user to investigate heterogeneity in the variant-specific estimates, which indicates potential pleiotropy in the analysis. The `mr_forest` function can also be used to plot estimates from different methods, either in addition to the variant-specific estimates or without them (Figure 2b):

```r
mr_forest(mr_input(ldlc, ldlcse, chdlodds, chdloddsse)) # Figure 2A
mr_forest(mr_input(ldlc, ldlcse, chdlodds, chdloddsse), # Figure 2B
  snp_estimates=FALSE,
  methods = c("ivw", "median", "wmedian", "egger", "maxlik", "mbe", "conmix"))
```

(For presentation purposes, in this and subsequent figures we provide plots for the first 9 variants in the package only.) The `mr_funnel` function is similar, except that the variant-specific estimates are plotted against their precision (that is, the reciprocal of their standard error). This plot also enables the user to investigate heterogeneity in the variant-specific estimates (Figure 3):

```r
mr_funnel(mr_input(ldlc, ldlcse, chdlodds, chdloddsse)) # Figure 3
```

The `mr_loo` function allows the user to investigate sensitivity of the IVW estimate to individual data points. This is implemented by calculating the IVW estimate omitting each variant from the analysis in turn (loo stands for ‘leave one out’). The IVW estimate based on all the variants is also plotted for reference (Figure 4):

```r
mr_loo(mr_input(ldlc, ldlcse, chdlodds, chdloddsse)) # Figure 4
```

Output from each of these commands is a `ggplot` object, and so basic graphical parameters can be changed using functions from the ggplot2 package. For example, the horizontal axis can be set to run from −5 to +5 using the following code:

```r
library(ggplot2)
forest = mr_forest(mr_input(ldlc, ldlcse, chdlodds, chdloddsse))
forest2 = forest + coord_cartesian(xlim=c(-5,5))
forest2
```

![Figure 2. Forest plots created by mr_forest command.](image)

Left panel: comparison of variant-specific estimates plus inverse-variance weighted (IVW) estimate (default options). Right panel: comparison of estimates from different methods with variant-specific estimates switched off. Points represent estimates and horizontal error bars are 95% confidence intervals (CI).
Figure 3. **Funnel plot created by mr_funnel command.** Points represent variant-specific estimates and horizontal error bars are 95% confidence intervals (CI).

Figure 4. **Leave-one-out plot created by mr_loo command.** Points represent estimates from the inverse-variance weighted (IVW) method, omitting the variant indicated. Horizontal error bars are 95% confidence intervals (CI).
Loading data from PhenoScanner
The initial release of the MendelianRandomization package included a function called `extract.pheno.csv`. This function took a .csv file previously downloaded by the user from the PhenoScanner webtool (http://www.phenoscaner.medschl.cam.ac.uk/) and converted the file into an `MRInput` object, extracting the relevant genetic associations with the risk factor and outcome. PhenoScanner\(^22,23\) is a database of genetic associations that contains over 65 billion associations for over 150 million unique genetic variants, including genetic associations reported by major consortia, as well as those for the UK Biobank study reported by Ben Neale’s team (http://www.nealelab.is/uk-biobank).

The `extract.pheno.csv` function is no longer maintained; however, it has been superseded by the `pheno_input` command, which calls PhenoScanner directly from R and creates an `MRInput` object. Using this command, the entire workflow of a Mendelian randomization analysis can be performed in a single line of code. For example:

```r
mr_ivw(pheno_input(snps=c("rs12916", "rs2479409", "rs217434", "rs1367117", "rs4299376", "rs629301", "rs4420638", "rs6511720"),
exposure = "Low density lipoprotein", pmidE = "24097068", ancestryE = "European",
outcome = "Coronary artery disease", pmidO = "26343387", ancestryO = "Mixed"))
```

This code first extracts data on eight genetic variants (their ‘rsid’ identifiers are listed above), and creates an `MRInput` object using the genetic associations with “low density lipoprotein” taken from the study with PubMed ID 24097068\(^24\) in individuals of European descent as the summarized associations with the risk factor, and genetic associations with “coronary artery disease” taken from the study with PubMed ID 26343387\(^25\) in a mixed ancestry sample as the summarized associations with the outcome. The triplet of trait name, PubMed ID, and ancestry is necessary to uniquely identify the correct dataset for genetic associations, as some publications report associations with multiple traits, or associations with the same trait in different ancestry groups. While the above code then implements the IVW method on this `MRInput` object, any other estimation or data visualization command that takes an `MRInput` object as input could be applied to the output of the `pheno_input` function.

**Summary**
In summary, the MendelianRandomization package has added a number of features since its initial release: to implement various robust estimation methods, to implement methods for multivariable Mendelian randomization, to enable a greater range of data visualization options, and to facilitate data entry. We conclude with the same warning that we stated at the end of the manuscript accompanying the initial package release: while this software simplifies the operational aspects of a Mendelian randomization, the truly difficult parts of an analysis are choosing sensible risk factors and outcomes, selecting genetic variants that are plausible instrumental variables, performing a reasonable range of analyses, and interpreting the results with care and caution\(^26\). Software code for these aspects of an analysis cannot be written\(^27\).

**Data availability**
Underlying data
All data used in this article are distributed in the software package described, or can be freely downloaded using commands in the software package that are detailed in the text of the article.

**Software availability**
The MendelianRandomization package is available via the Comprehensive R Archive Network (CRAN)

The software package is available here: [https://cran.r-project.org/web/packages/MendelianRandomization/index.html](https://cran.r-project.org/web/packages/MendelianRandomization/index.html).

Source code is available from GitHub: [https://github.com/sb452/MendelianRandomization/tree/v0.5.0](https://github.com/sb452/MendelianRandomization/tree/v0.5.0)


Software license: GPL-2 | GPL-3.

**Acknowledgments**
We thank users of the package for helpful feedback that has guided the development of the package.
References


Open Peer Review

Current Peer Review Status:  ?  ?  ?

Version 1

Reviewer Report 06 November 2020

https://doi.org/10.21956/wellcomeopenres.18010.r41072

© 2020 Cornelis M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Marilyn C. Cornelis
Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Broadbent et al. summarize updates made to their MendelianRandomization R package. The package is widely used and thus a detailed description of updates is justified and warranted. I have only very minor suggestions.

1. The R package manual (updated September 30, 2020) details all functions described in this paper. The introduction or summary should briefly mention what the current paper adds above and beyond the R package manual.

2. The paper can use some minor re-organizing or heading revisions. Currently, “Introduction”, “Methods”, “Use cases”, “Summary”, “Data availability” and “Software availability” are the major sections. The “Use cases” is well written and describes features in four categories which the authors introduce in the beginning and constitutes the key/main section of the paper. The “Methods” section, however, seems out of place or mislabeled for this descriptive paper. Perhaps use “Software overview”?

3. Table 1. Please annotate functions that are new and not discussed in reference 5.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets
and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** genetic epidemiology, nutrition, cardiometabolic traits, aging traits, coffee, caffeine, metabolomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

**Author Response 20 Nov 2020**

**Stephen Burgess**, University of Cambridge, Cambridge, UK

**Reviewer 3: Marilyn C. Cornelis**

**C0.** Broadbent et al. summarize updates made to their MendelianRandomization R package. The package is widely used and thus a detailed description of updates is justified and warranted. **I have only very minor suggestions.**
- We thank the reviewer for their positive view of this submission.

**C1.** The R package manual (updated September 30, 2020) details all functions described in this paper. The introduction or summary should briefly mention what the current paper adds above and beyond the R package manual.
- The reviewer is correct that all of the material in the current paper is also available through the R package manual. However, there are reasons why we believe that a publication is worthwhile. First, this paper presents the contents of the package in a clearer and more holistic way than the manual. Secondly, it provides examples of usage and interpretation of the output. Thirdly, it creates a citation in the scientific literature so that we can track usage of the package.
- We have added text to the introduction to more clearly define the scope of this paper. We have also referenced the manual more clearly within the text, for example to obtain lists of the input options and output slots for each method.

**C2.** The paper can use some minor re-organizing or heading revisions. Currently, “Introduction”, “Methods”, “Use cases”, “Summary”, “Data availability” and “Software availability” are the major sections. The “Use cases” is well written and describes features in four categories which the authors introduce in the beginning and constitutes the key/main section of the paper. The “Methods” section, however, seems out of place or mislabeled for this descriptive paper. Perhaps use “Software overview”?  
- Thank you for the suggestion. However, the format of the article is set by the journal to ensure homogeneity of papers across this article class. Hence we do not have complete
freedom to determine the structure of the article. We have reviewed the guidelines and re-arranged where possible according to the journal guidelines, but the original structure largely remains.

C3. **Table 1. Please annotate functions that are new and not discussed in reference 5.**
- We have added annotation of those functions that are new and those that are updated since the publication of Yavorska and Burgess 2017 (now reference 7).

**Competing Interests:** No competing interests were disclosed.

---

**Reviewer Report 05 November 2020**

https://doi.org/10.21956/wellcomeopenres.18010.r41070

© 2020 Spiller W. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Wes Spiller**
Department of Population Health Sciences, University of Bristol, Bristol, UK

This article demonstrates how a range of valuable Mendelian randomization approaches can be easily implemented using the MendelianRandomization R Package. It describes substantial additions to the existing software, which I believe would be of great value to researchers. This is especially true of integrating data from the PhenoScanner web-resource, which can be invaluable in identifying avenues for pleiotropic bias.

The material presented in the article is technically sound, though I believe it could be improved with minor revisions.

1. In the methods section, some of the wording is at times confusing, for example: “simple and weighted method methods)”.

2. There is an understandably difficult balance between summarising methods previously described, and providing sufficient detail so as for researchers to understand how options relate to methods implemented. For example, while the correl option is referenced several times, there is no indication of appropriate usage (such as providing a suitable correlation matrix). Especially in the *Use cases* section, the details provided at times read like the help file for various functions when using R. This shifts from providing insufficient detail overall for using such methods (specifically novel options provided in the package) with confidence, to providing at times a puzzling focus on elements such as the standard error options for the modal estimator. If these options are important enough to be included in the manuscript, then this would warrant more description of their usage.

3. A brief overview of multivariable MR would potentially be helpful, as there is often confusion over whether only SNPs correlated with all exposures should be used, or a set of
SNPs where each SNP is correlated with at least one exposure included in the model. I appreciate this has been explained in previous publications, but the clarification would be welcome.

Having performed an example analysis using this software as part of the review process, I commend the authors on providing helpful vignettes and documentation guiding the analysis process in the R package itself.

It is also very encouraging to see issues surrounding the appropriate selection of exposures and outcomes, as well as interpretation of results, afforded attention at times lacking in software showcasing novel statistical approaches.

With the above comments in mind, I believe the MendelianRandomization R package represents a valuable addition to a growing body of MR software ensuring emerging methods are easily implemented in applied research.

**Is the rationale for developing the new software tool clearly explained?**
Yes

**Is the description of the software tool technically sound?**
Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**
Partly

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**
No

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**
Yes

*Competing Interests:* No competing interests were disclosed.

*Reviewer Expertise:* Genetic Epidemiology, Applied Statistics, Mendelian Randomization, Transethnic MR

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

**Author Response 20 Nov 2020**

**Stephen Burgess,** University of Cambridge, Cambridge, UK
**Reviewer 2: Wes Spiller**

B0. This article demonstrates how a range of valuable Mendelian randomization approaches can be easily implemented using the MendelianRandomization R Package. It describes substantial additions to the existing software, which I believe would be of great value to researchers. This is especially true of integrating data from the PhenoScanner web-resource, which can be invaluable in identifying avenues for pleiotropic bias.

- We thank the reviewer for their positive view of this submission.

The material presented in the article is technically sound, though I believe it could be improved with minor revisions.

B1. In the methods section, some of the wording is at times confusing, for example: “simple and weighted method methods)".

- Thank you for pointing out that error. The text should read “simple and weighted median methods”.

B2. There is an understandably difficult balance between summarising methods previously described, and providing sufficient detail so as for researchers to understand how options relate to methods implemented. For example, while the correl option is referenced several times, there is no indication of appropriate usage (such as providing a suitable correlation matrix). Especially in the Use cases section, the details provided at times read like the help file for various functions when using R. This shifts from providing insufficient detail overall for using such methods (specifically novel options provided in the package) with confidence, to providing at times a puzzling focus on elements such as the standard error options for the modal estimator. If these options are important enough to be included in the manuscript, then this would warrant more description of their usage.

- We appreciate the sentiment of the reviewer's comment. As per the response to A3, our preference is for this paper to be brief and to illustrate the options available rather than to wade into deep discussions about the properties of the various methods or the reasoning for choosing between the different options available. Hence, we are happy to provide the various options that can be chosen, but we would rather not discuss the merits and demerits of the options at length – these are already addressed elsewhere. This is now explicitly stated in the Introduction.

- If there are other specific examples where the reviewer feels that additional detail is required, then we would be happy to consider adding these. As for the specific examples that the reviewer calls out: 1) we are not aware of any novel options that are provided in the package – all the methods and options have been introduced elsewhere (see citations in the new Table 2); 2) we have added explanation and an example of the “correl” option; 3) the discussion of standard errors for the mode-based estimation method is one sentence. We do not think this is excessive. The original version of the mode-based estimation method provided by the authors gave 12 different estimates based on different options for the method (3 options for the phi parameter, weighted and unweighted, simple and second-order standard errors): we have condensed these down into the three options discussed in the submission.

B3. A brief overview of multivariable MR would potentially be helpful, as there is often confusion over whether only SNPs correlated with all exposures should be used, or a set of SNPs where each
SNP is correlated with at least one exposure included in the model. I appreciate this has been explained in previous publications, but the clarification would be welcome.

- As per the response to A4, we have added a short description of multivariable MR to the manuscript. In particular, we clarify that genetic associations are required for all variants with each risk factor.

B4. Having performed an example analysis using this software as part of the review process, I commend the authors on providing helpful vignettes and documentation guiding the analysis process in the R package itself.

- We thank the reviewer for his positive feedback on the vignettes and documentation.

B5. It is also very encouraging to see issues surrounding the appropriate selection of exposures and outcomes, as well as interpretation of results, afforded attention at times lacking in software showcasing novel statistical approaches.

- We thank the reviewer for his positive feedback on our discussion of practical implementation of the methods and interpretation of the results.

B6. With the above comments in mind, I believe the MendelianRandomization R package represents a valuable addition to a growing body of MR software ensuring emerging methods are easily implemented in applied research.

- We thank the reviewer for his comments, which have strengthened and clarified this manuscript.

**Competing Interests:** No competing interests were disclosed.
2. Additionally, I would provide greater clarity for ‘summarised data’; perhaps, summary level data for the association of SNPs with a trait, or facilitate a two-sample MR design.

3. For the description of methods, I would firstly comment that the degree of customisability of these methods (mr_mbe, for example), is rather impressive and to be lauded. However, I would perhaps add one or two sentences to each of the method paragraphs further clarifying the use of these options. For example, the description of the priors of SNPs with causal effects may be quite unclear for those coming new to these methods. While I acknowledge that ultimately the papers behind each method should be read by users of the software, one or two extra helper sentences for each method would be an easy way to improve the ease of use for your software.

4. Similarly, it would be useful to add some extra description to the MVMR introduction to explicitly specify what is going on; for example that the basic MVMR is just a regression with first order weights \( y \sim x_1 + x_2 \) to make it less intimidating for users.

5. Is there a facility to report Cochran’s or ruckers Q within the software? The authors discuss the value of visually inspecting plots for heterogeneity but it would wonderful if it were possible to return the statistics for this.

6. I also note that it is enjoyable that the plots can be directly edited as ggplot objects.

7. Finally, it would be excellent to have a brief paragraph explaining the format of the data frame of results being reported. It can be quite frustrating when using a package to have to manually explore the data object to try and sleuth out what exactly is present and what it means. I expect this is in the vignette in R but I recommend a summary of this here, perhaps, even only for the components reported in common for all MR methods.

In general, this is a welcome update that is described in a clear manner. With the minor updates I suggest above it will be a very useful tool for genetic epidemiology.

**Is the rationale for developing the new software tool clearly explained?**
Yes

**Is the description of the software tool technically sound?**
Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**
Partly

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**
No

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genetic and Molecular Cancer Epidemiology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 20 Nov 2020

Stephen Burgess, University of Cambridge, Cambridge, UK

Reviewer 1: Karl Smith Byrne
A0. This article describes an invaluable addition to the set of tools facilitating MR analyses, and provides a more up-to-date suite of methods.
- We thank the reviewer for their positive view of this submission.

I believe the paper, however, would benefit from some minor revisions:

A1. In the introduction I advise the authors to amend the specification that MR is used on ‘modifiable’ risk factors and ‘clinical’ outcomes as MR can and often is conducted for risk factors that are not clearly modifiable, such as proteins, and similarly MR may be run on non-clinical outcomes.
- We have modified the language as suggested.

A2. Additionally, I would provide greater clarity for ‘summarised data’; perhaps, summary level data for the association of SNPs with a trait, or facilitate a two-sample MR design.
- We have added a brief description of what is meant by “summarized data”.

A3. For the description of methods, I would firstly comment that the degree of customisability of these methods (mr_mbe, for example), is rather impressive and to be lauded. However, I would perhaps add one or two sentences to each of the method paragraphs further clarifying the use of these options. For example, the description of the priors of SNPs with causal effects may be quite unclear for those coming new to these methods. While I acknowledge that ultimately the papers behind each method should be read by users of the software, one or two extra helper sentences for each method would be an easy way to improve the ease of use for your software.
- We appreciate the sentiment of the reviewer. However, as the reviewer states, detailed descriptions of these methods are already available in the literature, and a paper reviewing the methods has already been written (Slob and Burgess, 2020). Our strong preference is that this paper remains short and focuses on the software package, rather than deviating too strongly into technical detail. To this end, we have included a brief description of the methods in a separate table (Table 2), so that this information is available to the interested reader, but it does not add to the length of the text. The aim of this paper is now more clearly stated in the introduction.

A4. Similarly, it would be useful to add some extra description to the MVMR introduction to
explicitly specify what is going on; for example that the basic MVMR is just a regression with first order weights \( y \sim x_1 + x_2 \) to make it less intimidating for users.

- We have added a brief description of multivariable Mendelian randomization (MVMR), and describe how the multivariable IVW method is implemented using multivariable regression.

**A5.** Is there a facility to report Cochran's or Rücker's Q within the software? The authors discuss the value of visually inspecting plots for heterogeneity but it would be wonderful if it were possible to return the statistics for this.

- The `mr_ivw` and `mr_egger` functions provide the “Heterogeneity test statistic” as part of the outcome. This is the same as Cochrane's Q statistic for the IVW method and Rücker's Q statistics for the MR-Egger method. This is clarified in the help files for the `mr_ivw` and `mr_egger` functions, and now in the manuscript.

**A6.** I also note that it is enjoyable that the plots can be directly edited as ggplot objects.

- That is wonderful!

**A7.** Finally, it would be excellent to have a brief paragraph explaining the format of the data frame of results being reported. It can be quite frustrating when using a package to have to manually explore the data object to try and sleuth out what exactly is present and what it means. I expect this is in the vignette in R but I recommend a summary of this here, perhaps, even only for the components reported in common for all MR methods.

- We understand the reviewer's request, and sympathize with this problem. As suggested, we have provided details of some key outputs that are common to all (or at least, most) methods. A full description of the output of each method is readily available in the package documentation.

- For example, typing `?mr_ivw` opens the description of the `mr_ivw` function. Under the subheading “Value”, there is a clear list of the output for the IVW method. In this case, there are 15 slots: Model, Exposure, Outcome, Correlation, Robust, Penalized, Estimate, StdError, CI Lower, CI Upper, Alpha, Pvalue, SNPs, RSE, and Heter.Stat. We do not think it would be helpful to the reader to provide a detailed list of every slot for every method in this manuscript, but have added an explanation of how to access this information.

**A8.** In general, this is a welcome update that is described in a clear manner. With the minor updates I suggest above it will be a very useful tool for genetic epidemiology.

- We thank the reviewer for his comments, which have strengthened and clarified this manuscript.

**Competing Interests:** No competing interests were disclosed.