Metadata of the chapter that will be visualized online

Chapter Title	A Primer in Mendelian Randomization Methodology with a Focus on Utilizing Published Summary Association Data	
Copyright Year	2018	
Copyright Holder	Springer Science+Business Media, LLC, part of Springer Nature	
Author	Family Name	**Dimou**
	Particle	
	Given Name	**Niki L.**
	Suffix	
	Division	Department of Hygiene and Epidemiology
	Organization	University of Ioannina School of Medicine
	Address	Ioannina, Greece
Corresponding Author	Family Name	**Tsilidis**
	Particle	
	Given Name	**Konstantinos K.**
	Suffix	
	Division	Department of Hygiene and Epidemiology
	Organization	University of Ioannina School of Medicine
	Address	Ioannina, Greece
	Division	Department of Epidemiology and Biostatistics, School of Public Health
	Organization	Imperial College London
	Address	London, UK
	Email	ktsilidi@cc.uoi.gr

Abstract	Mendelian randomization (MR) is becoming a popular approach to estimate the causal effect of an exposure on an outcome overcoming limitations of observational epidemiology. The advent of genome-wide association studies and the increasing accumulation of summarized data from large genetic consortia make MR a powerful technique. In this review, we give a primer in MR methodology, describe efficient MR designs and analytical strategies, and focus on methods and practical guidance for conducting an MR study using summary association data. We show that the analysis is straightforward utilizing either the MR-base platform or available packages in R. However, further research is required for the development of specialized methodology to assess MR assumptions.

AU1

Keywords (separated by '-')	Mendelian randomization - Summarized data - Instrumental variable - Causal inference

Chapter 13

A Primer in Mendelian Randomization Methodology with a Focus on Utilizing Published Summary Association Data

Niki L. Dimou and Konstantinos K. Tsilidis

Abstract

Mendelian randomization (MR) is becoming a popular approach to estimate the causal effect of an exposure on an outcome overcoming limitations of observational epidemiology. The advent of genome-wide association studies and the increasing accumulation of summarized data from large genetic consortia make MR a powerful technique. In this review, we give a primer in MR methodology, describe efficient MR designs and analytical strategies, and focus on methods and practical guidance for conducting an MR study using summary association data. We show that the analysis is straightforward utilizing either the MR-base platform or available packages in R. However, further research is required for the development of specialized methodology to assess MR assumptions.

Key words Mendelian randomization, Summarized data, Instrumental variable, Causal inference

1 Introduction

Mendelian randomization (MR) is a technique that uses genetic variants to make causal inferences about the effect of an exposure on an outcome. MR is a special case of the instrumental variable (IV) methodology, initially introduced in econometrics, where genetic variants are used as IVs [1]. This approach is based on the principle of the random assignment of an individual's genotype from his or her parental genotypes that occurs at conception, and is analogous to the random allocation of a treatment in randomized controlled trials. The reason for utilising the MR approach is to overcome residual confounding, reverse causation or exposure measurement error, which occur frequently in observational studies and may bias their results [2]. Genetic variants are, in general, not associated with environmental confounders. Reverse causality is not an issue in genetic epidemiology, as the genotype does not usually change through life. Finally, genotypes may index the tendency for

Evangelos Evangelou (ed.), *Genetic Epidemiology: Methods and Protocols*, Methods in Molecular Biology, vol. 1793,
https://doi.org/10.1007/978-1-4939-7868-7_13,

lifetime concentrations of an environmental exposure and may thus circumvent exposure measurement error that is frequent when exposures are evaluated at one point in time in observational studies [3].

The use of MR is growing rapidly in popularity during the last 5 years [4]. MR studies have demonstrated causal effects of obesity and low-density lipoprotein cholesterol with cardiovascular disease, but lack of causal effects for high-density lipoprotein cholesterol and C-reactive protein [5–7]. Recent studies have also identified a number of potential causal associations between obesity and related metabolic traits with several cancers [8–11].

Many review articles on MR exist, which include descriptions of MR assumptions and evaluation methods [12–14], commentaries on available study designs [15], statistical models for deriving a causal effect [16, 17], and guidance for the reporting of the MR findings [18, 19]. Many of the aforementioned issues have been recently presented in a unified framework [20]. We seek to complement existing literature by contributing a review article to guide an interested reader to conduct an MR study with publicly available data. We begin with a general description of IV assumptions, MR statistical estimators and study designs when individual level data are available. We then switch to specific MR approaches used when summarized data are available. In particular, we describe general guidelines on the selection of IVs, statistical approaches for the estimation of causal effects and the assessment of IV assumptions. We proceed with demonstration of the MR-base platform [21], an online database of summary genetic association data and a tool to perform MR analyses, as well as popular R packages. We close with a discussion of the advantages and limitations of the MR approach.

2 IV Assumptions in MR

MR studies must fulfil IV assumptions. These assumptions are that (1) the genetic variant (G) is associated with the exposure (X); (2) the genetic variant is not associated with any confounder (U) of the exposure-outcome association; and (3) the genetic variant is conditionally independent of the outcome (Y) given the exposure and confounders (Fig. 1) [22, 23].

For the first assumption to hold, it necessitates the use of genetic variants as IVs that are strongly associated with the exposure. This is the only assumption that can be formally tested and could be satisfied if genome-wide statistically significant variants are selected as candidate IVs. The second assumption is violated if the IVs are associated with confounders, although genes are not in general correlated with environmental confounders. The third assumption implies that all causal pathways from the genetic variants to the outcome pass through the exposure, and that there are

no alternative pathways [23]. The second and third assumptions are not testable, but we could get some intuition of their validity based on existing biological knowledge. The second assumption is violated when population stratification exists, which is a type of confounding due to different ancestry. It often occurs in genetic epidemiology, when the population under analysis can be decomposed into different ancestries that have different allele frequencies for the genetic variant under study and different risks for the outcome under study. The third assumption can be violated by numerous phenomena including pleiotropy, linkage disequilibrium (LD), population stratification, and gene-environment or gene-gene interactions. There is evidence in the literature that several genetic variants have pleiotropic effects, which means that they are associated with several different phenotypes. Pleiotropy is often categorized as "balanced" if the average pleiotropic effects of the IVs that contribute in an MR analysis are zero or "directional" eitherwise. LD refers to the phenomenon that some genetic variants are jointly inherited due to their physical proximity on a chromosome. Overall, there is no way to prove that the second and third MR assumptions definitively hold. However, it is often possible to find empirical evidence suggesting that the putative IVs are invalid. One of the best ways to indirectly evaluate MR assumptions is if there is high reproducibility of the MR causal estimates in different studies.

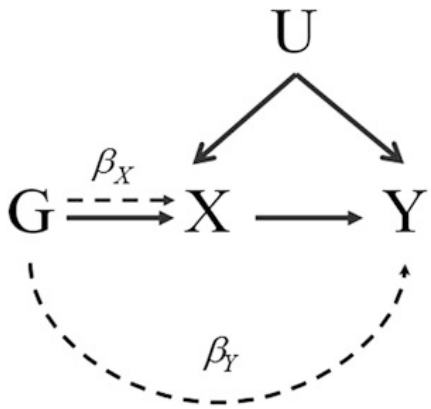

Fig. 1 Directed acyclic graph (DAG) for the instrumental variable assumptions in Mendelian randomization. The exposure (*X*) is causally associated with the outcome (*Y*) if: (1) the genetic variant (*G*) is associated with *X*; (2) *G* is independent of any confounding factors (*U*), and (3) there is no association between *G* and *Y*, except through *X*. The dashed lines represent the coefficient from the regression of the outcome on *G* (β_Y) and the coefficient from the regression of the exposure on *G* (β_X)

3 MR Estimators Using Individual Level Data

Several methods have been proposed for the estimation of the causal effect of the exposure on both continuous and binary outcomes using IVs, which include the ratio of the regression coefficients method, several two-stage methods, likelihood-based methods and semi-parametric models. A thorough overview of these methods as well as guidelines for their use have been recently published [17].

The ratio method, also known as the Wald method [24], is the simplest approach. The causal effect can be expressed as a ratio with nominator the coefficient from the regression of the outcome on the IV (β_Y) and denominator the coefficient from the regression of the exposure on the IV (β_X) (Fig. 1). Confidence intervals for the ratio estimator can be calculated using a normal approximation, which however may be suboptimal when normality assumptions are violated particularly when small sample sizes are available. Alternatively, one could use the Fieller's theorem (https://sb452.shinyapps.io/fieller/) [3, 25], bootstrapping [26], Anderson-Rubin test statistic [27] or the conditional likelihood ratio test statistic [28]. This approach can be extended to account for binary outcomes by simply employing log-linear or logistic regression models. The ratio method for calculating the MR estimator can be performed for single IVs. The sample size required under the ratio method for making causal inferences can be very large [29], and methods that can incorporate multiple IVs are preferable.

Two-stage methods are widely used in MR and are formulated in two separate regression stages: the first-stage involves a regression of the exposure on the IVs, and the second-stage a regression of the outcome on the predicted values of the exposure from the first stage. The first-stage regression model can incorporate multiple IVs. The causal estimate is the second-stage regression coefficient. However, although this estimate is valid, the standard error is estimated imprecisely as the variability of the first-stage regression is not accounted for. Thus, an alternative formula for the calculation of the variance of the two-stage estimator has been presented when the size of the error terms do not differ across values of the independent variables (homoscedasticity assumption) [30], or alternatively robust standard errors can be reported. Binary outcomes can also be accounted for by using log-linear or logistic regression models in the second-stage regression, although these methods have been criticized as the residuals from the second-stage regression may be correlated with the IVs [30]. Under a similar perspective, the "control function" estimator [31] follows the same principle as the two-stage estimator but also includes the estimated residuals from the first-stage regression in the second-stage.

As already pointed out, two-stage methods do not account for the variance of the first-stage regression, and likelihood-based methods are preferable since the two stages are performed simultaneously. These involve full information maximum likelihood (FIML) or limited information maximum likelihood (LIML) models [32] and Bayesian methods [33]. Finally, there are also semi-parametric methods, which make a parametric assumption for the model relating the exposure to the outcome, but make no assumption on the distribution of the errors. These methods include the generalized method of moments (GMM) [34, 35], continuous updating estimator (CUE) [36] and structural mean models (SMM) [37–39]. However, a drawback with all these semi-parametric models is that a unique causal estimate may not be estimated when binary outcomes are assessed. It should also be noted that with a single IV, causal estimates obtained via the ratio, two-stage methods, LIML, GMM and SMM coincide [17].

4 MR Study Designs Using Individual Level Data

When data on the IV(s), exposure, and outcome are available for all participants in a single sample, estimation of the causal effect is straightforward using the appropriate method(s) described in the previous sections. However, in practice, in the era of large scale genome-wide association studies (GWAS), exposure data may not always be available. This may be the case, when the exposure is a biomarker difficult or prohibitively expensive to measure in tens of thousands of disease cases and controls. Therefore, efficient designs of MR studies were recently proposed [40]. A "subsample" IV estimation design can be used, when data on the IV-exposure association are available for a subset of participants but data on the IV-outcome association are available for all participants in the same dataset. A "two-sample" IV estimation design can also be used, when data for the association between the IV and exposure and the IV and outcome are available from different independent datasets. Simulation studies suggest that subsample IV designs obtain statistical power estimates comparable with studies with complete exposure data [40]. In particular, it was shown that power exceeds 90% even when exposure data was available for 20% of the total sample size. Overall, power for MR studies is most efficiently increased by increasing the sample size of the gene-outcome association. Employing two-sample designs does not result in efficiency loss under the assumption that the two samples are selected randomly from the same underlying population.

5 MR Estimators Using Summary Association Data

In the previous sections, we discussed MR study designs and estimation methods of the causal effect when individual level data on the genetic variant(s), exposure and outcome are available. If individual-level data are not available, then valid statistical inference can still be obtained from summarized data on the associations between the genetic variants with the exposure and the outcome. The increasing number of publicly available summary data from GWAS is a valuable source for estimating the causal effect of the exposure on an outcome with greater precision. By using summarized data, one can avoid additional complications arising from confidentiality agreements, especially when it comes to large consortia. Moreover, accumulating evidence from GWAS involving multiple genetic variants can be used to derive an overall causal effect more efficiently [41]. Efficient designs described in Subheading 4 could be adopted. The subsequent sections will focus on details for designing and conducting summary data MR studies .

5.1 Selection of the IVs

Genetic variants used as IVs are selected on the basis of a strong association with the exposure of interest. Robustly and highly statistically significant variants can be selected if individual level data are available from GWAS. Alternatively, the GWAS Catalog, which is a curated repository of accumulating evidence from publicly available GWAS, is a valuable source for identifying summary gene-exposure associations [42]. There is a trade-off in including only genetic variants as IVs with a genome-wide significant association with the exposure, risking an underpowered estimate, or including all available variants with any association with the exposure, risking a biased estimate due to potential violation of the third MR assumption. The current safest recommendation is to select instruments that are genome-wide significantly associated with the exposure.

5.2 Estimation Methods

Two main statistical methods have been proposed for the estimation of causal effects when summarized data are available: the likelihood-based method and the inverse-variance weighted (IVW) method [43].

Let us denote the estimate of association for the genetic variant $k = 1,\ldots,K$ with the exposure by $\widehat{\beta}_{Xk}$ with standard error σ_{Xk}, and the estimate of association with the outcome by $\widehat{\beta}_{Yk}$ with standard error σ_{Yk}. Under the likelihood-based method assuming linearity of the exposure-outcome association, the causal effect $\left(\widehat{\beta}_L\right)$ is estimated by the following model:

$$\begin{aligned}\widehat{\beta}_{Xk} &\sim N\left(\xi_k, {\sigma_{Xk}}^2\right)\\ \widehat{\beta}_{Yk} &\sim N\left(\beta_L \xi_k, {\sigma_{Yk}}^2\right)\end{aligned} \tag{1}$$

The parameters of the model in Eq. (1) can be estimated under standard likelihood or Bayesian approaches. This basic model is valid for two-sample designs. Modification is required to account for the correlation structure of gene-exposure and gene-outcome associations if they are estimated in the same or overlapping participants [43].

Another approach is based on the idea to combine the ratio estimates of the causal effects $\frac{\widehat{\beta}_{Yk}}{\widehat{\beta}_{Xk}}$ from each genetic variant by employing an IVW meta-analysis [44]. The variance of the ratio estimate can be approximately estimated using the Delta method, as $\frac{\sigma^2_{Yk}}{\widehat{\beta}^2_{Xk}}$ [45]. Further terms could be also incorporated to account for the uncertainty in the gene-exposure association. Thus, the IVW estimate can be expressed as:

$$\widehat{\beta}_{\mathrm{IV}W} = \frac{\sum_k \widehat{\beta}_{Xk}\widehat{\beta}_{Yk}{\sigma_{Yk}}^{-2}}{\sum_k \widehat{\beta}_{Xk}{}^2 {\sigma_{Yk}}^{-2}} \tag{2}$$

with an approximated standard error given by:

$$\mathrm{se}\left(\widehat{\beta}_{\mathrm{IV}W}\right) = \sqrt{\frac{1}{\sum_k \widehat{\beta}_{Xk}{}^2 {\sigma_{Yk}}^{-2}}} \tag{3}$$

Thus, the IVW method is a weighted average of the causal effects derived from the genetic variants k. Of course, if only one IV is available Eq. (2) is simplified to the classical ratio estimator. If the IVs are not in LD (uncorrelated IVs), the causal estimate obtained from the IVW method is equivalent to a two-stage approach using individual-level data. This basic model assumes that any differences in the causal estimates derived from multiple IVs can be explained by their variances, as these are assumed to represent the same underlying quantity (homogeneity assumption). This assumption can be tested using the classical Cohran's Q heterogeneity statistic.

Both the likelihood-based and the IVW methods assume that gene-gene interactions among the selected IVs are negligible and that the IVs are not in LD. Extensive simulation studies reveal that gene-gene interactions have little impact on the estimates obtained using summarized data MR methods [43]. On the other hand, when correlated variants are used as IVs the standard errors are underestimated, which may result in invalid statistical inferences. This warrants the need of statistical approaches that account for the

correlation structure of the IVs [16]. Overall, the estimates obtained from the likelihood-based approach including robustly associated variants that are not in LD are unbiased and precise when compared against those derived by employing two-stage methods using individual-level data. The IVW approach also gives similar point estimates to two-stage methods [43].

The IVW and likelihood-based approaches are based on the idea to synthesize separate causal effects derived from multiple IVs in order to derive an overall causal estimate. Another perspective could be to combine multiple IVs and construct an allele score [46]. Unweighted or weighted scores can be constructed, and it has been shown that both approaches give unbiased results [46]. Formulas for the calculation of allele score based methods using summarized data have been presented and bias can be avoided if weights from an external population are used [47]. Allele scores accounting for variants in LD have been recently described allowing for the inclusion of multiple correlated IVs that further increase power of the causal estimates [47]. However, caution on the interpretation of the findings is necessary if the allele score is composed of invalid IVs. We conclude that synthesizing evidence using summarized data is a good compromise if individual-level data are not available.

5.3 Assessing IV Assumptions

We have discussed MR methods for estimating causal effects using summary association data. Here, we discuss methods that are used for assessing the validity of IV assumptions and present robust estimation methods that account for pleiotropy. Assessing IV assumptions is particularly challenging when individual level data are not available. To secure the non-violation of the first assumption of MR, genetic variants are selected that are robustly associated with the exposure of interest in GWAS. The strength of an instrument can be evaluated using the F statistic in the regression of the exposure on the IV [48]. In the case of summary data, $F = \frac{N-K-1}{K}\frac{R^2}{1-R^2}$, with R^2 approximately equal to $2\widehat{\beta}_{Xk} \times MAF \times (1 - MAF)$, where N is the sample size, K is the number of genetic variants, R^2 the proportion of the variance of the exposure explained by the IV and MAF the minor allele frequency. Thus, the F-statistic depends on the sample size, the number of IVs, MAF and the proportion of the variance explained by the IVs. If F is less than 10, this is an indication of weak instrument(s) [49]. However, this value is arbitrary and is valid only for two-stage methods [50]; it fluctuates according to the sample size and is calculated post-hoc [51]. Alternatively, the instrument strength could be quantified by a modification of the classical I^2 statistic [52], which is termed $I_{GX}{}^2$ [53] attributing the excess of variability of gene-exposure associations to measurement error. This statistic fluctuates from

0 to 1, and a value of 0.9 is equivalent to an F statistic of 10. The second assumption is testable only for known confounders. If summarized data are available, this assumption could be at least partially tested acquiring information from the literature or from checking for possible associations between the selected IVs and known confounders in the GWAS Catalogue [42].

Although the inclusion of multiple IVs derived from published GWAS can increase power to detect causal effects, it is more likely to introduce bias due to violation of the third MR assumption [46]. If the distinct causal estimates derived from each genetic variant differ, this may be an indication of pleiotropic effects. Formal statistical tests exist to test for those discrepancies including the classical Cohran's Q statistic, the I^2 statistic [54] or likelihood ratio tests. Heterogenous effects could also be detected by plotting causal estimates from the each included IV.

Moreover, pleiotropy could be tested applying the MR-Egger regression method [55]. We move back to the IVW estimate of Eq. (2), which is equivalent to fitting a weighted linear regression of the associations of the IVs with the outcome on the IVs with the exposure with no intercept term. This analysis assumes that all IVs are valid and no pleiotropic effects exist. The classical IVW method is a good fit only if pleiotropy is balanced. In order to account for directional pleiotropic effects, one could reformulate the aforementioned regression model with no constraint on the intercept term resulting in the so-called MR-Egger regression method [55]. This intercept term captures the average pleiotropic effects of the IVs and values away from zero are an indication of directional pleiotropy. The slope of the MR-Egger regression is a robust estimate of the causal effect. This approach assumes that the pleiotropic effects of the IVs on the outcome are distributed independently of the associations of the IVs with the exposure (InSIDE assumption). The InSIDE assumption is more likely to be satisfied if IVs are not associated with a confounder of the exposure-outcome association. An additional assumption that is required to hold is the so-called "NO Measurement Error" (NOME) where the variance of the IV-exposure association is negligible [43, 56]. In a two-sample MR design, violation of the NOME assumption results in underestimated causal effects and other approaches have been proposed [53], such as the simulation extrapolation approach (SIMEX) [57]. A limitation of the MR-Egger regression is the lack of power and poor performance when few instruments are used. MR-Egger regression is also more sensitive to the InSIDE assumption violation than IVW. Besides, the confidence intervals when the causal effect is not null are not precisely estimated and overinterpretation should be avoided.

Another robust method to account for directional pleiotropy is based on the simple idea to order the ratio estimates of the *k* genetic variants and report the median [58], which assumes that at least 50% of the variants are valid. The InSIDE assumption is not necessary. Violations of the second and the third assumptions are also allowed. One could also derive a weighted median estimate assigning weights proportionally to the precision of the causal estimate derived from each IV [59]. This approach requires at least 50% of the weights to originate from valid IVs. Simulation studies reveal that the weighted median approach results in more precise estimates, when compared against the MR-Egger regression method [59]. Extensions of the classical additive or multiplicative random effects models used in meta-analysis can accommodate both balanced and directional pleiotropy [60].

Statistical tests for assessing pleiotropy could be further enriched graphically. A typical graph presented in MR studies is to plot the gene-outcome against the gene-exposure associations. If pleiotropy is absent, we expect that a variant's association with outcome is proportional to its association with exposure, and therefore the plotted points fall along a line that passes through the origin and has a slope equal to the MR estimate. Additionally, one could create a funnel plot of the reciprocal of the standard error versus the MR causal estimate and check for any asymmetry.

Another aspect in an MR setting is that estimation methods of the causal effects assume linearity of the exposure-outcome association. This is important and most MR investigations do not check this, but future MR studies should check first if exposure-outcome relationships derived from multi-SNP scores are linear before using the suggested estimation methods.

5.4 MR in Practice

In this section we will describe step by step how one can perform an MR study using publicly available summary data. We will focus on the two most popular options, the MR-base platform (http://www.mrbase.org/) [21] and the MendelianRandomization package in R. MR-base is a database and an online platform that allows the user to run a two-sample MR analysis. Currently, it is a collective repertoire of 'complete summary data' from 1094 GWAS analyses from 44 consortia with approximately 4 billion associations between SNPs and phenotypes (i.e., diseases, risk factors, metabolites and immune system traits). This database populates information not only from the GWAS Catalog [42], but also gene expression quantitative trait loci (QTLs) [61], methylation level QTLs [62], metabolite level QTLs [63] and protein level QTLs [64]. In the first step, one has to select the exposure of interest from the appropriate source (GWAS catalog, gene expression QTLs, etc.), and robustly associated IVs (with the exposure) are extracted with respect to a

p-value threshold for inclusion and/or LD threshold for pruning IVs that can be modified by the user. Alternatively, the user can upload a specific list of IVs manually with pre-calculated effect sizes and standard errors. In a second step, the user chooses the outcome of interest. For instance, if we were interested in testing the causal association of body mass index (exposure) and lung cancer (outcome), originally published by Carreras-Torres and coworkers [8], we would select IVs for the exposure from the Genetic Investigation of ANthropometric Traits (GIANT) consortium and for the outcome from the International Lung Cancer Consortium (ILCCO) by clicking on the respective GWAS [65, 66]. The platform also offers the functionality to use proxies if a particular IV is not present, harmonise gene-exposure and gene-outcome effect alleles to ensure a common effect allele is used in both associations, and correct for palindromic SNPs. The user then selects the method of analysis (e.g., IVW, maximum likelihood, etc.) and is navigated to the results window. One can retrieve summary information on the studies used for the exposure and outcome associations, the number of variants extracted, and the MR causal estimates from each predefined method. The presence of pleiotropy can be evaluated using the reported heterogeneity statistics and the p-value of the intercept from the MR-Egger regression method. Using the example of body mass index and risk of lung cancer, we retrieved a total of 79 IVs using default settings (i.e., *p*-value threshold for including IVs at 5×10^{-8}, LD R^2 values for pruning IVs at 0.001, clumping distance at 10.000, LD R^2 values for proxies at 0.8 and a MAF threshold for aligning palindromes at 0.3). None of the IVW, MR-Egger or weighted median approaches yielded statistically significant causal estimates. There was some evidence for heterogeneity, but evidence for directional pleiotropy was not present. We urge investigators to check for associations of the selected IVs with known confounders such as smoking in the particular example, and re-evaluate MR estimates after excluding those variants. Four plots are also available (i.e., causal effects calculated from each IV, IV-outcome associations against IV-exposure, causal effects derived removing one IV sequentially and a funnel plot of the reciprocal of the standard error versus the MR causal estimate). MR results and associated plots can also be converted in an HTML format. An interested researcher can alternatively use the TwoSampleMR package in R to perform the analysis (Box 1).

Box 1. Estimating causal association of body mass index with lung cancer using TwoSampleMR R package.

```
# Load TwoSampleMR R package:
. library(TwoSampleMR)
# Obtain data from MR Base GWAS database:
. ao<- available_outcomes()
# Extract IVs for an exposure, for example to obtain IVs for body mass index
using Locke et al. 2015 GIANT study, specifying the study ID:
. exposure_dat<- extract_instruments(ao$id[c(2)])
*Options are also available:
p1 = P-value threshold for keeping a SNP (default=5e-08)
clump = Whether or not to return independent SNPs only (default=TRUE)
r2 = The maximum LD R-square allowed between returned SNPs (default=0.001)
kb = The distance in which to search for LD R-square values (default=10.000)
# Extract IVs for an outcome, for example to obtain IVs for lung cancer
using Wang et al. 2014 ILCCO study, specifying the study ID, LD Rsq values
for proxies at 0.8 and a MAF threshold for aligning palindromes at 0.3:
. outcome_dat<- extract_outcome_data(exposure_dat$SNP, c(966), proxies = 1,
rsq = 0.8, align_alleles = 1, palindromes = 1, maf_threshold = 0.3)
# Harmonise exposure-outcome data to match the same reference allele,
inferring forward strand using allele frequency:
. dat<- harmonise_data(exposure_dat, outcome_dat, action = 2)
# Perform an MR analysis:
. mr_results<- mr(dat)
## Sensitivity analyses:
# Obtain heterogeneity statistics:
. mr_heterogeneity<- mr_heterogeneity(dat)
# Test for directional pleiotropy:
. mr_pleiotropy_test<- mr_pleiotropy_test(dat)
# Obtain MR estimates for each of the selected IVs:
```

```
. res_single<- mr_singlesnp(dat)

# Obtain MR estimates excluding one IV at a time:

. res_loo<- mr_leaveoneout(dat)

## Creating plots:

# Create scatter plot of IV-outcome associations against IV-exposure:

. p1<- mr_scatter_plot(mr_result, dat)

# Create forest plot of causal effects calculated from each IV:

. p2<- mr_forest_plot(res_single)

# Create plot of causal effects derived removing one IV sequentially:

. p3<- mr_leaveoneout_plot(res_loo)

# Create funnel plot of of the reciprocal of the standard error versus the
MR causal estimate:

. p4<- mr_funnel_plot(res_single)
```


Another option is to use the MendelianRandomization package in R. This package offers the extra functionality to model the correlation of IVs that are in LD which is not feasible via the TwoSampleMR package. Moreover, overlapping samples for estimating IV-exposure and IV-outcome associations can be accounted for. The I_{GX}^2 statistic [53] can be also calculated in order to measure the instrument strength. The user has to specify appropriately vectors including IV-exposure and IV-outcome beta estimates along with their standard errors and this information is not automatically retrieved as in MR-base. However, the authors are willing to directly import information from genetic association studies available in PhenoScanner (http://phenoscanner.medschl.cam.ac.uk) in the package in the near future. Optionally, names of the genetic variants, effect or non-effect alleles and effect allele frequencies can be provided by the user. For demonstration purposes, we reanalysed the harmonised data extracted from the MR Base for estimating the potential causal association of body mass index with lung cancer risk assuming that IVs are independent (Box 2).

Box 2.Estimating causal association of body mass index with lung cancer using MendelianRandomization R package.

```
# Load MendelianRandomzation R package:
. library(MendelianRandomization)
# Create MRInput object from the harmonised body mass index using Locke et al. 2015 GIANT study and lung cancer using Wang et al. 2014 ILCCO study obtained from MR Base GWAS database:
. MRInputObject <- mr_input(bx = dat$beta.exposure, bxse = dat$se.exposure, by = dat$beta.outcome, byse = dat$se.outcome,exposure = "Body mass index", outcome = "Lung cancer", snps = dat$SNP)
# Run IVW MR method:
. IVW<- mr_ivw(MRInputObject,
model = "default",
robust = FALSE,
penalized = FALSE,
weights = "simple",
distribution = "normal",
alpha = 0.05)
*Options for IVW method:
model = "default", "random" or "fixed" (default=fixed-effect with 3 IVs or fewer)
robust = robust instead of standard regression can be performed (default=FALSE)
penalized = penalty can be applied to downweight the contribution of genetic variants with outlying ratio estimates (default=FALSE)
weights = "simple" or "delta", the latter option uses the delta method to calculate the variance of the ratio estimates (default=simple)
distribution = "normal" or "t-dist" (default=normal)
# Run MR Egger method:
. Egger<- mr_egger(MRInputObject,
```

```
                        robust = FALSE,
                        penalized = FALSE,
                        distribution = "normal",
                        alpha = 0.05)
*Options as in IVW method
# Run ML method:
. MaxLik<- mr_maxlik(MRInputObject,
                          model = "default",
                          distribution = "normal",
                          alpha = 0.05)
*Options as in IVW method
# Run Median based method:
. Median<- mr_median(MRInputObject,
                                  weighting = "weighted",
                                  distribution = "normal",
                                  alpha = 0.05,
                                  iterations = 10000,
                                  seed = 314159265)
*Options for median based methods:
weighting = "simple", "weighted" or "penalized" (default=weighted)
distribution = "normal" or "t-dist" (default=normal)
iterations = bootstrap samples for calculating standard errors
(default=10000)
seed = seed to use when generating bootstrap samples (default=314159265)
# Run all methods:
. MR_all<- mr_allmethods(MRInputObject, method = "all")
## Creating plots:
# Create scatter plot of IV-outcome associations against IV-exposure:
. p<- mr_plot(MRInputObject,
error = TRUE,
orientate = FALSE,
```

```
interactive = TRUE,

labels = TRUE,

line = "ivw")

*Options for scatter plot:

error = include error bars (default=TRUE)

orientate = convert negative gene-exposure associations to positive (default=FALSE)

interactive = produces interactive plots (default=TRUE)

labels = displays IV labels (default=FALSE)

line = "ivw" or "egger" (default=ivw)
```

6 Discussion

MR is a powerful approach for deriving causal inferences about the effect of an exposure on an outcome overcoming limitations of observational epidemiology (i.e., confounding and reverse causation). As the sharing of summary data from consortia becomes common practice, numerous genetic variants can be utilized as possible IVs resulting in greater efficiency and more powerful causal estimates. We described the methods for conducting an MR study using summary association data and provided practical guidance using available software.

The MR methodology has great promise for advancing biomedical research, but is also subject to assumptions and limitations caused by unsuitable IVs, population stratification, LD and pleiotropy. We showed that MR assumptions are difficult to be evaluated when individual level data are available, which becomes even more difficult when only summary association data are available. MR is a relatively new field, and additional methodology is warranted to increase the sensitivity and power to detect potential violation of IV assumptions. For instance, MR methods that allow for automatic identification of specific genetic variants with pleiotropic effects that could be excluded from subsequent analysis could strengthen the MR approach. Moreover, selected genetic variants usually explain a small proportion of the variance in the different exposures. Given that many of these environmental exposures/traits are highly heritable, further work using additional genetic variants as instruments, when they become available from future GWAS, will increase power of MR studies and will allow investigations in subgroups. As in all science, replication of results from MR studies is vital.

References

1. Thomas DC, Conti DV (2004) Commentary: the concept of 'Mendelian Randomization'. Int J Epidemiol 33(1):21–25. https://doi.org/10.1093/ije/dyh048
2. Smith GD, Ebrahim S (2003) Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 32 (1):1–22
3. Lawlor DA, Harbord RM, Sterne JA et al (2008) Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med 27 (8):1133–1163. https://doi.org/10.1002/sim.3034
4. Bochud M, Rousson V (2010) Usefulness of Mendelian randomization in observational epidemiology. Int J Environ Res Public Health 7 (3):711–728. https://doi.org/10.3390/ijerph7030711
5. Burgess S, Butterworth A, Malarstig A et al (2012) Use of Mendelian randomisation to assess potential benefit of clinical intervention. BMJ 345:e7325. https://doi.org/10.1136/bmj.e7325
6. Kivimaki M, Smith GD, Timpson NJ et al (2008) Lifetime body mass index and later atherosclerosis risk in young adults: examining causal links using Mendelian randomization in the cardiovascular risk in young finns study. Eur Heart J 29(20):2552–2560. https://doi.org/10.1093/eurheartj/ehn252
7. Voight BF, Peloso GM, Orho-Melander M et al (2012) Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 380(9841):572–580. https://doi.org/10.1016/S0140-6736(12)60312-2
8. Carreras-Torres R, Haycock PC, Relton CL et al (2016) The causal relevance of body mass index in different histological types of lung cancer: a Mendelian randomization study. Sci Rep 6:31121. https://doi.org/10.1038/srep31121
9. Dixon SC, Nagle CM, Thrift AP et al (2016) Adult body mass index and risk of ovarian cancer by subtype: a Mendelian randomization study. Int J Epidemiol 45(3):884–895. https://doi.org/10.1093/ije/dyw158
10. Gao C, Patel CJ, Michailidou K et al (2016) Genetically predicted body mass index and breast cancer risk: mendelian randomization analyses of data from 145,000 women of European descent. PLoS Med 13(8): e1002105. https://doi.org/10.1371/journal.pmed.1002105
11.
12. Didelez V, Sheehan N (2007) Mendelian randomization as an instrumental variable approach to causal inference. Stat Methods Med Res 16(4):309–330. https://doi.org/10.1177/0962280206077743
13. Glymour MM, Tchetgen Tchetgen EJ, Robins JM (2012) Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. Am J Epidemiol 175(4):332–339. https://doi.org/10.1093/aje/kwr323
14. Hernan MA, Robins JM (2006) Instruments for causal inference: an epidemiologist's dream? Epidemiology 17(4):360–372. https://doi.org/10.1097/01.ede.0000222409.00878.37
15. Lawlor DA (2016) Commentary: two-sample Mendelian randomization: opportunities and challenges. Int J Epidemiol 45(3):908–915. https://doi.org/10.1093/ije/dyw127
16. Burgess S, Scott RA, Timpson NJ et al (2015) Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. Eur J Epidemiol 30 (7):543–552. https://doi.org/10.1007/s10654-015-0011-z
17. Burgess S, Small DS, Thompson SG (2015) A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res. https://doi.org/10.1177/0962280215597579
18. Boef AG, Dekkers OM, le Cessie S (2015) Mendelian randomization studies: a review of the approaches used and the quality of reporting. Int J Epidemiol 44(2):496–511. https://doi.org/10.1093/ije/dyv071
19. Davies NM, Smith GD, Windmeijer F et al (2013) Issues in the reporting and conduct of instrumental variable studies: a systematic review. Epidemiology 24(3):363–369. https://doi.org/10.1097/EDE.0b013e31828abafb
20. Haycock PC, Burgess S, Wade KH et al (2016) Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr 103 (4):965–978. https://doi.org/10.3945/ajcn.115.118216
21. Hemani G, Zheng J, Wade KH et al (2016) MR-base: a platform for systematic causal inference across the phenome using billions of

genetic associations. bioRxivr. https://doi.org/10.1101/078972

22. Greenland S (2000) An introduction to instrumental variables for epidemiologists. Int J Epidemiol 29(4):722–729
23. Martens EP, Pestman WR, de Boer A et al (2006) Instrumental variables: application and limitations. Epidemiology 17(3):260–267. https://doi.org/10.1097/01.ede.0000215160.88317.cb
24. Wald A (1940) The fitting of straight lines if both variables are subject to error. Ann Math Stat 11(3):284–300
25. Fieller E (1954) Some problems in interval estimation. J R Stat Soc Series B Stat Methodology 16(2):175–185
26. Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall/CRC Press, Boca Raton, Florida
27. Anderson T, Rubin H (1949) Estimators of the parameters of a single equation in a complete set of stochastic equations. Ann Mathe Stat 21 (1):570–582
28. Moreira M (2003) A conditional likelihood ratio test for structural models. Econometrica 71(4):1027–1048
29. Ebrahim S, Davey Smith G (2008) Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? Hum Genet 123(1):15–33. https://doi.org/10.1007/s00439-007-0448-6
30. Angrist J, Pischke J (2009) Mostly harmless econometrics: an empiricist's companion. Chapter 4: instrumental variables in action: sometimes you get what you need. Princeton University Press, Princeton, New Jersey
31. Nagelkerke N, Fidler V, Bernsen R et al (2000) Estimating treatment effects in randomized clinical trials in the presence of non-compliance. Stat Med 19(14):1849–1864
32. Davidson R, MacKinnon J (1993) Estimation and inference in econometrics. Chapter 18: simultaneous equation models. Oxford University Press, Oxford
33. Kleibergen F, Zivot E (2003) Bayesian and classical approaches to instrumental variable regression. J Econom 114:29–72
34. Foster E (1997) Instrumental variables for logistic regression: an illustration. Soc Sci Res 26(4):487–504
35. Johnston KM, Gustafson P, Levy AR et al (2008) Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. Stat Med 27(9):1539–1556. https://doi.org/10.1002/sim.3036
36. Hansen LP, Heaton J, Yaron A (1996) Finite-sample properties of some alternative GMM estimators. J Bus Econ Stat 14(3):262–280
37. Bowden J, Vansteelandt S (2011) Mendelian randomization analysis of case-control data using structural mean models. Stat Med 30 (6):678–694. https://doi.org/10.1002/sim.4138
38. Greenland S, Lanes S, Jara M (2008) Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. Clin Trials 5(1):5–13. https://doi.org/10.1177/1740774507087703
39. Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. Math Model 7 (9–12):1393–1512
40. Pierce BL, Burgess S (2013) Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am J Epidemiol 178(7):1177–1184. https://doi.org/10.1093/aje/kwt084
41. Pierce BL, Ahsan H, Vanderweele TJ (2011) Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. Int J Epidemiol 40 (3):740–752. https://doi.org/10.1093/ije/dyq151
42. Welter D, MacArthur J, Morales J et al (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42 (Database issue):D1001–D1006. https://doi.org/10.1093/nar/gkt1229
43. Burgess S, Butterworth A, Thompson SG (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 37(7):658–665. https://doi.org/10.1002/gepi.21758
44. Johnson T (2011) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Technical report, Queen Mary University of London
45. Thomas DC, Lawlor DA, Thompson JR (2007) Re: estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. Ann Epidemiol 17 (7):511–513. https://doi.org/10.1016/j.annepidem.2006.12.005
46. Burgess S, Thompson SG (2013) Use of allele scores as instrumental variables for Mendelian randomization. Int J Epidemiol 42 (4):1134–1144. https://doi.org/10.1093/ije/dyt093

47. Burgess S, Dudbridge F, Thompson SG (2016) Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. Stat Med 35 (11):1880–1906. https://doi.org/10.1002/sim.6835
48. Stock J, Wright J, Yogo M (2002) A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat 20(4):518–529
49. Staiger D, Stock J (1997) Instrumental variables regression with weak instruments. Econometrica 65(3):557–586
50. Burgess S, Granell R, Palmer TM et al (2014) Lack of identification in semiparametric instrumental variable models with binary outcomes. Am J Epidemiol 180(1):111–119. https://doi.org/10.1093/aje/kwu107
51. Burgess S, Thompson SG, CRP CHD Genetics Collaboration (2011) Avoiding bias from weak instruments in Mendelian randomization studies. Int J Epidemiol 40(3):755–764. https://doi.org/10.1093/ije/dyr036
52. Higgins JP, Thompson SG, Deeks JJ et al (2003) Measuring inconsistency in meta-analyses. BMJ 327(7414):557–560. https://doi.org/10.1136/bmj.327.7414.557
53. Bowden J, Del Greco MF, Minelli C et al (2016) Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-egger regression: the role of the I2 statistic. Int J Epidemiol 45 (6):1961–1974. https://doi.org/10.1093/ije/dyw220
54. Greco MF, Minelli C, Sheehan NA et al (2015) Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. Stat Med 34(21):2926–2940. https://doi.org/10.1002/sim.6522
55. Bowden J, Davey Smith G, Burgess S (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. Int J Epidemiol 44 (2):512–525. https://doi.org/10.1093/ije/dyv080
56. Brion MJ, Shakhbazov K, Visscher PM (2013) Calculating statistical power in Mendelian randomization studies. Int J Epidemiol 42 (5):1497–1501. https://doi.org/10.1093/ije/dyt179
57. Cook JR, Stefanski LA (1994) Simulation-extrapolation estimation in parametric measurement error models. J Am Stat Assoc 89 (428):1314–1328. https://doi.org/10.2307/2290994
58. Han C (2008) Detecting invalid instruments using L1-GMM. Econ Lett 101(3):285–287
59. Bowden J, Davey Smith G, Haycock PC et al (2016) Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol 40(4):304–314. https://doi.org/10.1002/gepi.21965
60. Bowden J, Del Greco MF, Minelli C et al (2017) A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Stat Med. https://doi.org/10.1002/sim.7221
61. GTEx Consortium (2015) Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348(6235):648–660. https://doi.org/10.1126/science.1262110
62. Gaunt TR, Shihab HA, Hemani G et al (2016) Systematic identification of genetic influences on methylation across the human life course. Genome Biol 17:61. https://doi.org/10.1186/s13059-016-0926-z
63. Kettunen J, Demirkan A, Wurtz P et al (2016) Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat Commun 7:11122. https://doi.org/10.1038/ncomms11122
64. Deming Y, Xia J, Cai Y et al (2016) Genetic studies of plasma analytes identify novel potential biomarkers for several complex traits. Sci Rep 6:18092. https://doi.org/10.1038/srep18092
65. Locke AE, Kahali B, Berndt SI et al (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518 (7538):197–206. https://doi.org/10.1038/nature14177
66. Wang Y, McKay JD, Rafnar T et al (2014) Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet 46 (7):736–741. https://doi.org/10.1038/ng.3002

Author Queries

Chapter No.: 13 394545_1_En

Query Refs.	Details Required	Author's response
AU1	Please check whether the affiliation and correspondence details are presented correctly.	
AU2	Please provide complete bibliographic details of this ref. [11].	