Cardiovasc Prev Pharmacother. 2020 Jan;2(1):24-30
https://doi.org/10.36011/cpp.2020.2.e3
eISSN 2671-700X

Special Article

Check for updates

Basic Concepts of a Mendelian Randomization Approach

Tae-Hwa Go, MS[1], Dae Ryong Kang, PhD[1,2]

[1]Department of Biostatistics, Yonsei University Wonju College of Medicine, Wonju, Korea
[2]Department of Precision Medicine, Yonsei University Wonju College of Medicine, Wonju, Korea

OPEN ACCESS

Received: Dec 19, 2019
Accepted: Dec 29, 2019

Correspondence to
Dae Ryong Kang, PhD
Department of Biostatistics, Yonsei University Wonju College of Medicine, 20 Ilsan-ro, Wonju 26426, Korea.
Email: dr.kang@yonsei.ac.kr
kdr.bmc@gmail.com

ORCID iDs
Tae-Hwa Go
https://orcid.org/0000-0003-4386-0134
Dae Ryong Kang
https://orcid.org/0000-0002-8792-9730

Conflict of Interest
The authors have no financial conflicts of interest.

Author Contributions
Writing - original draft: Go TH; Writing - review & editing: Kang DR.

ABSTRACT

The Mendelian Randomization (MR) approach is a method that enables causal inference in observational studies. There are 3 assumptions that must be satisfied to obtain suitable results: 1) The genetic variant is strongly associated with the exposure, 2) The genetic variant is independent of the outcome, given the exposure and all confounders (measured and unmeasured) of the exposure-outcome association, 3) The genetic variant is independent of factors (measured and unmeasured) that confound the exposure-outcome relationship. This analysis has been used increasingly since 2011, but many researchers still do not know how to perform MR. Here, we introduce the basic concepts, assumptions, and methods of MR analysis to enable better understanding of this approach.

Keywords: Causality; Epidemiologic studies; Genetic association studies; Mendelian Randomization analysis; Observational study

INTRODUCTION

The primary goals of medical research include the identification of progression, specific consequences, and risk factors of a disease.[1)] In this regard, a randomized controlled trial (RCT) is required to establish a causal relationship between exposure and outcome, and therefore, it is the most representative study design method when conducting medical research. However, an RCT cannot always be performed; consequently, many medical studies are observational instead.

In observational studies, it can be difficult to rule out the effects of confounding variables between exposure and outcomes, and there is a possibility of false causal inferences, regardless of the use of an appropriate study design and statistical methods.[2)] To reduce these errors, the instrumental variable (IV) method has been proposed as an alternative statistical method for investigating the causal relationship between exposure and outcome, while controlling for confounding variables. The IV method was first introduced by econometricians and later applied in Mendelian randomization (MR) analysis in medical statistics. The MR approach was suggested by Katan in 1986,[3)] wherein it was explained how various apolipoprotein E isoforms could be used as IVs for investigating the association

Table 1. Steps for MR analysis

Step	Description
1	Define hypotheses for causal inference
2	Select genetic IV using GWAS or supporting information in the literature
3	Identify the MR assumptions for the selected IV
4	Perform an MR analysis
5	Interpret and discuss results

GWAS = genome wide association studies; IV = instrumental variable; MR = Mendelian Randomization

between serum cholesterol levels and cancer risk.[4] However, MR research did not start to gain popularity until 2011; in 2015, a report on MR was published in a special issue of the *International Journal of Epidemiology*. In the same year, a book regarding MR was also published.[5] Nevertheless, many researchers remain uncertain of the approach toward MR studies. Here, we introduce the basic concept of MR, covering analysis and extension methods (**Table 1**).

BASIC CONCEPT

MR was derived from the concept of genetic variants randomly assigned according to Mendel's second law.[1] Whereas RCTs are randomized to control for confounders during the clinical trial period, in MR studies, genes are assigned to individuals prior to exposure to other factors. Since these genetic factors cannot be modified, genetic variants, such as single-nucleotide polymorphisms (SNPs), are used as IVs for MR analysis.[2][6-8] The general aim of the MR approach is to estimate the causal effect of an exposure (X) on an outcome (Y) using genetic variants (G) for X (**Figure 1**).[9]

Two-stage least squares regression

Two-stage least squares (2SLS) is a two-step method that uses continuous outcomes and a linear model.[10][11] The first step estimates the effect on exposure ($\hat{x}_i$) to the instruments, and the second step estimates the effect on the outcome (y_i) through the estimated exposure ($\hat{x}_i$). Thus, the variation in exposure described by IV identifies the effect on the outcome, and the causal estimate is reflected as a regression coefficient (β_{IV}) for the change in outcome due to the unit change in exposure.[12]

For example, let us suppose that we have an IV available.[11] With data on an individual indexed by i=1, ⋯, N who have exposure x_i and outcome y_i and assuming an additive linear model for the IVs g_{ik} indexed by k=1, ⋯, K, the first-stage regression model is represented as follows:

$$x_i=\alpha_0+\sum_k \alpha_k g_{ik}+\varepsilon_{xi} \quad (1)$$

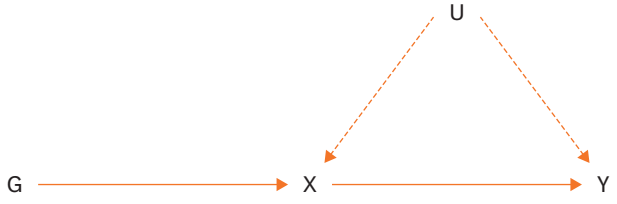
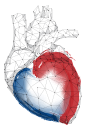


Figure 1. Causal diagram for a Mendelian randomization study.

https://doi.org/10.36011/cpp.2020.2.e3



The fitted values $x_i=\hat{\alpha}_0+\sum_k \hat{\alpha}_k z_{ik}$ are then used in the following second-stage regression model:

$$y_i=\beta_0+\beta_{IV}\hat{x}_i+\varepsilon_{yi} \quad (2)$$

where ε_{Xi} and ε_{Yi} are independent error terms. If both models are estimated by standard least-squares regression, both the error terms are implicitly assumed to be homoscedastic and normally distributed. Estimating the causal effect in a 2-stage method provides the correct point estimate; however, uncertainty in the first-stage regression is not considered. Thus, the standard error from the second-stage regression is not correct.[13]

If the genetic instrument is a path from X to Y, the direct effect of the instrument on the outcome Y (β_{GY}) is equal to the product of the effects on the pathway mediated by exposure (i.e., $\beta_{GY}=\beta_{GX}\times\beta_{\hat{X}Y}$). Accordingly, the causal effect can be estimated by dividing the effect of IV on the outcome (β_{GY}) by the effect of IV on the exposure (β_{GX}) as follows.[9]

$$\beta_{\hat{X}Y}=\frac{\beta_{GY}}{\beta_{GX}} \quad (3)$$

Since the formula is calculated as the ratio of 2 IV-based effect estimates, it is also called a ratio estimate or Wald estimate. The variance of $\beta_{\hat{X}Y}$ is estimated through the delta-method based on Taylor series expansion and can be approximated as follows:

$$\mathrm{var}(\beta_{\hat{X}Y})=var\left(\frac{\beta_{GY}}{\beta_{GX}}\right)\cong\frac{var(\beta_{GY})}{\beta_{GX}^2}+\frac{\beta_{GY}^2}{\beta_{GX}^4}var(\beta_{GX})-2\frac{\beta_{GY}}{\beta_{GX}^3}cov(\beta_{GY},\beta_{GX}) \quad (4)$$

Approximations can be easily identified through statistical programs, such as R, SAS, or STATA.

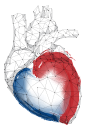
The 2SLS regression method is also applicable when outcome Y is binary. In this case, an approximate normal distribution of X is required, and the causal relative risk or odds ratio parameter can be estimated using a log-linear or logistic regression model in the second-stage regression.[11] However, even in binary outcomes, uncertainty in the first-stage regression is not accounted for, such that the standard error of first-stage coefficients has slightly less precision. This can be solved using a likelihood-based or bootstrap method.[14] The 2SLS method using a non-linear second-stage regression model has been criticized for being called "forbidden regression" because it does not guarantee un-correlation of residuals and IVs in the second-stage regression.[14][15] Debate on the interpretation and validity of such estimates is ongoing.[11] In addition to 2SLS, a limited information maximum likelihood method is available for calculating the IV estimate,[16] although here, we have only explained 2SLS.

ASSUMPTIONS OF MR

Multiple IVs can be used when conducting MR studies via 2SLS. In this case, rather than choosing an IV that is likely to be associated with unconditional research, it is necessary to meet the assumptions. To infer correct causality, the choice of genetic IV in MR studies must be carefully considered, and the 3 main assumptions for allowing IV are as follows.[1][8][15][17-19]

1) The genetic variant is strongly associated with the exposure.
2) The genetic variant is independent of the outcome, given the exposure and all confounders (measured and unmeasured) of the exposure-outcome association.

https://doi.org/10.36011/cpp.2020.2.e3



3) The genetic variant is independent of factors (measured and unmeasured) that confound the exposure-outcome relationship.

These assumptions cannot be easily tested because not all confounders are observed, although they should be confirmed based on the subject matter or background knowledge.[18] Assumption 1 confirms the degree of association between the IV and exposure. F-statistics and R^2 are commonly used for identification.[12] The relationship between the genetic variants selected for exposure and exposure can be confirmed by linear regression. For F-statistics >10, there is a strong association between genetic variants and exposure.[20] Additionally, only genetic variants with p-values $<5\times10^{-8}$ are used for analysis according to genotype using Cuzick's test for trend.[16][21][22] Assumption 2 can be demonstrated by showing that IV affects outcome through exposure, although it can be difficult to verify the same directly due to SNPs in linkage disequilibrium or horizontal pleiotropy of SNPs.[1] However, the assumption that there is no association between the IV and confounder owing to random allocation of alleles is often difficult to prove indirectly by empirically evaluating the association. Finally, Assumption 3 is also difficult to directly prove due to pleiotropy. Indirect tests, including the Sargan and Hansen tests, analyze over-identifying restrictions,[23] identifying the residual effects of genetic instruments on an outcome.

GENETIC RISK SCORES

If the analysis is based on multiple genes and it is known that different biological pathways function between the genes and traits, it is important to include all related information. Using multiple IVs, rather than one IV, can help solve weak instrument bias. Genetic risk score (GRS) is used to enhance the quantitative effect of IVs on risk factors.[24] GRS analysis is based on SNPs selected from genetic information analysis, and SNPs included in the analysis are strongly associated with the exposure and are stratified for low linkage disequilibrium.

There are 2 methods of GRS, namely, counted GRS (cGRS) and weighted GRS (wGRS). cGRS is a simple method of adding the number of risky alleles in each SNP, while weighted GRS first multiplies the weight and number of risky alleles of each SNP and then adds them. For a multi-SNP risk score depending on k chosen SNPs, the value of the risk score for the i-th subject is as follows:

$$cGRS_i=\sum_k x_{ij} \quad (5)$$

$$wGRS_i=\sum_k x_{ik}\times w_k \quad (6)$$

where x_{ij} is the dose of the coded allele at the k-th SNP in the i-th subject and w_k is a chosen coefficient or weight for the k-th SNP.[25] The GRS obtained in this way fits the first-stage regression model as an IV in 2SLS.

EXTENSIVE MR

The MR method using 2SLS is a standard for estimating one-sample MR or single-sample MR. However, some extensions to the MR approach have been developed in recent years. In the following sections, we introduce two-sample MR and bidirectional MR; in addition, 2-step MR, multivariable MR, and factorial MR have been developed.[1][2][19]

https://doi.org/10.36011/cpp.2020.2.e3

Two-sample MR

Standard MR is analyzed using only one-sample data. However, 2-sample MR is a method for estimating the causal effect when exposure and outcome data are measured in different samples.[2)26)27)] This can be used when it is difficult or expensive to measure both exposure and outcome for the same data. In addition, 2-sample MR has become increasingly popular as the scope of analysis has been extended by using summary data of publicly available genome-wide association studies (GWAS).[28)]

Bidirectional MR

Bidirectional MR can determine whether exposure causes an outcome or whether an outcome causes exposure.[29)] The study is first conducted in the direction of exposure to outcome and then in the opposite direction. This method determines whether exposure affects outcomes in the opposite direction or by potential confounding between exposure and outcome.[30)] Nonetheless, the complexity of biological systems may make interpretation of such analytical results difficult.[17)]

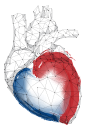
LIMITATION

The first limitation of MR is that it requires large sample sizes.[31)] In many cases, genetic variant proxying for exposure or traits can only account for a very small proportion of the variance in exposure or traits. To obtain an accurate risk estimate, thousands of samples are generally needed. The second limitation of MR is population stratification. Spurious associations may arise in MR where the genetic variant and outcome are associated with ancestral background in an admixed or stratified sample.[2)] To address limitations, there are methods for performing analyses only on homogenous populations or for controlling populations appropriately using principal components analysis or linear mixed models. The third limitation of MR is winner's curse. In the case of single-sample MR, if the discovery GWAS and MR analysis for the genetic instrument use the same sample, the IV and exposure estimates may be biased upward.[26)] This can be averted by using the aforementioned 2-sample MR. The final limitations are trait heterogeneity, horizontal pleiotropy, and linkage disequilibrium. These limitations break existing MR assumptions and require understanding of genetic variants and biological knowledge. Special methods for detecting pleiotropy include MR-Egger regression[32)] and weighted median approaches.[33)]

CONCLUSION

When conducting MR analysis, it is important to be aware of the validity of the assumptions supporting the study and how previous studies have been reported. The most important aspect of MR is the choice of IV.[19)] When conducting studies with some considerations clarified, these findings potentially provide unbiased information on exposure and IV, which can then be used to assess new causal relationships or verify the results studied in RCTs. MR will be applied further in the future as a statistical method to identify causal effects. In addition, the extension of the MR approach may provide a potentially fruitful method for strengthening causal inference in epigenetic studies, and these tools can be applied to contemporary large-scale epigenetic studies. Therefore, efforts must be made to overcome the limitations of MR analysis to ensure precise studies.

https://doi.org/10.36011/cpp.2020.2.e3



REFERENCES

1. Sekula P, Del Greco M F, Pattaro C, Köttgen A. Mendelian randomization as an approach to assess causality using observational data. J Am Soc Nephrol 2016;27:3253-65.
PUBMED | CROSSREF
2. Zheng J, Baird D, Borges MC, Bowden J, Hemani G, Haycock P, Evans DM, Smith GD. Recent developments in Mendelian randomization studies. Curr Epidemiol Rep 2017;4:330-45.
PUBMED | CROSSREF
3. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. 1986. Int J Epidemiol 2004;33:9.
PUBMED | CROSSREF
4. Wells D. Mendelian randomisation: a minireview. Winnower 2015;2015:3073.
5. Burgess S, Thompson S. Mendelian Randomization. New York, NY: Chapman and Hall/CRC; 2015.
6. Emdin CA, Khera AV, Kathiresan S. Mendelian randomization. JAMA 2017;318:1925-6.
PUBMED | CROSSREF
7. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 2003;32:1-22.
PUBMED | CROSSREF
8. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med 2008;27:1133-63.
PUBMED | CROSSREF
9. Teumer A. Common methods for performing Mendelian randomization. Front Cardiovasc Med 2018;5:51.
PUBMED | CROSSREF
10. Baum C, Schaffer M, Stillman S. Instrumental variables and GMM: estimation and testing. Stata J 2003;3:1-31.
CROSSREF
11. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res 2017;26:2333-55.
PUBMED | CROSSREF
12. Pierce BL, Ahsan H, Vanderweele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. Int J Epidemiol 2011;40:740-52.
PUBMED | CROSSREF
13. Angrist JD, Pischke JS, editors. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press; 2009.
14. Foster EM. Instrumental variables for logistic regression: an illustration. Soc Sci Res 1997;26:487-504.
CROSSREF
15. Angrist JD, Pischke JS. Instrumental variables in action: sometimes you get what you need. In: Angrist JD, Pischke JS, editors. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press; 2009. pp.113-220.
16. Burgess S, Thompson SG; CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. Int J Epidemiol 2011;40:755-64.
PUBMED | CROSSREF
17. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr 2016;103:965-78.
PUBMED | CROSSREF
18. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. Stat Methods Med Res 2007;16:309-30.
PUBMED | CROSSREF
19. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 2018;362:k601.
PUBMED | CROSSREF
20. Staiger D, Stock JH. Instrumental variables regression with weak instruments. Econometrica 1997;65:557-86.
CROSSREF
21. Nordestgaard AT, Nordestgaard BG. Coffee intake, cardiovascular disease and all-cause mortality: observational and Mendelian randomization analyses in 95 000–223 000 individuals. Int J Epidemiol 2016;45:1938-52.
PUBMED | CROSSREF

22. Palmer TM, Sterne JA, Harbord RM, Lawlor DA, Sheehan NA, Meng S, Granell R, Smith GD, Didelez V. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. Am J Epidemiol 2011;173:1392-403.
PUBMED | CROSSREF

23. Wehby GL, Ohsfeldt RL, Murray JC. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. Stat Med 2008;27:2745-9.
PUBMED | CROSSREF

24. Jung KJ, Kim S, Yun M, Jeon C, Jee SH. Review on genetic risk score and cancer prediction models. J Health Info Stat 2014;39:1-15.

25. Johnson T. Efficient Calculation for Multi-SNP Genetic Risk Scores. American Society of Human Genetics Annual Meeting 2012.

26. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet 2014;23:R89-98.
PUBMED | CROSSREF

27. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG; EPIC- InterAct Consortium. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. Eur J Epidemiol 2015;30:543-52.
PUBMED | CROSSREF

28. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, Tan VY, Yarmolinsky J, Shihab HA, Timpson NJ, Evans DM, Relton C, Martin RM, Davey Smith G, Gaunt TR, Haycock PC. The MR-Base platform supports systematic causal inference across the human phenome. eLife 2018;7:e34408.
PUBMED | CROSSREF

29. Timpson NJ, Nordestgaard BG, Harbord RM, Zacho J, Frayling TM, Tybjærg-Hansen A, Smith GD. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. Int J Obes 2011;35:300-8.
PUBMED | CROSSREF

30. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet 1986;1:507-8.
PUBMED | CROSSREF

31. Wang LN, Zhang Z. Mendelian randomization approach, used for causal inferences. Zhonghua Liu Xing Bing Xue Za Zhi 2017;38:547-52.
PUBMED | CROSSREF

32. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol 2015;44:512-25.
PUBMED | CROSSREF

33. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic. Int J Epidemiol 2016;45:1961-74.
PUBMED | CROSSREF