

Exploration of errors in variance caused by using the first-order approximation in Mendelian randomization

Hakin Kim¹, Kunhee Kim^{2,3,4}, Buhm Han^{1,2*}

¹Interdisciplinary Program of Bioengineering, Seoul National University College of Engineering, Seoul 08826, Korea

²Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul 03080, Korea

³Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

⁴UTHealth Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Mendelian randomization (MR) uses genetic variation as a natural experiment to investigate the causal effects of modifiable risk factors (exposures) on outcomes. Two-sample Mendelian randomization (2SMR) is widely used to measure causal effects between exposures and outcomes via genome-wide association studies. 2SMR can increase statistical power by utilizing summary statistics from large consortia such as the UK Biobank. However, the first-order term approximation of standard error is commonly used when applying 2SMR. This approximation can underestimate the variance of causal effects in MR, which can lead to an increased false-positive rate. An alternative is to use the second-order approximation of the standard error, which can considerably correct for the deviation of the first-order approximation. In this study, we simulated MR to show the degree to which the first-order approximation underestimates the variance. We show that depending on the specific situation, the first-order approximation can underestimate the variance almost by half when compared to the true variance, whereas the second-order approximation is robust and accurate.

Keywords: computer simulation, delta method, Mendelian randomization analysis

Introduction

It is important to understand the causality between two phenotypes to uncover the pathogenesis of diseases. Some strategies exist for assessing causality in epidemiological studies. Mendelian randomization (MR) is a technique that uses genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure on an outcome [1]. In accordance with Mendel's laws of inheritance, alleles are randomly inherited from parents. Therefore, the genotypes of offspring can be considered independent of confounding factors. Furthermore, the fact that genotypes are fixed and are not affected by phenotypes obviates the reverse causation problem. For these reasons, genetic variants naturally meet many of the basic assumptions of IVs.

Summary statistics released from large genome-wide association studies recently began to facilitate MR by providing exposure effect sizes for multiple genetic variants [2]. The

type of MR analysis using an external dataset for quantifying exposure effect is called a two-sample MR design (2SMR). An advantage of 2SMR is that the statistical power can be increased by merging summary statistics from various sources including large consortia such as the UK Biobank [3]. The causal effect between an exposure and an outcome is estimated by the ratio between the reported genetic effect to the exposure in an external dataset and the observed genetic effect to the outcome in the target dataset. Since there are multiple variants, the ratio estimates over multiple variants are usually combined into a single estimate via the inverse-variance weighted method.

In 2SMR, the standard error of the estimated ratio is conventionally approximated by the first-order term from the delta method. As stated by Thomas et al. [4], however, this approximation can lead to an underestimation of the variance. This underestimation can lead to both increased power and an increased false-positive rate (FPR). An alternative is to use the second-order approximation of the standard error, which can considerably correct for the deviation of the first-order approximation.

In this study, we extensively simulate MR to show the impact of this first-order approximation on the FPR and power of MR. We simulate several different situations to evaluate which study design parameters affect the errors of the first-order approximation, and also compare the errors of the first-order approximation to those of the second-order approximation.

Methods

Genetic variants as instrumental variables

Genetic variants such as single-nucleotide polymorphisms (SNPs) have several properties that make them appropriate as an instrument of exposure. The random inheritance of the alleles makes the genotype distribution independent of socio-economic factors and lifestyle factors such as income [5]. Inherited alleles are not changed from birth by diseases or conditions, except in rare cases of somatic mutations. However, some assumptions still need to be satisfied to ensure the validity of a genetic variant as an IV (Fig. 1). Three basic assumptions must hold for a genetic variant to be used as an IV for MR [6].

- IV1. The genetic variant is associated with the exposure.
- IV2. The genetic variant influences the outcome only through the exposure.
- IV3. The genetic variant is independent of confounding factors affecting the exposure–outcome relationship.

Whether these assumptions are satisfied in various conditions has been discussed elsewhere [7]. Herein, we simply accept these

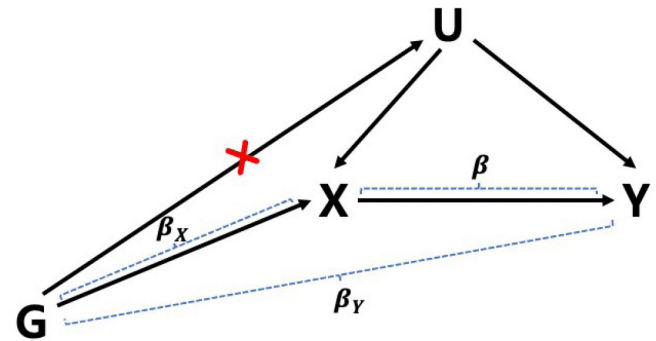


Fig. 1. Diagram depicting the relationships of instrumental variable (IV), exposure, outcome and confounder. Under the assumption of Mendelian randomization, IV should not affect the confounder (red cross). G denotes an IV which is SNP in our case, X denotes an exposure, Y denotes an outcome and U denotes the confounder such as smoking.

assumptions and proceed to the description of MR.

Basic model of MR and the first-order approximation of variance

In this section, we describe the basic model of MR along with the commonly used first-order variance approximation (Fig. 1). Let G be an IV (e.g., a SNP), X be an exposure such as body mass index, and Y be an outcome, such as disease. We can set the relationships between variables (G, X, and Y) via a linear regression model.

$$\begin{aligned} X &| G_X = \beta_{X0} + \beta_X G_X + \varepsilon_X \\ Y &| G_Y = \beta_{Y0} + \beta_Y G_Y + \varepsilon_Y \end{aligned}$$

If we assume that all IV assumptions are satisfied, then $\beta_X \neq 0$ because of IV1 and $\beta_Y = \beta_X \times \beta$ because of IV2 and IV3. That is, G (Fig. 1) affects Y (outcome) only through X (exposure). It is assumed that the error terms ε_X and ε_Y follow normal distributions and are independent in the case of 2SMR of two disjoint samples. Even in the case of two non-overlapping samples, a report has stated the sample correlation between $\hat{\beta}_X$ and $\hat{\beta}_Y$ can be ignored [8]. The ratio estimate $\hat{\beta} = \hat{\beta}_X / \hat{\beta}_Y$ reflects the causal effect between exposure and outcome, and is consistent asymptotically.

To test whether $\beta \neq 0$, it is essential to obtain the variance estimate of $\hat{\beta}$. The commonly used first-order approximation is $\text{Var}(\hat{\beta}) \approx \frac{\text{Var}(\hat{\beta}_Y)}{\hat{\beta}_X^2}$. The first-order approximation method involves treating the denominator $\hat{\beta}_X$ as a constant. However, because of the

innate uncertainty in $\hat{\beta}_X$, we can expect that $\frac{\text{Var}(\hat{\beta}_Y)}{\hat{\beta}_X^2}$ tends to underestimate the true variance of $\hat{\beta}$.

The second-order approximation method of variance of estimated causal effects

Thomas et al. [4] suggested a second-order approximation of the variance of $\hat{\beta}$. With the delta method, one can approximate the variance of causality $\hat{\beta}$ as follows.

$$(\hat{\beta}) \approx \frac{\text{Var}(\hat{\beta}_Y)}{\hat{\beta}_X^2} - 2 \frac{\hat{\beta}_Y}{\hat{\beta}_X^3} \text{Cov}(\hat{\beta}_X, \hat{\beta}_Y) + \frac{\hat{\beta}_Y^2}{\hat{\beta}_X^4} \text{Var}(\hat{\beta}_X)$$

Since we use different samples (2SMR), we can set $\text{Cov}(\hat{\beta}_X / \hat{\beta}_Y) = 0$, as X and Y are from non-overlapping samples. Therefore, we obtain the following approximation.

$$\text{Var}(\hat{\beta}) \approx \frac{\text{Var}(\hat{\beta}_Y)}{\hat{\beta}_X^2} + \frac{\hat{\beta}_Y^2}{\hat{\beta}_X^4} \text{Var}(\hat{\beta}_X)$$

The second term is always positive. Therefore, if researchers use only the first term from this approximation for the variance, this can lead to an underestimation of the standard error.

Simulation design

We designed simulations to evaluate the magnitude of error in the first-order approximation method. We assumed specific true values for β and β_X , which also gave us the true value of $\beta_Y = \beta \times \beta_X$. We assumed the intercepts $\beta_{X0} = 0.03$ and $\beta_{Y0} = 0.03$, and the errors $\sqrt{\text{Var}(\epsilon_X)} = \text{sd}(\epsilon_X) = 0.3$ and $\sqrt{\text{Var}(\epsilon_Y)} = \text{sd}(\epsilon_Y) = 0.3$. We independently generated genotypes (SNP alleles) G_X and G_Y , which are composed of 0, 1, and 2 from the distribution Binomial(2, MAF), where MAF denotes the minor allele frequency. We generated $(X|G_X, Y|G_Y)$ by adding noise with mean 0 and variance $(\text{Var}(\epsilon_X), \text{Var}(\epsilon_Y))$ to $(\beta_{X0} + \beta_X G_X, \beta_{Y0} + \beta_Y G_Y)$. Then we obtained $\hat{\beta}_X$ and $\hat{\beta}_Y$ via simple linear regression. We can expect $\hat{\beta}_X$ and $\hat{\beta}_Y$ to be randomly distributed by

$$\hat{\beta}_X \sim N\left(\beta_X, \frac{\text{Var}(\epsilon_X)}{\text{SS}_{GG_X}}\right)$$

$$\hat{\beta}_Y \sim N\left(\beta_Y, \frac{\text{Var}(\epsilon_Y)}{\text{SS}_{GG_Y}}\right)$$

$$\text{SS}_{GG_X} = \sum G_{Xi}^2 - \frac{(\sum G_{Xi})^2}{N_X} \text{ and } \text{SS}_{GG_Y} = \sum G_{Yi}^2 - \frac{(\sum G_{Yi})^2}{N_Y}$$

where N_X is the size of the reference dataset used in 2SMR and N_Y is the size of the target sample.

To approximate $\text{Var}(\hat{\beta})$, we can use either the first-order or the second-order approximation:

First-order: $\text{Var}(\hat{\beta}) \approx \frac{\text{Var}(\hat{\beta}_Y)}{\hat{\beta}_X^2}$

Second-order: $\text{Var}(\hat{\beta}) \approx \frac{\text{Var}(\hat{\beta}_Y)}{\hat{\beta}_X^2} + \frac{\hat{\beta}_Y^2}{\hat{\beta}_X^4} \text{Var}(\hat{\beta}_X)$

Our simulation allowed us to empirically obtain a very accurate estimate of $\text{Var}(\hat{\beta})$ by repeating the simulation many times (we set the number of simulations as 100,000 in our study) with the same assumptions and calculating the observed variance of $\hat{\beta}$. This allowed us to compare the first and second-order approximations to the empirically obtained values.

We provide the R script code to run the entire simulation pipeline as Supplementary Data.

Results

We performed empirical simulations to compare the two types of analytical approximations: the classical way, in which only the first-order term is used, and the recently suggested way [4], which includes up to the second-order term. We also obtained an accurate estimate of the variance by empirically repeating simulations 100,000 times. Assuming that the empirically obtained variance is the gold standard, we calculated the ratio of the estimated variance to the gold standard.

In our simulations, we varied multiple parameters. We varied the N-ratio (N_X/N_Y), we also varied β (the magnitude of causal effect) and MAF. Fig. 2 shows that the analytical approximation that contained variance up to the second-order term was almost as accurate as the empirical estimate, whereas the first-order approximation method was often largely inaccurate depending on the situation.

Fig. 2A shows that the error due to the first-order approximation decreased as the number of individuals (N_Y) decreased from 200,000 to 2,000 (as the N-ratio increased from 1 to 100). The ratio was 0.84 when N_Y was 100,000, which is equal to $N_X/2$ (N-ratio = 2). The ratio rose to 0.99 when N_Y was 2,000 (N-ratio = 100). The mean of the ratios was 0.98, which translates to a reduced SE($\hat{\beta}$) by $\sqrt{0.98} = 0.99$ times in the first-order approximation. Fig. 2B shows that the errors increased when the actual causal effect (β) between the exposure and outcome increased from 0.01 to 1. Therefore, if there is not a strong causal effect between the exposure and outcome in MR, the error from the first-order approximation would be small. The mean of the ratios of the first-order approximation was 0.93. Fig. 2C shows that, interestingly, the ratio appeared to be independent of the MAF of the variant. The mean of the ratios in this simulation was 0.93 in the first-order case.

We then analyzed the impact of the underestimated variance. If the variance is underestimated, the FPR can increase. We assumed the null hypothesis of no causal effect and generated 100,000 samples under an environment equivalent to that of Fig. 2A. We calculated the FPR based on the significance threshold of $\alpha = 0.05$. Fig.

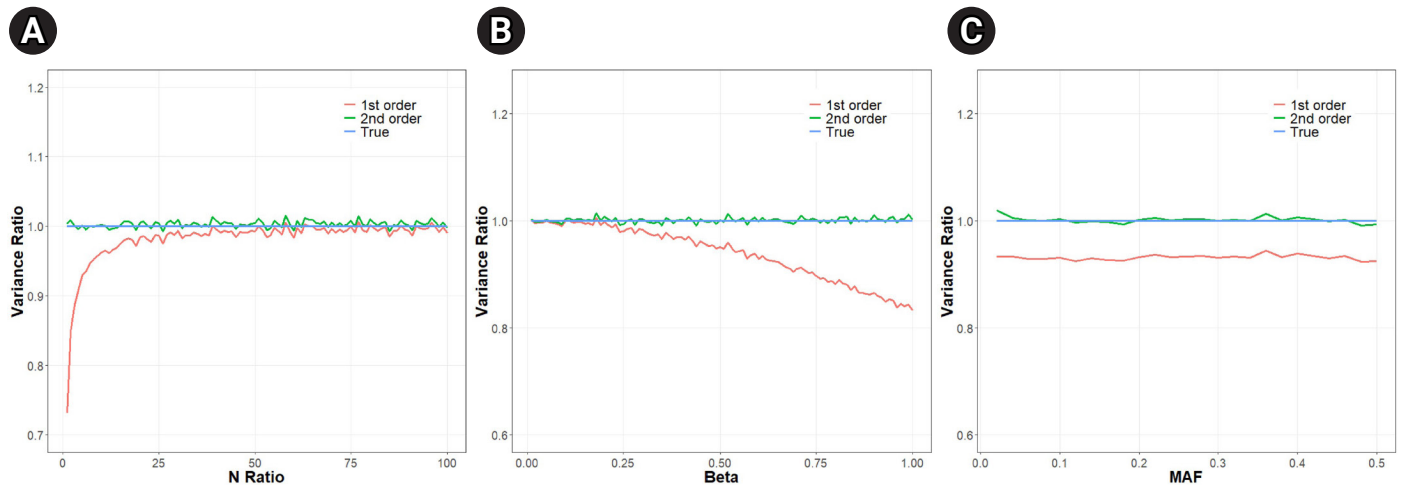


Fig. 2. The ratio of the approximation of the variance of causal effect estimate to the true value. (A) We varied the N-ratio (N_x/N_y) value from 1 to 100 assuming $N_x = 200,000$, $\beta_x = 0.02$, $\beta = 0.6$, and minor allele frequency (MAF) = 0.2. (B) We varied the value β , i.e. the ratio of β_x and β_y from 0.01 to 1 assuming $\beta_x = 0.02$, $N_x = 200,000$, $N_y = 10,000$ and MAF = 0.2. (C) We varied the MAF from 0.02 to 0.5 assuming $\beta = 1$, $\beta_x = 0.02$, $N_x = 200,000$ and $N_y = 10,000$. The true value was estimated by empirical simulations ($N_{sim} = 100,000$).

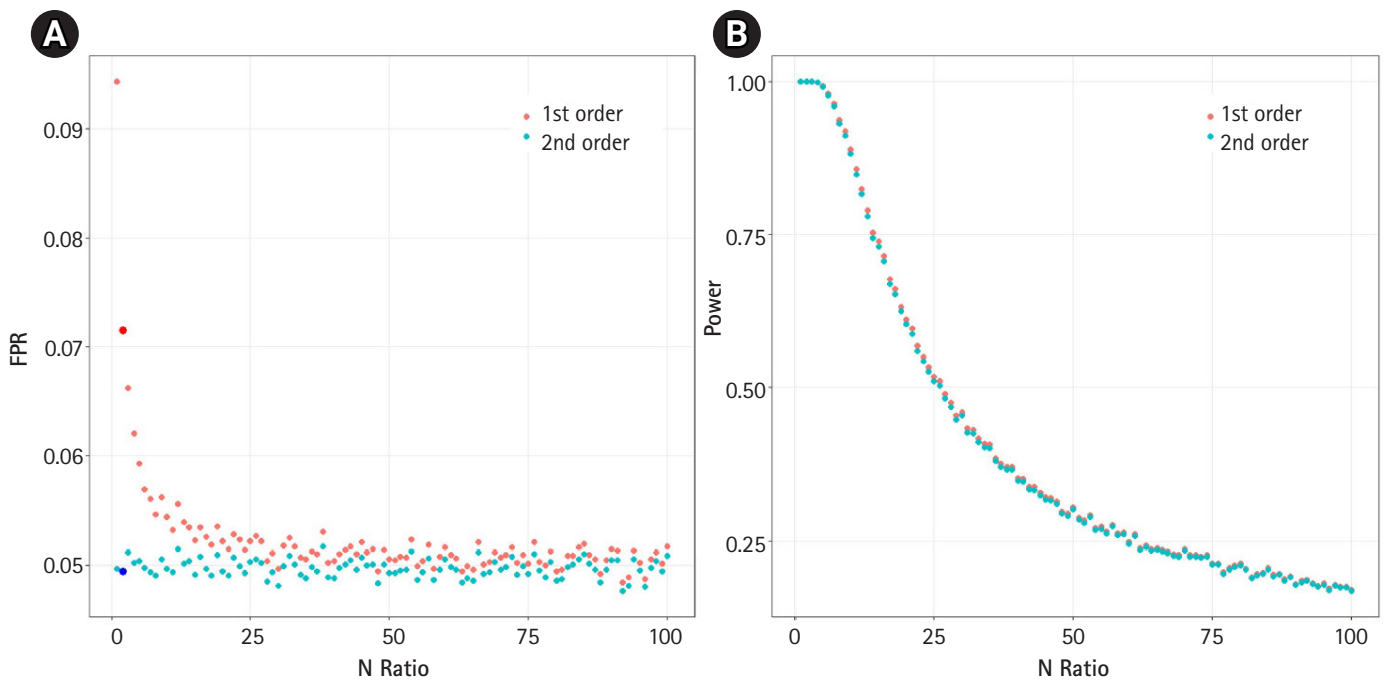


Fig. 3. The scatter plots of false-positive rate (FPR) and statistical power. (A) The scatter plot of relationship between the N-ratio in the simulation of Fig. 2A and FPR. If N-ratio = 2 ($N_y = 100,000$ and $N_x = 200,000$), the FPRs were 0.071 (the dark red colored large dot) for the first-order approximation and 0.049 (the dark blue colored large dot) for the second-order approximation. To calculate FPR values, we generated 100,000 samples from the null hypothesis of no causal effect. (B) The scatter plot of relationship between the N-ratio the power. We generated 100,000 samples with $\beta = 0.6$ which is the causal effect of the exposure on the outcome.

3A shows the relationship between the N-ratio and the FPR. Notably, when the variance was underestimated by a factor of 0.84, as shown in Fig. 2A (for the case of an N-ratio = 2—that is, $N_y = 100,000$ and $N_x = 200,000$), the FPR of the first-order approxima-

tion method increased to 0.071 (the dark red colored large dot in Fig. 3A), while the FPR of the second-order approximation method was 0.049 (the dark blue colored large dot in Fig. 3A), corresponding to approximately 0.7 times that of the first-order case.

The average FPR in the second-order approximation method was 0.049, whereas the average FPR in the first-order approximation was 0.052. These findings indicate that the second-order approximation can be a good choice to prevent inflation of the FPR.

We also analyzed the statistical power (Fig. 3B). Since the variance of $\hat{\beta}$ is underestimated, the first-order approximation method may also tend to increase the power (or underestimate the false-negative rate). To compare the powers of the first and the second-order approximation methods, we generated 100,000 samples under an environment equivalent to that of Fig. 2A, with $\beta = 0.6$, which denotes the causal effect of the exposure on the outcome. Under this setting, the power of the first-order approximation was similar to that of the second-order approximation (on average 1.01 times greater).

Discussion

In this study, we performed simulations to evaluate the errors in the variance estimate of causal effects in 2SMR. We simulated a range of study parameters and showed that the commonly used first-order approximation can be inaccurate depending on the situation, while the second-order approximation is consistently accurate. We then showed that the underestimated variance can lead to a significant increase in the FPR.

In our simulations, the variance errors due to the first-order approximation were dependent on parameters such as the N-ratio and the β -ratio. When the number of samples in the target study increased while the number of samples in the external dataset for exposure association was fixed, the errors became larger. This suggested that in future studies, a larger study size may correspond to increased error from the first-order approximation method. Furthermore, as the true causal effect increased, so did errors. Interestingly, the errors appeared to be independent of the MAF.

In this study, we simply assumed the use of a single SNP as an IV in 2SMR. The causal effect between an exposure and an outcome is usually obtained by merging the ratio per variant (β) via the inverse-variance weighted method over a large number of variants. In this extended multi-variant model, we expect that the variance of the final estimate will also be affected by the errors induced by the first-order approximation, because the ratio for all variants is affected regardless of MAF. Then, the standard error of the causal effect, $\hat{\beta}$, would be dependent on the same parameters (N-ratio and the magnitude of beta) as in the extended model. Some other issues, such as linkage disequilibrium and pleiotropy, should also be addressed in the extended multivariate model.

Overall, our study suggests that the use of the second-order ap-

proximation is always preferable, since it provides an accurate estimate of the variance regardless of the situation. However, when the IV-exposure association is much greater than the IV-outcome association (i.e., β is very small), we observed no significant difference between the first- and second-order approximations. Therefore, we expect that whether one must apply the second-order approximation to avoid an increased FPR will depend on many factors, including the actual range of β .

ORCID

Hakin Kim: <https://orcid.org/0000-0003-2186-015X>

Kunhee Kim: <https://orcid.org/0000-0001-7899-7961>

Buhm Han: <https://orcid.org/0000-0002-2266-5164>

Authors' Contribution

Conceptualization: BH. Data curation: HK. Formal analysis: HK. Funding acquisition: BH. Methodology: BH, KK, HK. Writing - original draft: HK, KK, BH. Writing - review & editing: HK, BH.

Conflicts of Interest

Buhm Han is the CTO of Genealogy Inc.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) (grant number 2022R1A2B5B02001897) funded by the Korean government, Ministry of Science and ICT. This work was supported by the Creative-Pioneering Researchers Program funded by Seoul National University (SNU).

Supplementary Materials

Supplementary data can be found with this article online at <http://www.genominfo.org/>.

References

1. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013;37:658-665.
2. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;32:1-22.

3. Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. *Int J Epidemiol* 2016;45:908-915.
4. Thomas DC, Lawlor DA, Thompson JR. Re: Estimation of bias in nongenetic observational studies using “Mendelian triangulation” by Bautista et al. *Ann Epidemiol* 2007;17:511-513.
5. Teumer A. Common methods for performing Mendelian randomization. *Front Cardiovasc Med* 2018;5:51.
6. Glymour MM, Tchetgen Tchetgen EJ, Robins JM. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol* 2012;175:332-339.
7. VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. Methodological challenges in mendelian randomization. *Epidemiology* 2014;25:427-435.
8. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res* 2017;26:2333-2355.