Two-Sample Mendelian Randomization with Summary Data

Xinyi Pan
Master in Interdisciplinary Data Science
Duke University
xinyi.pan@duke.edu

Abstract

This project reviews the two-sample Mendelian randomization (MR) methods with GWAS summary data to estimate causal effects in the presence of unmeasured confounding. In an applied experiment, the random-effect inverse-variance weighted (IVW) method is employed to estimate the effect of body mass index on coronary heart disease, which is 0.45 with a standard error of 0.06. Compared with other MR methods, IVW provides the causal estimate with the lowest uncertainty. The sensitivity analysis indicates that no single SNP was dramatically driving the overall causal estimate in the leave-one-out analysis, but that heterogeneity among SNPs exists.

1 Introduction

Understanding the causal mechanisms of disease is at the heart of epidemiologic research. If a factor is known to be causally responsible for an adverse health outcome, an intervention can be developed to reduce the factor and improve population health [1]. The randomized controlled trial (RCT) is the gold standard of causal inference because it takes care of unmeasured confounding by design. However, many causal questions in epidemiology are not amenable to evaluations with RCTs [2]. Observational data are used to infer causality in the absence of RCTs. Conventional observational studies is to control for confounding variables between the exposure and outcome, which, however, can lead to biased results [2] due to unmeasured confounding. Therefore, different tools are in an urgent need to bypass these biases and therefore understand causal mechanisms. Mendelian randomization (MR) studies provide an affordable approach to evaluating causality and have increased rapidly in the spectrum of applications owing to the increasing availability of data.

1.1 Mendelian randomization

Mendelian randomization (MR) imitates the RCT design by using a genetic variation to address causal questions on how modifiable exposures influence different outcomes [3, 2]. MR is applied with exposures that reliably associate with individuals' genetic variants, including measurable characteristics such as body mass index (BMI), although MR is usually implemented with single-nucleotide polymorphisms (SNPs), which act as instrumental variables (IVs) [2]. The principles of MR are based on Mendel's first and second laws of inheritance and IV estimation methods.

The MR method draws on Mendel's first and second laws: the law of segregation and the law of independent assortment. The law of segregation states that offspring randomly inherit one allele among those separated at meiosis. The law of independent assortment indicates that the alleles for separate traits are transmitted independently of each other to offspring [4]. Given Mendel's laws of inheritance, the alleles an offspring receives at the corresponding SNP are expected to be random with regard to the potential confounders and the causal upstream from the exposure to the outcome. Observing an individual's genetic variation at this SNP is similar to randomizing individuals to

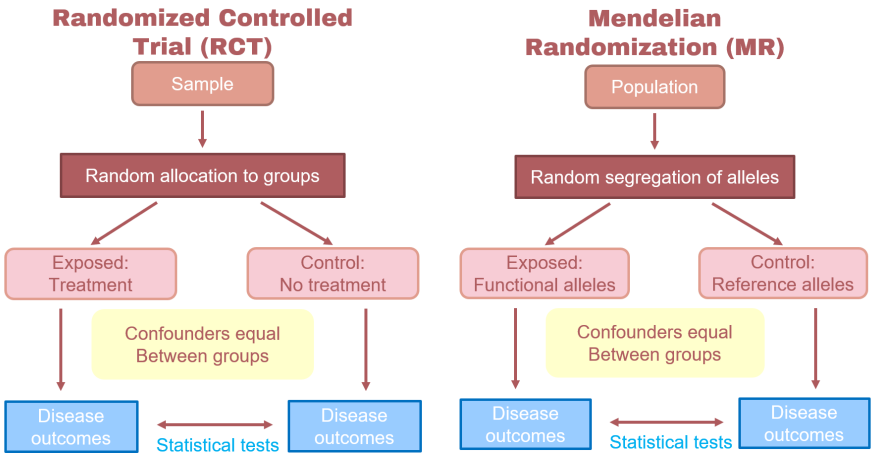

Figure 1: Comparison between MR and RCTs. Randomization in MR is owing to the random allocation of alleles [5].

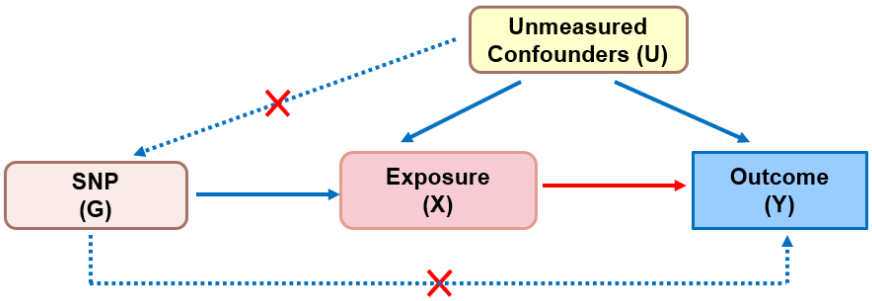

Figure 2: Illustration of instrumental variable estimation of MR. The arrow highlighted in red represents the causal effect of interest.

treatment or control groups in RCTs [5]. Therefore, the SNP can be considered as an IV to conduct causal analysis in this natural experiment. The overview of MR studies is summarized in Figure 1.

1.2 IV and the resulting assumptions of MR

Statistical methodology for MR is generally based on instrumental variable (IV) analysis. The IV approach was proposed to estimate causal effect in the presence of unmeasured confounding of the exposure and outcome. Assumptions are required for an IV analysis to be valid, the first three of which are stated as follows.

1. **Relevance assumption**: the IV is associated with the exposure.
2. **Independence assumption**: the IV is not associated with confounding variables.
3. **Exclusion restriction assumption**: the IV is only associated with the outcome through exposure.

The plausibility of MR estimation also relies on these three assumptions [2]. The first relevance assumption is easy to satisfy in the genotype case, while the other two are not provable but can be examined through sentiment analysis. The instrumental variable estimation of MR is illustrated in Figure 2. At least one additional assumption is required for point-estimate identification, which however remains an area of debate and research development [2]. Usually monotonicity, which assumes that the direction of the effect of the genetic variant on the exposure is the same for everyone [2], is relevant for MR estimation among other point-estimate-identifying conditions.

1.3 GWAS summary data

MR estimates can be made using summary-level data on the association of each genetic instrument with the exposure and with the outcome phenotype of interest [2]. Summary data can be obtained from genome-wide association studies (GWAS), which estimate the association of SNPs with the exposure and with the outcome trait. GWAS are hypothesis-free experimental designs in which a set of (from hundreds of thousands to millions) of gene variants is systematically tested for association with a single trait or disease outcome [6]. The main goal of GWAS is to identify specific variants that can be used to predict traits, as well as to highlight genes or loci associated with the etiology of traits or diseases [7]. GWAS are conducted using a strict significance threshold after Bonferroni correction. The threshold is so-called genome-wide significance, which is defined as having a statistical association with a p-value $< 5.0 \times 10^{-8}$ [6]. All genetic variants significantly associated with the trait at genome-wide significance are commonly used as instrumental variables in MR studies. In this project, GWAS summary data are obtained from MR-Base, a database platform developed by Hemani et al. [5].

1.4 Two-Sample MR

When summary data are used, it is natural to consider the two-sample MR analysis for causality estimation, where the association of SNPs with the exposure can be estimated in a sample other than that used to estimate the association of SNPs with the outcome [8]. The ability to perform MR analyses using two different samples largely broadens the scope of MR studies, as it can be difficult to find a dataset with data on the genetic instrument, exposure, and outcome are available simultaneously [2]. Two-sample MR requires genetic instruments to be robustly associated with the exposure. Many MR studies select SNPs identified in GWAS as genome-wide significant predictors of the exposure of interest for estimation [5]. A further assumption of two-sample MR estimation is that the two samples are from the same underlying population. To meet this assumption, two-sample methods typically use data from the most similar populations possible, involving genetic ancestry and, the prevalence of environmental exposures and the time frame over which measurements were made [2].

This study will next elaborate the process to conduct two-sample MR methods with GWAS summary data for causal estimation in Section 2. An applied example is illustrated in Section 3 to estimate the effect of body mass index on coronary heart disease. Section 4 discusses the strengths and the limitations of the existing MR methods in brief.

2 Methodology

This section details the steps and models to conduct the two-sample MR analysis. The analysis utilizes GWAS summary data and is supported by R packages 'TwoSampleMR' and 'MRInstruments' curated by MR-Base.

2.1 Define instruments

First, genetic instruments are identified as SNPs that robustly associated with the exposure at a genome-wide significance (p-value $< 5.0 \times 10^{-8}$). Such instruments are obtained from the MR-Base database and the MRInstruments package. Meanwhile, an approach is adopted to ensure that the genetic instruments selected for an exposure are independent. It is often the case that a locus includes only one independently associated, or causal, variant, while other variants are associated because they are genetically linked to the causal variant. This genetic linkage is expressed as linkage disequilibrium (LD), a measure of how related each pair of alleles are. An LD value of 1 indicates that the alleles are perfectly related, which means they are always inherited together, while an LD value of 0 shows that the alleles are in linkage equilibrium and are inherited completely independently [9]. To assess which SNP instruments are independent, the LD clumping approach is taken to report the most significant genetic associations in a region with a small number of clumps of genetically linked SNPs [5].

2.2 IV analysis

To determine whether and how the treatment influences the outcome, it can be initially regarded as an intention-to-treat (ITT) analysis by comparing the outcome measures among individuals randomly assigned to the treatment and among those randomly assigned to the control [10]. However, beyond ITT analysis, we should also consider the extent to which the instrument influences the exposure to derive a valid estimate of the effect of the exposure on the outcome. Therefore, MR can take IV as an alternative approach, which estimates the ratio between SNP-outcome association and SNP-exposure estimation:

$$\gamma_1 = \frac{\mathbb{E}(Y \mid G = 1) - \mathbb{E}(Y \mid G = 0)}{\mathbb{E}(X \mid G = 1) - \mathbb{E}(X \mid G = 0)}, \tag{1}$$

where γ_1 is the Wald ratio estimator, Y is the outcome, X is the exposure, and G is the instrument. As discussed in Section 1.1, the randomization in MR is due to the random allocation of alleles [2]. Under the IV assumptions described in Section 1.2, this estimator provides a statistical test of whether there is a causal effect of the exposure on the outcome.

2.3 Harmonize effects

Next, the SNP-outcome and SNP-exposure effects are estimated with the corresponding standard errors and harmonized to reflect the same effect alleles. If an SNP of interest is not present in the outcome sample, SNPs highly correlated to this risk variant in the MR-Base database are used as LD proxies [11] for this risk variant. Moreover, it is important to ensure that the SNP-outcome and SNP-exposure associations obtained are from the same effect alleles [12], which is done by harmonization. The following scenarios are considered to be harmonized, wrong effect alleles, strand issues, palindromic SNPs, and incomplete alleles. In the first case of wrong effect alleles, an SNP with effect/non-effect alleles G/T for exposure and T/G for the outcome is harmonized by flipping the sign of the SNP outcome effect. As for strand issues, the alleles from the outcome sample are flipped to match those from the exposure sample. To identify the effect alleles from the exposure and outcome samples, the presence of palindromic SNPs also introduces ambiguity since their alleles are represented by the same pair of letters but on the strands of opposite directions [5]. The frequency of effect alleles is used to switch the direction of the effect alleles in either the exposure or outcome sample to make sure the alleles are aligned and thereby resolve the ambiguity. If there are incompatible alleles, however, the SNP will be excluded from the MR analysis.

2.4 IVW meta-analysis

The generated summary sets can then be analyzed by a range of methods. Multiple independent instruments can further improve the MR analysis by explaining a larger proportion of variance in the exposure [5]. In this project, the major approach to obtain the MR estimate using multiple SNPs is the inverse variance weighted (IVW) meta-analysis of each Wald ratio. After calculating the estimated effect of genetic variant l on the exposure, $\widehat{\pi}_l$, with variance $\sigma^2_{x,l}$ and the estimated effect of genetic variant l on the outcome, $\widehat{\Gamma}_l$, with variance $\sigma^2_{y,l}$, the variant-specific Wald ratio is estimated by Equation 2 for each variant l.

$$\hat{\beta}_l = \frac{\widehat{\Gamma}_l}{\widehat{\pi}_l}. \tag{2}$$

These Wald ratios are then weighted by their corresponding variance, which gives the IVW estimator $\hat{\beta}_{\mathrm{IVW}}$ [5] in Equation 3.

$$\hat{\beta}_{\mathrm{IVW}} = \frac{\sum_{l=1}^{L} \widehat{\pi}_l \widehat{\Gamma}_l \sigma^{-2}_{y,l}}{\sum_{l=1}^{L} \widehat{\pi}_l{}^2 \sigma^{-2}_{y,l}}, \tag{3}$$

where L is the total number of genetic variants identified as potentional IVs. A random-effect IVW model is used in this project to take into account the heterogeneity between SNPs, which is implemented by MR-Base.

Method	MR Causal Estimate	Std. Err.	P-value
MR Egger	0.5024935	0.14396056	8.012590e-04
Inverse variance weighted	0.4459091	0.05898302	4.032020e-14
Weighted median	0.3870065	0.07179142	7.018111e-08
Weighted mode	0.3888249	0.09642239	1.275765e-04

Table 1: Two-sample MR estimates of the effect of BMI on CHD with different MR methods.

2.5 Sensitivity analyses

In addition to IVW, other methods are considered to compare the estimate results with the IVW estimation, including MR Egger analysis [13], weighted-medium estimator [14], and weighted-mode estimator [15].

Moreover, leave-one-out meta-analysis is conducted to analyze the sensitivity of MR estimation. Specifically, the causal effect is re-estimated by leaving out exactly one SNP at a time to evaluate if the MR estimate is driven or biased by a single SNP, which may have a horizontal pleiotropic effect [5]. Leave-one-out analysis shows how a single SNP affects the overall estimate, and identifying SNPs that result in a significant change in the estimate when dropped is helpful in evaluating the estimation sensitivity to outliers.

3 Experiment

In this project, an MR study is conducted on the causal effect of body mass index (BMI) on coronary heart disease (CHD) using GWAS summary data. There are 79 SNPs selected as instruments in this MR study. The results of two-sample MR analysis with different methods are summarized in Table 1. It can be seen that the point estimate of the causal effect approximately ranges between 0.4 and 0.5, which is further illustrated in Figure 4. Figure 4 relates the effect sizes of the SNP-BMI association (x-axis) and the SNP-CHD associations (y-axis). The error bars around each point show the standard error of the estimated SNP-BMI association and the SNP-CHD association. The intercepts of the lines fitted by the four methods are all positive. The slopes of the lines show the causal effects estimated by each of the four different methods. Among the causal estimates, MR Egger returns the largest estimate with the highest uncertainty, while IVW gives the causal estimate of 0.45 with the smallest standard error of 0.06. It is reasonable since MR Egger is based on the IVW analysis but relaxes the IV assumption of the exclusion restriction [13]. Similar results to the IVW estimate were provided by the weighted median and weighted mode estimators.

In terms of sensitivity analysis, a forest plot 5 is first drawn to diagnose if there is any heterogeneity among SNPs. In Figure 5, each black point stands for the log odds ratio for CHD per standard deviation increase in BMI, produced using each SNP as separate instruments, and the red points show the aggregated causal estimate using all SNPs together with MR Egger and IVW. Also, the Cochran's Q value of the random-effect IVW is 143.6508 with p-value 8.72e-06, which suggests strong evidence for heterogeneity among SNPs. In a following leave-one-out analysis reffig:leaveoneout, one genetic variant (SNP) is sequentially dropped at a time to examine the sensitivity of the results to individual SNPs. Each black point stands for the IVW estimate of the causal effect of BMI on CHD, excluding that particular variant from the analysis. The red point represents the IVW causal estimate using all SNPs. Figure reffig:leaveoneout shows that no single SNP was dramatically driving the overall effect of BMI on CHD.

4 Conclusion

To conclude, this project reviews the two-sample Mendelian randomization methods with summary data to estimate causal effects in the presence of unmeasured confounding. In an applied experiment, the random-effect inverse-variance weighted method is employed to estimate the effect of body mass index on coronary heart disease, which is 0.45 with a standard error of 0.06. Compared with other MR methods, IVW provides the causal estimate with the lowest uncertainty. The sensitivity analysis indicates that there is strong heterogeneity among SNPs, but that no single SNP was dramatically driving the overall effect in the leave-one-out analysis.

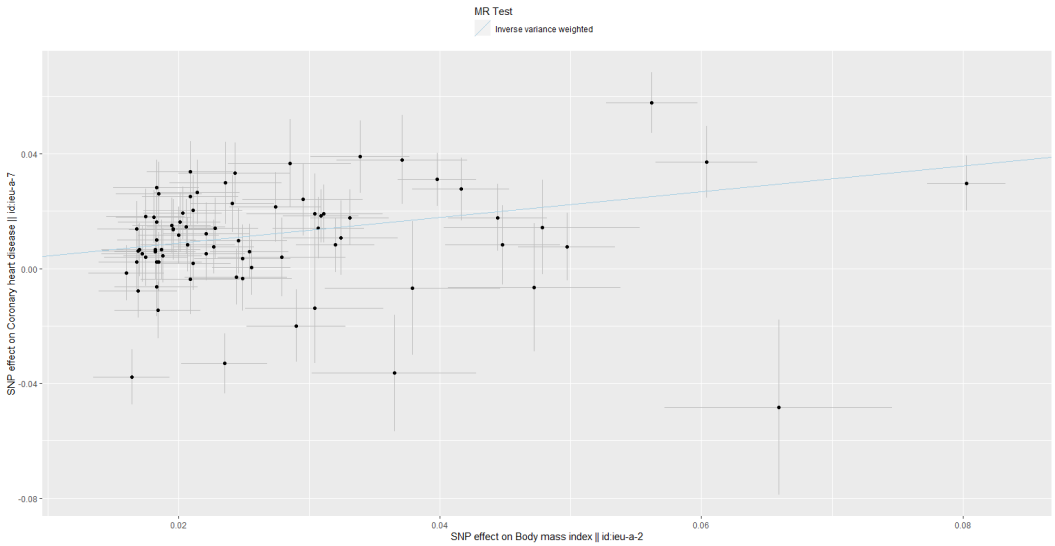

Figure 3: A scatter plot of the SNP-BMI and SNP-CHD associations for each SNP with an IVW-estimated line fitted

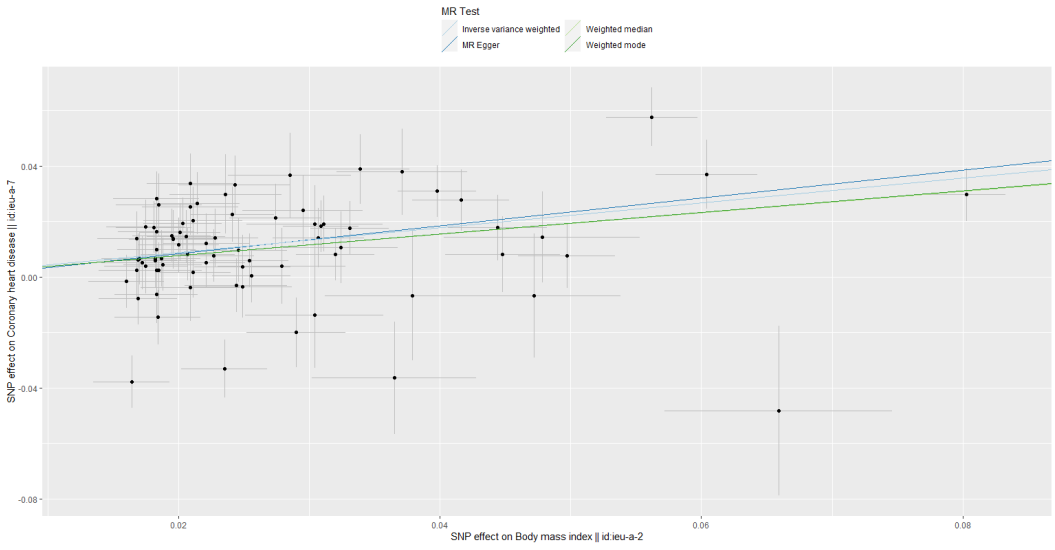

Figure 4: A scatter plot with lines fitted by the approaches of IVW, weighted mode, weighted median, and MR Egger

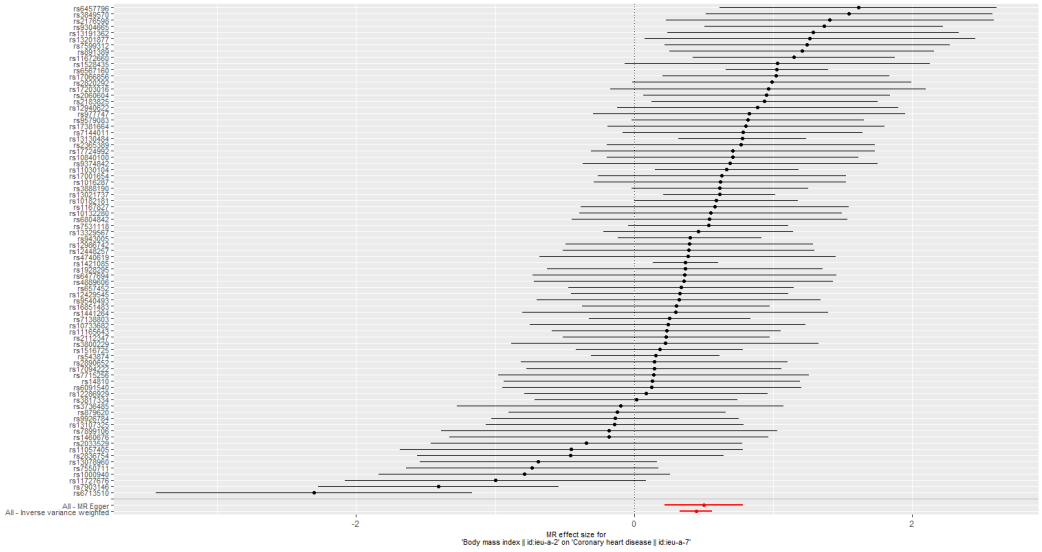

Figure 5: A forest plot of two-sample MR on each SNP individually as well as pooled results from IVW and MR Egger respectively

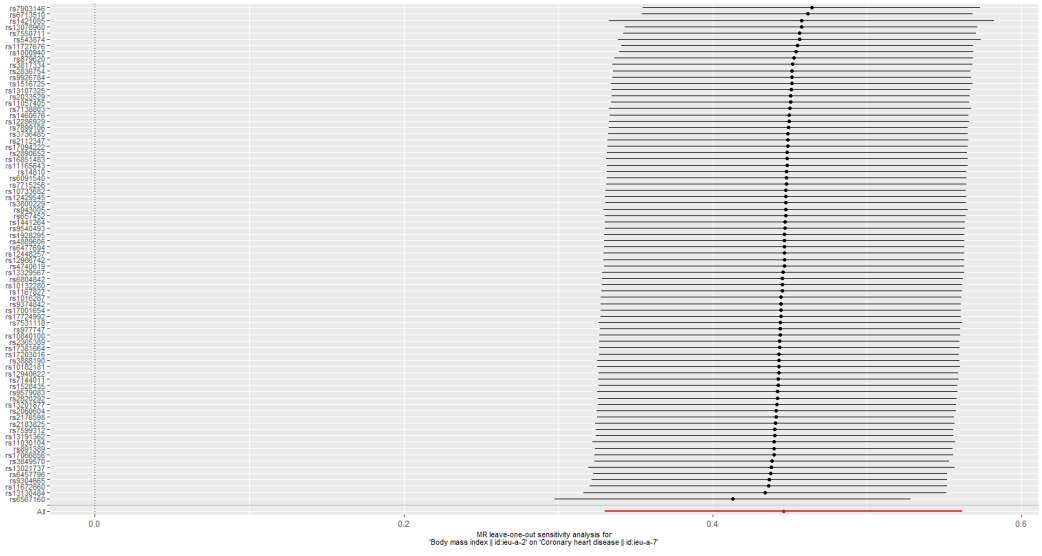

Figure 6: A leave-one-out analysis where the IVW estimate is recalculated

MR with summary data has greatly expanded the range of MR analyses that can be done and overcomes the original barriers to MR caused by lack of data. Further, using multiple genetic instruments in combination enhances the statistical power owing to the increased fraction of the exposure variance explained by the instruments. However, none of the MR methods employed in this study is truly robust to all types of pleiotropy. Also, the majority of MR assumptions are not provable but can only be examined via sentiment analysis. Thus, methods should be selected based on the most relevant assumptions for the question of interest and tested by sentiment analysis to determine the robustness of MR results.

References

[1] Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S Small. Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, 48(3):1742–1769, 2020.

[2] Eleanor Sanderson, M Maria Glymour, Michael V Holmes, Hyunseung Kang, Jean Morrison, Marcus R Munafò, Tom Palmer, C Mary Schooling, Chris Wallace, Qingyuan Zhao, et al. Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):1–21, 2022.

[3] George Davey Smith and Shah Ebrahim. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22, 2003.

[4] George Smith, Holmes, et al. Mendel's laws, mendelian randomization and causal inference in observational data: substantive and nomenclatural issues. *European Journal of Epidemiology*, 35, 02 2020.

[5] Gibran Hemani, Jie Zheng, Benjamin Elsworth, Kaitlin H Wade, Valeriia Haberland, Denis Baird, Charles Laurin, Stephen Burgess, Jack Bowden, Ryan Langdon, et al. The mr-base platform supports systematic causal inference across the human phenome. *elife*, 7:e34408, 2018.

[6] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[7] William S Bush and Jason H Moore. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[8] Stephen Burgess, Robert A Scott, Nicholas J Timpson, George Davey Smith, and Simon G Thompson. Using published data in mendelian randomization: a blueprint for efficient identification of causal risk factors. *European journal of epidemiology*, 30(7):543–552, 2015.

[9] Trine Folseraas, Espen Melum, Philipp Rausch, Brian D Juran, Eva Ellinghaus, Alexey Shiryaev, Jon K Laerdahl, David Ellinghaus, Christoph Schramm, Tobias J Weismüller, et al. Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. *Journal of hepatology*, 57(2):366–375, 2012.

[10] Sandeep K Gupta. Intention-to-treat concept: a review. *Perspectives in clinical research*, 2(3):109, 2011.

[11] Richard J.L. Anney. Chapter 2.3 - common genetic variants in autism spectrum disorders. In Joseph D. Buxbaum and Patrick R. Hof, editors, *The Neuroscience of Autism Spectrum Disorders*, pages 155–167. Academic Press, San Diego, 2013.

[12] Fernando Pires Hartwig, Neil Martin Davies, Gibran Hemani, and George Davey Smith. Two-sample mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique, 2016.

[13] Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, George Davey Smith, Nuala A Sheehan, and John R Thompson. Assessing the suitability of summary data for two-sample mendelian randomization analyses using mr-egger regression: the role of the i2 statistic. *International journal of epidemiology*, 45(6):1961–1974, 2016.

[14] Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American statistical Association*, 111(513):132–144, 2016.

[15] Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data mendelian randomisation via the zero modal pleiotropy assumption. *bioRxiv*, 2017.