Review: Basic Concepts and Practical Considerations for Mendelian Randomization

Sungho Won
Graduate School of Public Health
Seoul National University
Most Science Studies Appear to Be Tainted By Sloppy Analysis

September 14, 2007; Page B1

We all make mistakes and, if you believe medical scholar John Ioannidis, scientists make more than their fair share. By his calculations, most published research findings are wrong.

Dr. Ioannidis is an epidemiologist who studies research methods at the University of Ioannina School of Medicine in Greece and Tufts University in Medford, Mass. In a series of influential analytical reports, he has documented how, in thousands of peer-reviewed research papers published every year, there may be so much less than meets the eye.
Main Biases in Classical Epidemiological Studies

- Types of biases
  - Confounding
  - Information bias (regression dilution bias)
  - Selection bias (collider)

Bennet et al, Heart, 2017
Mendelian Randomization

- The fundamental idea: If we cannot randomize the exposure, we can find a randomized instrumental variable to disentangle
  - Confounding
  - Reverse causation

![Diagram showing the relationship between Genetic Variant, Intermediate exposure, Disease outcome, and Confounders](image-url)
Confounding: One key point is that the distribution of such polymorphisms is largely unrelated to the sorts of confounders socioeconomic or behavioural that were identified above as having distorted interpretations of findings from observational epidemiological studies.”

- Mendel’s second law, the law of independent assortment.
- variants can be viewed as if “randomized” conditional on parental genotypes.

- But it is an approximation in the population!
Figure 3  Comparison of a conventional trial with a Mendelian randomisation study. This illustrates the analogy between a conventional randomised controlled trial and a Mendelian randomisation study. CV, cardiovascular.
low cholesterol $\iff$ increased cancer risk

- Apo E sequence variation $\Rightarrow$ low cholesterol, this relationship is established since inheritance of Apo E sequence variation
- Apo E sequence variation $\Rightarrow$ increase cancer risk, cancer occurs later stage of life

- Two underlying assumptions
  - Apo E genetic effect on cancer risk can be unbiasedly assessed
  - Apo E genetic variation does not increase cancer risk through other pathways.

Example: Katan M. Lancet 1986
Instrumental Variable

- Condition: IV Z does not affect Y except through X, meaning:
  - Z is correlated with Y but does not affect Y directly (called “exclusion restriction”).
  - Z is also correlated with X but not perfectly.

- It’s very hard to find a good Z.

- Idea of Instrumental Variables attributed to Philip Wright (1861-1934).
Figure 1. Use of MR and instrumental variable approaches in the literature increases over time. PubMed Search strategy (June of 2016): for MR analysis, “mendelian random*[tiab]” or “Mendelian Randomization Analysis” (medical subject headings [MeSH]); for instrumental variable analysis or MR analysis, “instrumental variable*[tiab],” “mendelian random*[tiab],” or “Mendelian Randomization Analysis” (MeSH). Note that the MeSH term “Mendelian Randomization Analysis” was introduced by MEDLINE in 2010.
Three Core Assumptions

- Core assumptions

Figure 2. Conceptual illustration of the MR method and its three underlying core assumptions as directed acyclic graphs. (A) Conceptual model. (B) Assumption 1. (C) Assumption 2. (D) Assumption 3.

Pleiotropy

- **Types of Pleiotropy**
  
  a) **Horizontal pleiotropy**
  
  - SNP(s)
  - Telomere length
  - Cancer therapy
  - CHD

  b) **Vertical pleiotropy**
  
  - SNP(s)
  - BMI
  - SBP
  - CHD

- Horizontal pleiotropy is a violation of the instrumental variable assumptions and this is problematic for MR studies.
- Vertical pleiotropy is in general not problematic.

Bennet et al, Heart, 2017
Use inherited genetic variants to infer causal relationship of an exposure and a disease outcome.

Statistical methods for MR
  - One sample Mendelian randomization: Two-stage least squares (2SLS) method
  - Two sample Mendelian randomization
  - MR Egger’s method
One Sample MR
(2SLS)
Concept & Assumption
IV Estimation in Linear Models

- Two more assumptions required for estimation:
  - The effect of $X$ on $Y$ is linear.
  - No interaction between $X$ and $U$.
  - Suppose the data generating models are

\[
X = \alpha_0 + \alpha_1 G + \alpha_2 U + \varepsilon_1, \\
Y = \theta_0 + \theta_1 X + \theta_2 U + \varepsilon_2.
\]

- We can fit the following reduced models

\[
E[X|G] = \alpha_0 + \alpha_1 G, \\
E[Y|G] = \beta_0 + \beta_1 G,
\]
One sample MR: 2SLS Estimator

- **2SLS estimator**
  \[
  \hat{\beta}_{IV} = \left( X^T G (G^T G)^{-1} G^T X \right)^{-1} \left( X^T G (G^T G)^{-1} G^T Y \right)
  \]
  \[
  \sqrt{n}(\hat{\beta}_{IV} - \beta) \sim N(0, \sigma^2 (Q_{GX} Q_{GG}^{-1} Q_{XG})^{-1})
  \]

- First we need to compute
  \[
  \hat{X} = G (G^T G)^{-1} G^T X
  \]

- Then regress \( Y \) on \( \hat{X} \)
  \[
  \hat{\beta}_{IV} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T Y
  \]
  \[
  = [X^T G (G^T G)^{-1} G^T X]^{-1} X^T G (G^T G)^{-1} G^T Y
  \]

- Any regression software can be used to get 2SLS estimator, just compute the variance
Weak Instrument

- It can occur when using one or more genetic variants that only explain a small proportion of the variation in the risk factor, coupled with a small sample size.

- For weak instrument, we have the following problems:
  - Very little statistical power to test hypotheses.
  - Bias due to violations of the core instrumental variable assumptions, such as horizontally pleiotropic effects of variants, will be amplified.
  - Even when using very large samples, results using weak instruments are biased towards the outcome-risk factor association in the single sample setting and towards the null in the two sample setting. Precision (assessed with confidence intervals) is underestimated.

- It can be tested by the F statistics and the thumb of rule is the F statistic > 10.
We can detect pleiotropy and the validity of IV if

- The number of IVs \( I \) is more than the number of causal effects \( p \) to be estimated; not all \( I \) equations can be exactly zero.
- The null hypothesis is \( G \perp (Y - X\beta) \)
  - Instrument is orthogonal to the error term.
  - There is no direct effect left once conditional on \( X \).
- Sargan’s test (1958, JASA) for 2SLS for \( I \) instrumental variables and 1 causal effect

\[
\{G(Y - \hat{\theta}_{2SLS}X)\}^T \{\hat{\sigma}_2(G^T G)\}^{-1} \{G(Y - \hat{\theta}_{2SLS}X)\} \rightarrow \chi^2(I - 1)
\]

under the null that all instruments are valid.
Hansen (1982) gave general results

\[ J_n(\hat{\beta}) = n\bar{g}_n(\hat{\beta})^T \hat{W}_n\bar{g}_n(\hat{\beta}) \rightarrow \chi^2(I - p) \]

As long as \( W_n \) converges to the optimal \( W_0 \) and is efficient GMM estimator.

- Large J-statistic will reject null hypothesis so that at least one instrument might be invalid.
- Report this J-statistic whenever there are overidentifying condition for IV.
Equality between IV and OLS

- If there is no unmeasured confounding, OLS estimator will be consistent and efficient; IV is consistent under null or alternative.

- Large discrepancy between estimated coefficients for OLS and IV suggests that there is confounding and OLS cannot be trusted.

- Wu-Hausman test (test for endogeneity)

  \[ H_0: \text{cov}(X, U) = 0 \; \text{vs} \; H_1: \text{cov}(X, U) \neq 0 \]

- If \( H_0 \) is rejected, then IV approach should be utilized. However if it is not, OLS is preferred.
Let

\[ X = \alpha_0 + \alpha_1 G + \alpha_2 U + \varepsilon_1, \]
\[ Y = \theta_0 + \theta_1 X + \theta_2 U + \varepsilon_2. \]

Fit

\[ Y = \tau_0 + \tau_1 X + \tau_2 \hat{\varepsilon}_1 + \varepsilon_3, \quad \hat{\varepsilon}_1 = X - \hat{\alpha}_1 G \]

then test whether \( \tau_2 = 0 \).

If \( \tau_2 = 0 \), then there is no correlation between \( X \) and \( U \).
Examples

- The main goal is to identify whether circulating plasma CRP levels are associated with a range of metabolic and cardiovascular diseases (all continuous outcomes in the paper) but not necessarily causal.
- CRP haplotype (most likely ones) was used as instrumental variables (likely no other pathway other than circulating CRP)
- CRP haplotype is not associated with potential confounding variables, such as smoking, alcohol, physical activity etc.

C-reactive protein and its role in metabolic syndrome: mendelian randomisation study


Summary
Background Circulating C-reactive protein (CRP) is associated with the metabolic syndrome and might be causally linked to it. Our aim was to generate estimates of the association between plasma CRP and metabolic syndrome phenotypes that were free from confounding and reverse causation, to assess the causal role of this protein.
Examples

- Strong association between CRP haplotypes and plasma CRP (F-stat>10) was found (it is not weak instrument).
- It would be nice to perform a Sargan’s test for validity of instruments.

<table>
<thead>
<tr>
<th>Estimated frequency (SE)</th>
<th>Plasma CRP (mg/L) (geometric mean, 95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGC 0.37 (0.006)</td>
<td>1.81 (1.66–1.96)</td>
</tr>
<tr>
<td>CGT 0.26 (0.005)</td>
<td>1.70 (1.58–1.83)</td>
</tr>
<tr>
<td>CAC 0.30 (0.006)</td>
<td>2.03 (1.90–2.18)</td>
</tr>
<tr>
<td>GGT 0.07 (0.003)</td>
<td>1.39 (1.23–1.56)</td>
</tr>
</tbody>
</table>

Global ANOVA for differences in CRP concentration by haplotype p<0.0001. Haplotypes CAT, GGC, GAC, GAT excluded from table because of inferred frequencies of <1%.

Table 3: Common haplotypes for the CRP region
IV estimators are computed by 2SLS.
Wu-Hausman test for equality of IV and OLS.
These results suggest that there is no causal association between CRP and the metabolic syndrome phenotypes.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Change with doubling of CRP concentrations (linear regression)</th>
<th>Change with doubling of CRP concentration (instrumental variables)</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI (kg/m²)</td>
<td>1.04 (0.94 to 1.14)</td>
<td>-0.44 (-1.34 to 0.46)</td>
<td>0.0002</td>
</tr>
<tr>
<td>Systolic blood pressure (mm Hg)</td>
<td>1.4 (0.9 to 1.9)</td>
<td>-0.9 (-5.3 to 3.5)</td>
<td>0.3003</td>
</tr>
<tr>
<td>Waist-to-hip ratio</td>
<td>0.011 (0.099 to 0.013)</td>
<td>0.005 (-0.007 to 0.016)</td>
<td>0.2388</td>
</tr>
<tr>
<td>HDL cholesterol (mmol/L)</td>
<td>-0.064 (-0.073 to -0.055)</td>
<td>0.006 (-0.072 to 0.084)</td>
<td>0.0668</td>
</tr>
<tr>
<td>Triglycerides (mmol/L)†</td>
<td>1.08 (1.07 to 1.09)</td>
<td>0.99 (0.92 to 1.08)</td>
<td>0.0313</td>
</tr>
<tr>
<td>HOMA-R†</td>
<td>1.09 (1.07 to 1.10)</td>
<td>0.94 (0.84 to 1.07)</td>
<td>0.0139</td>
</tr>
</tbody>
</table>

*p* Test of equality of linear regression and instrumental variables estimates. †Ratios of geometric means by a doubling in plasma CRP concentration. Instrumental variables are two-stage least squares estimates with p values to compare between these and ordinary linear regression estimates obtained from Durbin-Wu-Hausman test. Results were similar with two other instrumental variable estimators and corresponding tests.
Two Sample MR

Concept & Assumption
Observe that

\[ \beta_1 = E[Y | G = g + 1] - E[Y | G = g] \]
\[ = \theta_1(E[X | g + 1] - E[X | g]) + \theta_2(E[U | g + 1] - E[U | g]) \]
\[ = \theta_1 \alpha_1. \]

Therefore \[ \theta_1 = \hat{\beta}_1 / \hat{\alpha}_1 \]

When there is one causal effect, one instrument, the IV estimator can be written as the ratio of two OLS estimator

\[ \hat{\beta}_{IV} = \hat{\beta}_1 / \hat{\alpha}_1 \]

The variance of \( \hat{\alpha}_1 \) is important; it can be highly variable in small samples!
Two Sample MR

- Two sample MR with a single variant

\[
\text{WALD} = \frac{\beta_{GY}}{\beta_{GX}}
\]

Causal estimate using Wald method:

\[
\frac{\Gamma_j}{\gamma_j} = \frac{\beta \gamma_j}{\gamma_j} = \beta.
\]
Two Sample MR

- Two sample MR with multiple variants
  - Inverse variance weighted (IVW) method from summarized data is utilized.

\[
\frac{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Y_j}^{-2} \hat{\beta}_j}{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Y_j}^{-2}} = \beta.
\]

where \( \hat{\beta}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_j} \) is the ratio method estimate for variant \( j \), and \( \sigma_{Y_j} \) is the standard error in the regression of the outcome on the \( j \)th genetic variant, assumed to be known.
MR Egger
Concept & Assumption
What is the Problem of IVW?

- Pleiotropic variants affect biological pathways other than the exposure under investigation and therefore can lead to biased causal estimates and false positives under the null.
- IVW method has a Increased power but greater potential for pleiotropy
Main Idea of MR Egger

\[ Y_i = \Gamma_j G_{ij} + \varepsilon_i^Y \]
\[ = (\alpha_j + \beta \gamma_j) G_{ij} + \varepsilon_i^Y. \]

Single variant Wald estimate:
\[ \beta_j = \beta + \frac{\alpha_j}{\gamma_j}. \]

Multiple variant
2SLS / IVW :
\[ \beta + \frac{\sum_{j=1}^{J} \gamma_j \sigma_{Y_j}^{-2} \alpha_j}{\sum_{j=1}^{J} \gamma_j^2 \sigma_{Y_j}^{-2}} = \beta + \text{Bias}(\alpha, \gamma). \]
Egger Regression: Central Concept

- MR Egger method
  - For $L$ different SNPs, we assume $\hat{\gamma}_j, \hat{\Gamma}_j$ ($j = 1, \ldots, L$) are available. Then, MR Egger fits the following regression model:

  $$\hat{\Gamma}_j = \beta_{0E} + \beta_{E} \hat{\gamma}_j + \epsilon_j, \epsilon_j \sim N(0, var(\hat{\Gamma}_j))$$

- $\beta_{0E}$: Identify the presence of ‘directional’ pleiotropy (biasing the IV estimate)

- $\beta_{E}$ provide a biased causal estimate and in the presence of pleiotropy, it is less biased compared to IVW method.

Bowden et al, Int J Epi, 2015
InSIDE Assumption

InSIDE assumption: MR Egger method requires relaxing the third assumption for IV approach

\[ Y_i = \Gamma_j G_{ij} + \epsilon_{ij}^Y \]
\[ = (\alpha_j + \beta \gamma_j)G_{ij} + \epsilon_{ij}^Y. \]

We explore the condition that the correlation between the genetic associations with the exposure (the \( \gamma_j \) parameters) and the direct effects of the genetic variants on the outcome (the \( \alpha_j \) parameters) is zero. We refer to the condition that the distributions of these parameters are independent as InSIDE (Instrument Strength Independent of Direct Effect). It can be viewed as a weaker version of the exclusion restriction assumption.
\( I^2 \) Statistics

- IVW and MR Egger use weights that consider SNP-exposure associations to be known, rather than estimated.
- When a causal effect exists, the MR-Egger estimate of causal effect is biased towards the null when NOME is violated.
- If \( I^2 < 0.9 \), the causal effect should be interpreted with caution.

\[
I_{GX}^2 = \frac{(Q_{GX} - (L - 1))/Q_{GX}}{} \\
Q_{GX} = \frac{\sum_{j=1}^{L} (\bar{y}_j - \bar{y})^2}{\sigma_{Xj}^2}
\]

Example

**ALL INVALID INSTRUMENTS INSIDE ASSUMPTION SATISFIED**

---

**SNP** – exposure association

SNP – outcome association

Increasing instrument strength

Bowden et al, Int J Epi, 2015
Software

- STATA: ivreg2
- R package: AER (ivreg function), gmm, Mendelian Randomization, TwoSampleMR, MR-Base, PhenoScanner

<table>
<thead>
<tr>
<th>Data source</th>
<th>Description</th>
<th>Number of traits</th>
<th>Integrated with statistics package?</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR-Base</td>
<td>A curated database of genome-wide association study results with integrated R package for MR(^{23})</td>
<td>Over 1000</td>
<td>Yes</td>
</tr>
<tr>
<td>PhenoScanner</td>
<td>A curated database of genome-wide association study results with integrated R package for MR(^{37})</td>
<td>Over 500</td>
<td>Yes</td>
</tr>
<tr>
<td>GWAS catalog</td>
<td>Searchable database of genome-wide association study results(^{38})</td>
<td>Over 24,000</td>
<td>No</td>
</tr>
</tbody>
</table>
Database

- Publicly available data sources

<table>
<thead>
<tr>
<th>Consortium name</th>
<th>Description</th>
<th>Most recent sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCAC\textsuperscript{24}</td>
<td>Breast cancer</td>
<td>256 123</td>
</tr>
<tr>
<td>CARDioGRAMplusC4D\textsuperscript{25}</td>
<td>Coronary artery disease and myocardial infarction</td>
<td>184 305</td>
</tr>
<tr>
<td>CKDGen\textsuperscript{26}</td>
<td>Chronic kidney disease</td>
<td>111 666</td>
</tr>
<tr>
<td>DIAGRAM\textsuperscript{27}</td>
<td>Diabetes</td>
<td>159 208</td>
</tr>
<tr>
<td>EAGLE\textsuperscript{28}</td>
<td>Antenatal and early life and childhood phenotypes</td>
<td>47 541</td>
</tr>
<tr>
<td>EGG\textsuperscript{29}</td>
<td>Early growth</td>
<td>153 781</td>
</tr>
<tr>
<td>GIANT\textsuperscript{30}</td>
<td>Height, BMI, and other adiposity traits</td>
<td>693 529</td>
</tr>
<tr>
<td>GLGC\textsuperscript{31}</td>
<td>Global lipids genetics consortium</td>
<td>331 368</td>
</tr>
<tr>
<td>ISGC\textsuperscript{32}</td>
<td>Stroke</td>
<td>84 961</td>
</tr>
<tr>
<td>MAGIC\textsuperscript{33}</td>
<td>Glucose and insulin related traits</td>
<td>224 459</td>
</tr>
<tr>
<td>PGC\textsuperscript{34,35}</td>
<td>Psychiatric genetics, alcohol and tobacco, and other related traits</td>
<td>&gt;500 000</td>
</tr>
<tr>
<td>SSGAC\textsuperscript{36}</td>
<td>Educational attainment and wellbeing</td>
<td>293 723</td>
</tr>
</tbody>
</table>

Davis et al, BMJ, 2018
For the present, however, it is probably fair to say that the method offers a more robust approach to understand the effect of some modifiable exposures on health outcomes than dose much conventional observational epidemiology. Where possible randomized controlled trials remain the final arbiter of the effects of interventions intended to influence health, however.

Smith & Ebrahim 2003 IJE.
References

Questions??