Statistical inference in Mendelian randomization: From genetic association to epidemiological causation

Qingyuan Zhao

Department of Statistics, The Wharton School, University of Pennsylvania

March 1st, 2018 @ UMN

Based on joint work with

- Jingshu Wang, Dylan Small (Penn).
- Jack Bowden, Gibran Hemani (University of Bristol).
Epidemiology = Why certain people are getting ill?

Central problem: Causal inference from observational data.

Methodological challenge: “Correlation does not imply causation”.

Conventional solution: Conditioning on confounding variables.
- Matched case-control studies.
- Cohort studies.

Problem: Biased if there is unmeasured confounder.
Motivating example

Question
What is the causal effect of Body Mass Index (BMI) on Systolic Blood Pressure (SBP)?

Exposure variable \( X \): BMI.

Outcome variable \( Y \): SBP.

Confounders \( C \): age, gender, lifestyle (food, smoking, ...) ...

Fundamental problem
Where is the end of this list?
Mendelian randomization (MR) is an alternative design of observational studies. Instead, MR uses additional “instrumental variables” $Z$. Not necessary to enumerate $C$.

**Brief history of MR**

- Hernán and Robins (2006): An Epidemiologist’s dream?
- Publications per year: only 5 in 2003, almost 400 in 2017.
This talk: MR using summary data

### Individual-level data (restricted access)

**Exposure variable** $X$: BMI; **Outcome variable** $Y$: SBP; **Instrumental variables** $Z_1, \ldots, Z_p$: Genetic variants (SNPs).

### Summary-level data (public)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BMI–FEM</th>
<th>BMI–MAL</th>
<th>SBP–UKBB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>GIANT (female)</td>
<td>GIANT (male)</td>
<td>UKBB</td>
</tr>
<tr>
<td>Sample size</td>
<td>171977</td>
<td>152893</td>
<td>317754</td>
</tr>
<tr>
<td>GWAS</td>
<td>$\text{lm}(X \sim Z_j)$</td>
<td>$\text{lm}(X \sim Z_j)$</td>
<td>$\text{lm}(Y \sim Z_j)$</td>
</tr>
<tr>
<td>Coefficient</td>
<td>Selection</td>
<td>$\hat{\gamma}_j$</td>
<td>$\hat{\Gamma}_j$</td>
</tr>
<tr>
<td>Std. Err.</td>
<td>$\sigma_{Xj}$</td>
<td>$\sigma_{Yj}$</td>
<td></td>
</tr>
</tbody>
</table>

**Step 1** Use BMI–FEM to select significant ($p$-value $\leq 5 \times 10^{-8}$) and uncorrelated SNPs ($p = 25$).

**Step 2** Use BMI–MAL to obtain $(\hat{\gamma}_j, \sigma_{Xj}), j = 1, \ldots, p$.

**Step 3** Use SBP–UKBB to obtain $(\hat{\Gamma}_j, \sigma_{Yj}), j = 1, \ldots, p$. 
Model: Errors-in-variables regression

- Observe mutually independent random variables:
  \[ \hat{\gamma}_j \sim N(\gamma_j, \sigma^2_{X_j}) \text{ and } \hat{\Gamma}_j \sim N(\Gamma_j, \sigma^2_{Y_j}). \]
- For some real number \( \beta_0 \),
  \[ \Gamma_j \approx \beta_0 \gamma_j \text{ for almost all } j. \]
- \( \beta_0 \) can be regarded as the “causal effect” of \( X \) on \( Y \).
**Example**

**Figure:** Scatter plot of $\hat{f}_j$ versus $\hat{\gamma}_j$ in the BMI-SBP example ($p = 25$). The goal is to fit a straight line across the origin.

- Observation: effects are tiny but many are significant.
Reasonable model?

Key assumptions

1. Measurement error is normal.
2. \((\hat{\gamma}_1, \ldots, \hat{\gamma}_p) \perp \perp (\hat{\Gamma}_1, \ldots, \hat{\Gamma}_p)\).
3. \(\hat{\gamma}_j \perp \perp \hat{\gamma}_k\) and \(\hat{\gamma}_j \perp \perp \hat{\Gamma}_k\) if \(j \neq k\).
4. (Implicitly) \(\gamma \neq 0\). Otherwise no information for \(\beta_0\).
5. \(\hat{\gamma}_j\) is unbiased: \(E[\hat{\gamma}_j] = \gamma_j\).

Pre-processing guarantees these assumptions

1. Large sample size \(\Rightarrow\) CLT.
2. Two-sample MR: \(\hat{\gamma}\) and \(\hat{\Gamma}\) are computed from independent samples (GIANT males and UKBB).
3. Selecting uncorrelated and significant SNPs using another independent dataset (GIANT females).
MR estimates **causal** effect

### Two remaining questions

1. Why should the model $\Gamma_j \approx \beta_0 \gamma_j$ be linear?
2. Why is $\beta_0$ causal?

Answer: MR is using SNPs as **instrumental variables (IV)**.

### Heuristic

Consider the following linear model:

\[
X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X C + E_X,
\]

\[
Y = \beta_0 X + \eta_Y C + E_Y = \sum_{j=1}^{p} (\beta_0 \gamma_j) Z_j + \ldots
\]

**Key assumption:** $Z \perp (C, E_X, E_Y)$, i.e. $Z$ are **valid** instruments.
Background: What is an IV?

Figure: Causal effect of $X$ and $Y$ is confounded by $C$. $Z$ is an instrumental variable.

Core IV assumptions

1. $Z$ is associated with the exposure ($X$).
2. $Z$ is independent of the unmeasured confounder ($C$).
3. $Z$ cannot have any direct effect on the outcome ($Y$).
MR = using genetic variants as IVs

Examine the core IV assumptions for MR

1. Criterion 1 is not a problem.
2. Criterion 2 almost comes free due to Mendel’s Second Law of random assortment.
   - A minor concern is population stratification
3. Criterion 3 is not clear.
   - There is a wide-spread phenomenon in genetics called pleiotropy, a.k.a. multi-function of genes.
Linear model (for summary data) is very robust

- Surprisingly, \( \Gamma_j \approx \beta_0 \gamma_j \) is true if the individual-level model for \((X, Y, Z, C)\) are non-linear.

### Heuristic

- Each SNP corresponds to a very small randomized trial.
- Maybe the relationship between \( \Gamma \) and \( \gamma \) is nonlinear, but it is locally linear near \( \gamma = 0 \).
- Can formalize this argument (see the paper) and show:
  \[
  \beta_0 \approx \lim_{\Delta x \to 0} \frac{E[Y(X + \Delta x) - Y(X)]}{\Delta x}.
  \]
- \( Y(x) \) is the counterfactual or “potential outcome”.
- The approximate linear model is still true even if the \( \hat{\Gamma}_j \) are computed from glm(\( Y \sim Z_j \)).
Recap

Two-sample summary-data MR

$$(\hat{\gamma}_j, \hat{\Gamma}_j, \sigma^2_{\chi_j}, \sigma^2_{\gamma_j})_{j=1}^p$$

Genetic associations $\xrightarrow{\text{Inference}}$ Epidemiological causation

Statistical model

1. Observe mutually independent random variables:
   $$\hat{\gamma}_j \sim N(\gamma_j, \sigma^2_{\chi_j})$$ and $$\hat{\Gamma}_j \sim N(\Gamma_j, \sigma^2_{\gamma_j})$$.

2. For some real number $\beta_0$, $\Gamma_j \approx \beta_0 \gamma_j$ for almost all $j$.

Three concrete models for $\alpha_j = \Gamma_j - \beta_0 \gamma_j$

1. No direct effect: $\alpha_j \equiv 0$.
2. Random effects model: $\alpha_j \sim N(0, \tau^2_0)$.
3. Contamination model: $\alpha_j \sim (1 - \epsilon)N(0, \tau^2_0) + \epsilon G$. 
Model 1

Model (No direct effect)

The linear model \( \Gamma_j = \beta_0 \gamma_j \) is true for every \( j = 1, \ldots, p \).

- Log-likelihood of the data (up to additive constant):
  \[
  l(\beta, \gamma_1, \ldots, \gamma_p) = -\frac{1}{2} \left[ \sum_{j=1}^{p} \frac{(\hat{\gamma}_j - \gamma_j)^2}{\sigma^2 X_j} + \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \gamma_j \beta)^2}{\sigma^2 Y_j} \right].
  \]

Profile likelihood

\[
  l(\beta) = \max_{\gamma} l(\beta, \gamma) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta \hat{\gamma}_j)^2}{\sigma^2 X_j \beta^2 + \sigma^2 Y_j}.
  \]

- Profile score (PS): \( \psi(\beta) = (d/d\beta)l(\beta) \).
- Estimator: \( \hat{\beta}_{PS} = \arg \max_{\beta} l(\beta) \) solves \( \psi(\beta) = 0 \).
Some theory

**Assumption**

\[ \sigma^2_{X_j} = O(1/n) = \sigma^2_{Y_j}. \|\gamma\|_2 \text{ is bounded (even if } p \rightarrow \infty). \]

**Heuristic**

Linear structural model: 

\[ X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X C + E_X. \]

- \( \|\gamma\|_2 \) is like the variance of \( X \) explained by \( Z \).

**Theorem (Consistency)**

Under Model 1, \( \hat{\beta}_{PS} \xrightarrow{p} \beta_0 \text{ if } p/n^2 \rightarrow 0. \)
Some theory (continued)

Theorem

Suppose $n \to \infty$, and further $p$ is fixed or $p \to \infty$ but $\|\gamma\|_3/\|\gamma\|_2 \to 0$, then

\[
\frac{V_2}{\sqrt{V_1}} (\hat{\beta}_{PS} - \beta_0) \xrightarrow{d} N(0, 1),
\]

where

\[
V_1 = \sum_{j=1}^{p} \frac{\gamma_j^2 \sigma_{Y_j}^2 + \Gamma_j^2 \sigma_{X_j}^2 + \sigma_{X_j}^2 \sigma_{Y_j}^2}{(\sigma_{X_j}^2 \beta_0^2 + \sigma_{Y_j}^2)^2}, \quad V_2 = \sum_{j=1}^{p} \frac{\gamma_j^2 \sigma_{Y_j}^2 + \Gamma_j^2 \sigma_{X_j}^2}{(\sigma_{X_j}^2 \beta_0^2 + \sigma_{Y_j}^2)^2}.
\]

- $V_1$ and $V_2$ can be consistently estimated from data.
- Necessity IV selection: adding $Z_{p+1}$ hurts if $\gamma_{p+1} = 0$. But adding even very weak SNPs usually reduces variance.
Diagnostic plots

Residual Q-Q plot

Under model 1, standardized residuals follow $N(0, 1)$ if $\hat{\beta} = \beta_0$:

$$\hat{t}_j = \frac{\hat{\Gamma}_j - \hat{\beta}\hat{\gamma}_j}{\sqrt{\hat{\beta}^2 \sigma^2_{Xj} + \sigma^2_{Yj}}}.$$

Influence of a single SNP

- Asymptotic expansion:

$$\hat{\beta}_{PS} = \frac{1 + o_p(1)}{V_2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta_0\hat{\gamma}_j)(\hat{\Gamma}_j\sigma^2_{Xj}\beta_0 + \hat{\gamma}_j\sigma^2_{Yj})}{(\sigma^2_{Xj}\beta_0^2 + \sigma^2_{Yj})^2}.$$

- In practice, we can also use leave-one-out.
Same three datasets, selected 160 uncorrelated SNPs that have $p$-values $\leq 10^{-4}$ in BMI–FEM.

Diagnostic plots:

- Clear overdispersion!
The most likely explanation of the lack of fit is pleiotropy (direct effects).

**Heuristic (Direct effect of $Z_j$ is $\alpha_j$)**

\[
X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X C + E_X,
\]
\[
Y = \beta_0 X + \sum_{j=1}^{p} \alpha_j Z_j + \eta_Y C + E_Y = \sum_{j=1}^{p} (\beta_0 \gamma_j + \alpha_j) Z_j + \ldots
\]

**Model (Random direct effects)**

Assume $\alpha_j = \Gamma_j - \beta_0 \gamma_j \sim i.i.d. \ N(0, \tau_0^2)$ for all $j$.

- $\tau_0^2$ should be small (scales like $1/p$).
We assumed the normal random effects model because of:
- The observed overdispersion.
- Statistical convenience.

This is consistent (or not inconsistent) with genetics:
- Fisher’s (1918) infinitesimal model.
- New perspective on pleiotropy:

  In summary, we conclude that there is an extremely large number of causal variants with tiny effect sizes on height and, moreover, that these are spread very widely across the genome, such that most 100-kb windows contribute to variance in height. More generally, the heritability of complex traits and diseases is spread broadly across the genome (Loh et al., 2015; Shi et al., 2016), implying that a substantial fraction of all genes contribute to variation in disease risk. These observations seem inconsistent.

---

Back to statistics: Failure of profile likelihood

- The profile likelihood of Model 2 is given by

$$l(\beta, \tau^2) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\gamma}_j - \beta \hat{\gamma}_j)^2}{\sigma^2 X_j \beta^2 + \sigma^2 Y_j + \tau^2} + \log(\sigma^2 Y_j + \tau^2),$$

- Easy to verify

$$E\left[ \frac{\partial}{\partial \beta} l(\beta_0, \tau^2_0) \right] = 0.$$

- But the other score function is biased:

$$\frac{\partial}{\partial \tau^2} l(\beta, \tau^2) = \frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\gamma}_j - \beta \hat{\gamma}_j)^2}{(\sigma^2 X_j \beta^2 + \sigma^2 Y_j + \tau^2)^2} - \frac{1}{\sigma^2 Y_j + \tau^2}.$$

- This is not too surprising as we are profiling out \(p\) nuisance parameters \(\gamma_1, \cdots, \gamma_p\) (the Neyman-Scott problem).
There exist many ways to modify the profile likelihood.

We will take the approach of McCullagh and Tibshirani (1990).

Adjusted profile score (APS)

\[ \psi_1(\beta, \tau^2) = -\frac{\partial}{\partial \beta} l(\beta, \tau^2), \]

\[ \psi_2(\beta, \tau^2) = \sum_{j=1}^{p} \sigma_{Xj}^2 \left[ \frac{(\hat{\Gamma}_j - \beta \hat{\gamma}_j)^2}{(\sigma_{Xj}^2 \beta^2 + \sigma_{Yj}^2 + \tau^2)^2} - \frac{1}{\sigma_{Xj}^2 \beta^2 + \sigma_{Yj}^2 + \tau^2} \right]. \]

- Trivial root: \( \beta \rightarrow \pm \infty \) or \( \tau^2 \rightarrow \infty \).
- Let \( \hat{\beta}_{APS} \) be the non-trivial finite solution.
Some theory

**Theorem**

Assume that Model 2 is true, $(\beta_0, p\tau_0^2)$ is in a bounded set $B$, $p \to \infty$ and $p/n^2 \to 0$. Then

1. With probability going to 1 there exists a solution in $B$.
2. All solutions in $B$ are consistent: $\hat{\beta}_{APE} \xrightarrow{p} \beta_0$ and $p\hat{\tau}_{APS}^2 - p\tau_0^2 \xrightarrow{p} 0$.

Can obtain similar asymptotic normality assuming $p/n \to \lambda \in (0, \infty)$ (see the paper).
Example (continued)

- Same 160 SNPs.

- A clear outlier: rs11191593.
- Slightly underdispersed.
Model 3

Model (Random direct effects with outliers)

\[ \alpha_j \sim (1 - \epsilon)N(0, \tau_0^2) + \epsilon G \text{ for some unknown } G. \]

- Recall the expansion:

\[ \hat{\beta}_{PS} = \frac{1 + o_p(1)}{V_2} \sum_{j=1}^{p} \frac{(\hat{r}_j - \beta_0 \hat{\gamma}_j)(\hat{r}_j \sigma_{Xj}^2 \beta_0 + \hat{\gamma}_j \sigma_{Yj}^2)}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2)^2}. \]

- Problem: a single SNP can have unbounded influence (same for APS).

- Our solution: robustify the adjusted profile score.
A method robust to outliers

- Standardized residual:

\[ t_j(\beta, \tau^2) = \frac{\hat{\Gamma}_j - \beta \hat{\gamma}_j}{\sqrt{\sigma^2 X_j \beta^2 + \sigma^2 Y_j + \tau^2}}. \]

**Robust adjusted profile score (RAPS)**

\[
\psi_1^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \rho'(t_j) \cdot \frac{\partial}{\partial \beta} t_j,
\]

\[
\psi_2^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \sigma^2 X_j \frac{t_j \cdot \rho'(t_j) - \delta}{\sigma^2 X_j \beta^2 + \sigma^2 Y_j + \tau^2}.
\]

- \( \rho \) is robust loss, \( \delta = \mathbb{E}[T \rho'(T)] \) for \( T \sim N(0, 1) \).
- Reduces to solution 2 (APS) when \( \rho(t) = t^2 / 2 \).
Some theory

- General theory is very difficult because we are solving nonlinear equations.

**Theorem (Local identifiability)**

\[
\mathbb{E}[\psi^{(\rho)}(\beta_0, \tau_0^2)] = 0 \quad \text{and} \quad \mathbb{E}[\nabla \psi^{(\rho)}] \quad \text{has full rank.}
\]

- Asymptotic normality can be established assuming consistency and additional technical conditions (see the paper).
Example (continued)

- Same 160 SNPs.
- Huber’s loss function.

- Influence of the outlier is limited.
Same 160 SNPs.

Tukey’s biweight function.

Influence of the outlier is essentially 0.
Comparison of all the methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile score (PS)</td>
<td>0.61</td>
<td>0.05</td>
</tr>
<tr>
<td>Adjusted PS (APS)</td>
<td>0.30</td>
<td>0.16</td>
</tr>
<tr>
<td>Robust APS (RAPS, Huber)</td>
<td>0.38</td>
<td>0.12</td>
</tr>
<tr>
<td>Robust APS (RAPS, Tukey)</td>
<td><strong>0.40</strong></td>
<td><strong>0.11</strong></td>
</tr>
<tr>
<td>Inverse variance weighted (IVW)</td>
<td>0.47</td>
<td>0.18</td>
</tr>
<tr>
<td>Weighted median</td>
<td>0.33</td>
<td>0.11</td>
</tr>
<tr>
<td>MR Egger</td>
<td>0.51</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Existing meta-analysis methods for MR

1. IVW: weighted average of $\hat{\Gamma}_j/\hat{\gamma}_j$.
2. Weighted median of $\hat{\Gamma}_j/\hat{\gamma}_j$.
3. MR Egger: weighted least squares of $\hat{\Gamma}_j$ against $\hat{\gamma}_j$.

They all ignore measurement error in $\hat{\gamma}_j$ and weak IV bias.
Simulation setting

- Create simulated datasets to mimic the BMI-SBP example with known $\beta_0$.
- $\hat{\gamma}_j \overset{i.d.}{\sim} N(\gamma_j, \sigma_{\chi_j}^2)$ where $(\gamma_j, \sigma_{\chi_j}^2)$ come from the real data.
- $\Gamma_j$ are generated in 6 different ways ($\beta_0 = 0.5$):
  - Model 1: $\Gamma_j = \gamma_j \beta_0$ (i.e. $\alpha_j \equiv 0$);
  - Model 2: $\alpha_j \overset{i.i.d.}{\sim} N(0, \tau_0^2)$, where $\tau_0 = 2 \cdot (1/p) \sum_{j=1}^p \sigma Y_j$;
  - Model 3: Same as 2 but add $5 \cdot \tau_0$ to $\alpha_1$.
- Heavy-tailed $\alpha$: $\alpha_j \overset{i.i.d.}{\sim} \tau_0 \cdot \text{Lap}(1)$.
- Heteroskedastic $\alpha$: $\text{Var}(\alpha_j) \propto \gamma_j^2$ and normal.
- Many outliers: Add $5 \cdot \tau_0$ to 10% randomly selected $\alpha_j$. \\[35/44\]
Simulation results 1 ($\rho = 25$)

- Black: boxplot of point estimate over 1000 realizations.
- Red: quartile using median standard error.
Simulation results 1 \((p = 160)\)

- Black: boxplot of point estimate over 1000 realizations.
- Red: quartile using median standard error.
Another real data example: BMI-BMI

**BMI-GIANT:** full dataset from the GIANT consortium (i.e. combining BMI-FEM and BMI-MAL), used to select SNPs.

**BMI-UKBB-1:** half of the UKBB data, used as the “exposure”.

**BMI-UKBB-2:** another half of UKBB data, used as the “outcome”.

- Because exposure and outcome, $\gamma_j \equiv \Gamma_j$.
- So true $\beta_0 = 1$ and there is no “direct effect”.
### BMI-BMI results (GIANT, UKBB-1, UKBB-2)

<table>
<thead>
<tr>
<th>$p_{sel}$</th>
<th># SNPs</th>
<th>Mean $F$</th>
<th>IVW</th>
<th>W. Median</th>
<th>W. Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e-9</td>
<td>48</td>
<td>78.6</td>
<td>0.983 (0.026)</td>
<td>0.945 (0.039)</td>
<td>0.941 (0.042)</td>
</tr>
<tr>
<td>1e-8</td>
<td>58</td>
<td>69.2</td>
<td>0.983 (0.024)</td>
<td>0.945 (0.039)</td>
<td>0.939 (0.044)</td>
</tr>
<tr>
<td>1e-7</td>
<td>84</td>
<td>55.0</td>
<td>0.988 (0.024)</td>
<td>0.945 (0.036)</td>
<td>0.933 (0.041)</td>
</tr>
<tr>
<td>1e-6</td>
<td>126</td>
<td>44.1</td>
<td>0.986 (0.022)</td>
<td>0.944 (0.034)</td>
<td>0.931 (0.038)</td>
</tr>
<tr>
<td>1e-5</td>
<td>186</td>
<td>34.3</td>
<td>0.986 (0.019)</td>
<td>0.943 (0.033)</td>
<td>0.928 (0.039)</td>
</tr>
<tr>
<td>1e-4</td>
<td>287</td>
<td>26.1</td>
<td>0.981 (0.017)</td>
<td>0.941 (0.031)</td>
<td>0.929 (0.035)</td>
</tr>
<tr>
<td>1e-3</td>
<td>474</td>
<td>18.8</td>
<td>0.955 (0.015)</td>
<td>0.903 (0.027)</td>
<td>0.917 (0.231)</td>
</tr>
<tr>
<td>1e-2</td>
<td>812</td>
<td>12.7</td>
<td>0.928 (0.014)</td>
<td>0.879 (0.023)</td>
<td>0.739 (7.130)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$p_{sel}$</th>
<th># SNPs</th>
<th>Median $F$</th>
<th>Egger</th>
<th>PS</th>
<th>RAPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e-9</td>
<td>48</td>
<td>51.8</td>
<td>0.926 (0.055)</td>
<td>0.999 (0.023)</td>
<td>0.998 (0.026)</td>
</tr>
<tr>
<td>1e-8</td>
<td>58</td>
<td>42.0</td>
<td>0.928 (0.050)</td>
<td>0.999 (0.023)</td>
<td>0.998 (0.025)</td>
</tr>
<tr>
<td>1e-7</td>
<td>84</td>
<td>32.1</td>
<td>0.905 (0.048)</td>
<td>1.012 (0.021)</td>
<td>1.004 (0.025)</td>
</tr>
<tr>
<td>1e-6</td>
<td>126</td>
<td>27.4</td>
<td>0.881 (0.043)</td>
<td>1.017 (0.019)</td>
<td>1.009 (0.023)</td>
</tr>
<tr>
<td>1e-5</td>
<td>186</td>
<td>21.0</td>
<td>0.874 (0.036)</td>
<td>1.020 (0.018)</td>
<td>1.013 (0.020)</td>
</tr>
<tr>
<td>1e-4</td>
<td>287</td>
<td>15.8</td>
<td>0.921 (0.031)</td>
<td>1.023 (0.017)</td>
<td>1.018 (0.018)</td>
</tr>
<tr>
<td>1e-3</td>
<td>474</td>
<td>10.8</td>
<td>0.913 (0.027)</td>
<td>1.010 (0.016)</td>
<td>1.006 (0.016)</td>
</tr>
<tr>
<td>1e-2</td>
<td>812</td>
<td>5.6</td>
<td>0.909 (0.022)</td>
<td>1.010 (0.015)</td>
<td>1.005 (0.015)</td>
</tr>
</tbody>
</table>
### Selection bias (UKBB-1, UKBB-1, UKBB-2)

**Mean F**

<table>
<thead>
<tr>
<th>$p_{sel}$</th>
<th># SNPs</th>
<th>Mean F</th>
<th>IVW</th>
<th>W. Median</th>
<th>W. Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e-9</td>
<td>110</td>
<td>68.63</td>
<td>0.851 (0.02)</td>
<td>0.83 (0.025)</td>
<td>0.896 (0.046)</td>
</tr>
<tr>
<td>1e-8</td>
<td>168</td>
<td>57.00</td>
<td>0.823 (0.017)</td>
<td>0.8 (0.022)</td>
<td>0.885 (0.053)</td>
</tr>
<tr>
<td>1e-7</td>
<td>228</td>
<td>50.08</td>
<td>0.799 (0.016)</td>
<td>0.768 (0.019)</td>
<td>0.886 (0.058)</td>
</tr>
<tr>
<td>1e-6</td>
<td>305</td>
<td>43.92</td>
<td>0.761 (0.015)</td>
<td>0.736 (0.019)</td>
<td>0.865 (0.079)</td>
</tr>
<tr>
<td>1e-5</td>
<td>443</td>
<td>36.98</td>
<td>0.721 (0.013)</td>
<td>0.667 (0.016)</td>
<td>0.824 (0.12)</td>
</tr>
<tr>
<td>1e-4</td>
<td>652</td>
<td>30.68</td>
<td>0.678 (0.012)</td>
<td>0.616 (0.015)</td>
<td>0.593 (0.122)</td>
</tr>
<tr>
<td>1e-3</td>
<td>929</td>
<td>25.36</td>
<td>0.629 (0.011)</td>
<td>0.57 (0.014)</td>
<td>0.576 (0.096)</td>
</tr>
<tr>
<td>1e-2</td>
<td>1289</td>
<td>20.70</td>
<td>0.592 (0.01)</td>
<td>0.528 (0.013)</td>
<td>0.554 (0.093)</td>
</tr>
</tbody>
</table>

**Median F**

<table>
<thead>
<tr>
<th>$p_{sel}$</th>
<th># SNPs</th>
<th>Median F</th>
<th>Egger</th>
<th>PS</th>
<th>RAPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e-9</td>
<td>110</td>
<td>49.20</td>
<td>1.071 (0.051)</td>
<td>0.871 (0.015)</td>
<td>0.862 (0.021)</td>
</tr>
<tr>
<td>1e-8</td>
<td>168</td>
<td>41.12</td>
<td>1.018 (0.046)</td>
<td>0.848 (0.014)</td>
<td>0.831 (0.018)</td>
</tr>
<tr>
<td>1e-7</td>
<td>228</td>
<td>37.12</td>
<td>1.016 (0.043)</td>
<td>0.824 (0.012)</td>
<td>0.803 (0.016)</td>
</tr>
<tr>
<td>1e-6</td>
<td>305</td>
<td>33.68</td>
<td>1.006 (0.041)</td>
<td>0.793 (0.011)</td>
<td>0.763 (0.016)</td>
</tr>
<tr>
<td>1e-5</td>
<td>443</td>
<td>28.74</td>
<td>0.957 (0.037)</td>
<td>0.762 (0.01)</td>
<td>0.716 (0.015)</td>
</tr>
<tr>
<td>1e-4</td>
<td>652</td>
<td>23.23</td>
<td>0.89 (0.033)</td>
<td>0.724 (0.009)</td>
<td>0.66 (0.014)</td>
</tr>
<tr>
<td>1e-3</td>
<td>929</td>
<td>19.12</td>
<td>0.823 (0.03)</td>
<td>0.687 (0.008)</td>
<td>0.594 (0.013)</td>
</tr>
<tr>
<td>1e-2</td>
<td>1289</td>
<td>15.26</td>
<td>0.749 (0.025)</td>
<td>0.657 (0.008)</td>
<td>0.541 (0.012)</td>
</tr>
</tbody>
</table>
This talk

Paper:

Software:
- R package mr.raps is currently on CRAN.
- Can be directly called from the TwoSampleMR platform (https://github.com/MRCIEU/TwoSampleMR).
Overview of MR (Epidemiology):


Overview of MR (more Stats):


Human genetics:


Robust MR methods (individual-level data):

Robust MR methods (summary-level data):
- Bowden, J. et al. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*.

Pleiotropy: