Mendelian Randomization: Old and New Insights

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

March 26, 2021

Acknowledgement

This talk benefited enormously from multi-disciplinary collaboration with

- Jingshu Wang (Chicago);
- Dylan Small, Nancy Zhang (UPenn);
- Matt Tudball, Gibran Hemani, George Davey Smith (Bristol);
- Jack Bowden (Exeter).

Outline

What is MR?

- A year ago, I would give you this definition by Wikipedia:

  In epidemiology, Mendelian randomization is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in observational studies.

What is MR?

- A year ago, I would give you this definition by Wikipedia:

  In epidemiology, Mendelian randomization is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in observational studies.

- Or I would tell you

  MR = Using genetic variation as instrumental variables.

What is MR?

- A year ago, I would give you this definition by Wikipedia:

  In epidemiology, Mendelian randomization is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in observational studies.

- Or I would tell you

  $$\text{MR} = \text{Using genetic variation as instrumental variables.}$$

- But now I think this view is too narrow.

It all goes back to

It all goes back to

It all goes back to

I am joking... Not quite to the dawn of humankind, but definitely to the dawn of modern statistics and genetics.

Original ideas of (Mendelian) randomization

(a) Gregor Mendel (1822-1884).

(b) Charles Sanders Perice (1839-1914).

(c) Sewall Wright (1889-1988).

(d) Ronald Aylmer Fisher (1890-1962).

Old ideas

Gregor Mendel (1822-1884)

- Mendel conducted a series of pea plant experiments between 1856 and 1863 and established several rules of heredity (now called laws of Mendelian inheritance).
- However, the profound significance of his work was not recognized until 1900.

Old ideas

Gregor Mendel (1822-1884)

- Mendel conducted a series of pea plant experiments between 1856 and 1863 and established several rules of heredity (now called laws of Mendelian inheritance).
- However, the profound significance of his work was not recognized until 1900.

Charles Sanders Perice (1839-1914)

- With Joseph Jastrow, Peirce first introduced blinded, controlled randomized experiments to psychology in 1884.

Old ideas

Gregor Mendel (1822-1884)

- Mendel conducted a series of pea plant experiments between 1856 and 1863 and established several rules of heredity (now called laws of Mendelian inheritance).
- However, the profound significance of his work was not recognized until 1900.

Charles Sanders Perice (1839-1914)

- With Joseph Jastrow, Peirce first introduced blinded, controlled randomized experiments to psychology in 1884.

Sewall Wright (1889-1988)

- Wright introduced causal diagrams and path analysis in 1918.
- In 1920, he used selective inbreeding to investigate genetic causes. In a later defense, he argued that "the universality of Mendelian inheritance under sexual reproduction" justifies causal inference.

Old ideas

Ronald Aylmer Fisher (1890-1962)

- Fisher first put these ideas together and formally introduced randomization as the "reasoned basis for inference" in 1925.

Old ideas

Ronald Aylmer Fisher (1890-1962)

- Fisher first put these ideas together and formally introduced randomization as the "reasoned basis for inference" in 1925.
- He later revealed in the 1951 Bateson Lecture that his factorial method of experimentation derives "its structure and its name from the simultaneous inheritance of Mendelian factors".

Old ideas

Ronald Aylmer Fisher (1890-1962)

- Fisher first put these ideas together and formally introduced randomization as the "reasoned basis for inference" in 1925.
- He later revealed in the 1951 Bateson Lecture that his factorial method of experimentation derives "its structure and its name from the simultaneous inheritance of Mendelian factors".
- Most relevant quote:

  The different genotypes possible from the same mating have been beautifully randomised by the meiotic process. A more perfect control of conditions is scarcely possible, than that of different genotypes appearing in the same litter.

Old ideas

Ronald Aylmer Fisher (1890-1962)

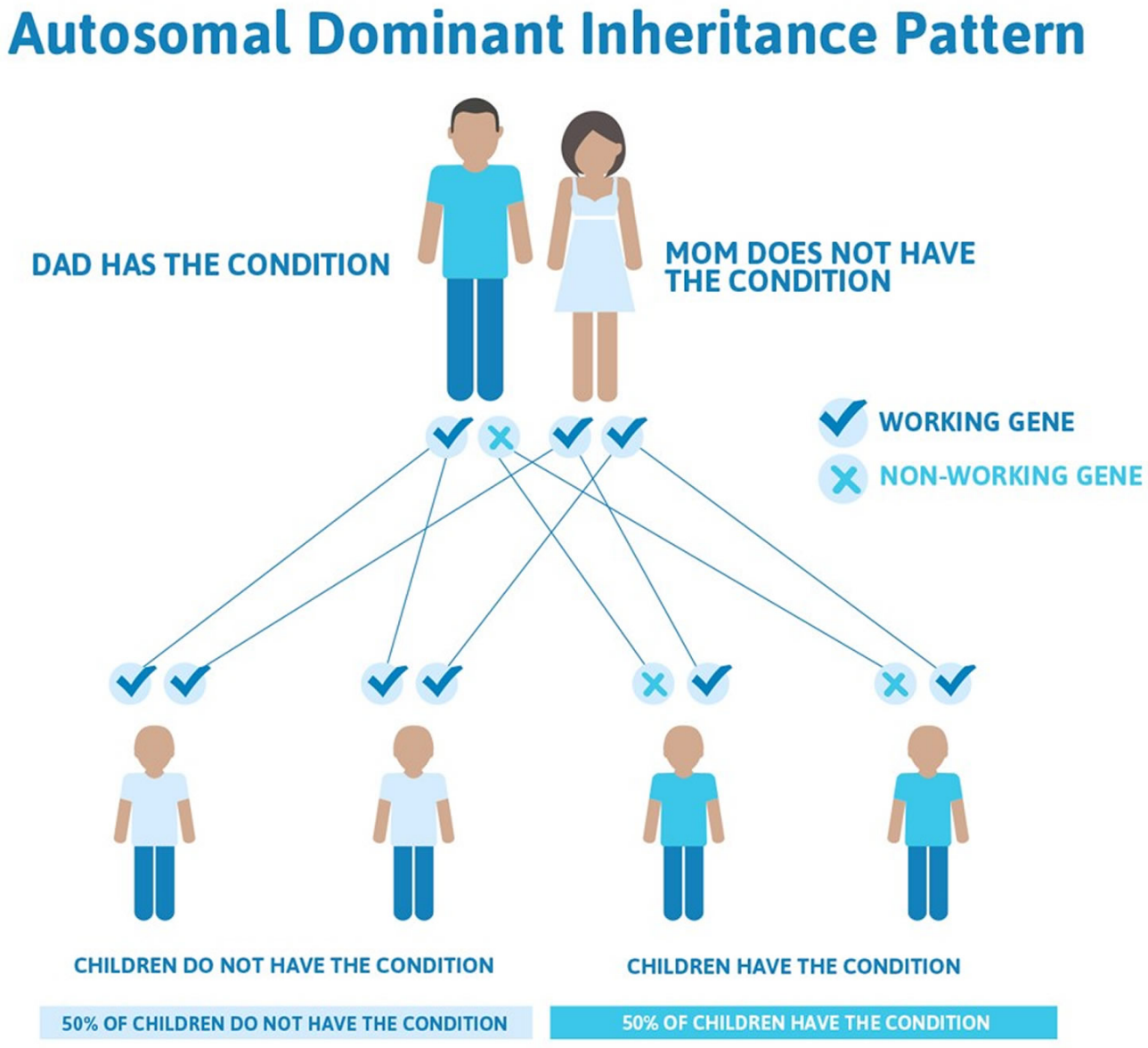
- Fisher first put these ideas together and formally introduced randomization as the "reasoned basis for inference" in 1925.
- He later revealed in the 1951 Bateson Lecture that his factorial method of experimentation derives "its structure and its name from the simultaneous inheritance of Mendelian factors".
- Most relevant quote:

  The different genotypes possible from the same mating have been beautifully randomised by the meiotic process. A more perfect control of conditions is scarcely possible, than that of different genotypes appearing in the same litter.

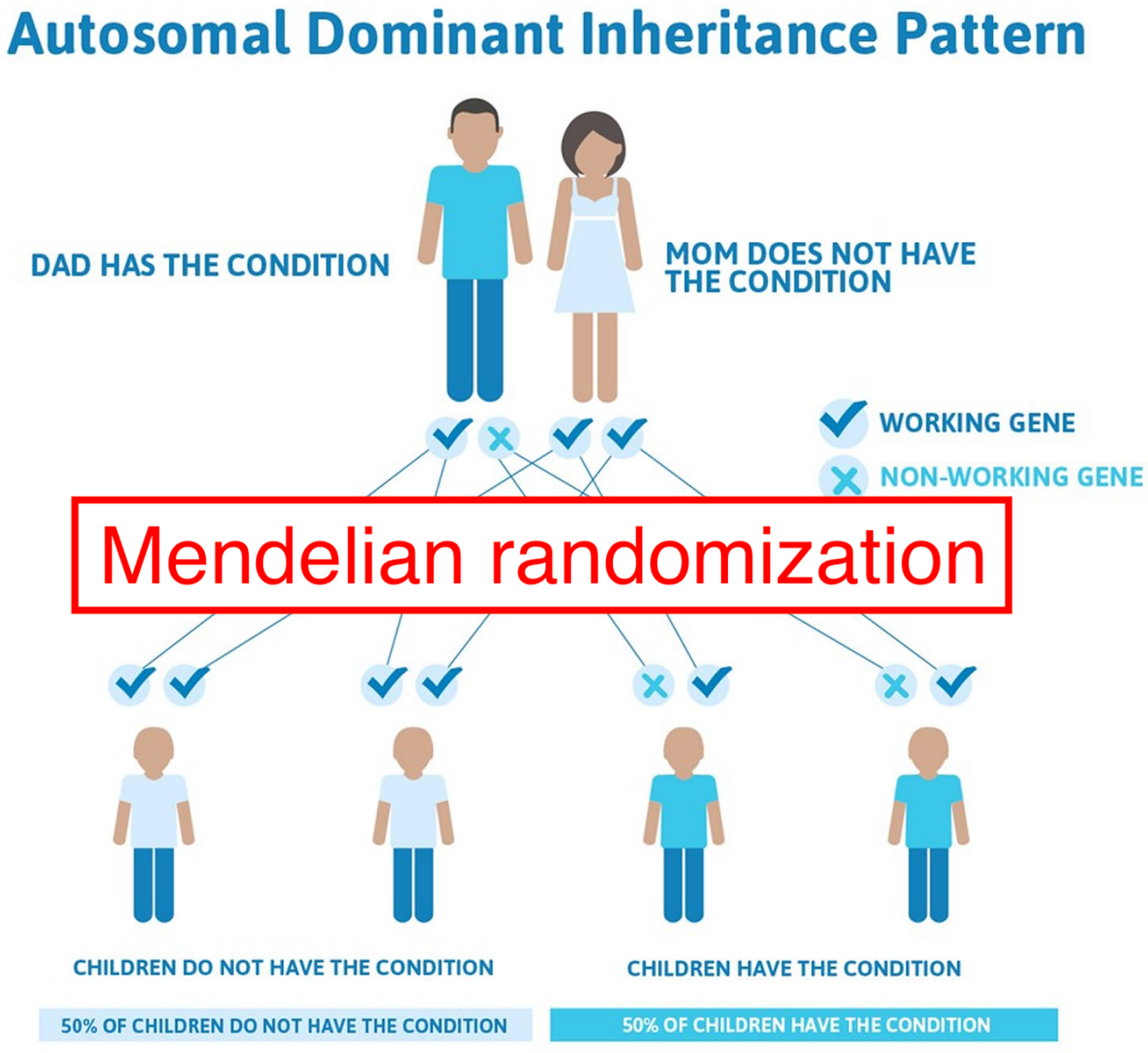
My current definition of MR

MR = Base causal inference on randomness in Mendelian inheritance.

Heredity as a natural experiment



Heredity as a natural experiment



More (not so old) ideas

- The current literature on (narrow sense) MR often attribute the idea to Katan (1986), who proposed to address reverse causation in the hypothesised effect of low serum cholesterol on cancer risk via polymorphisms in the *APOE* gene.

More (not so old) ideas

- The current literature on (narrow sense) MR often attribute the idea to Katan (1986), who proposed to address reverse causation in the hypothesised effect of low serum cholesterol on cancer risk via polymorphisms in the *APOE* gene.
- The same reasoning is applied to study the effectiveness of bone marrow transplantation in treating leukaemia by Gray and Wheatley (1991). They also coined the term "Mendelian randomization".

More (not so old) ideas

- The current literature on (narrow sense) MR often attribute the idea to Katan (1986), who proposed to address reverse causation in the hypothesised effect of low serum cholesterol on cancer risk via polymorphisms in the *APOE* gene.
- The same reasoning is applied to study the effectiveness of bone marrow transplantation in treating leukaemia by Gray and Wheatley (1991). They also coined the term "Mendelian randomization".
- MR becomes more widely known after the seminal lecture and article by Davey Smith and Ebrahim (2003).

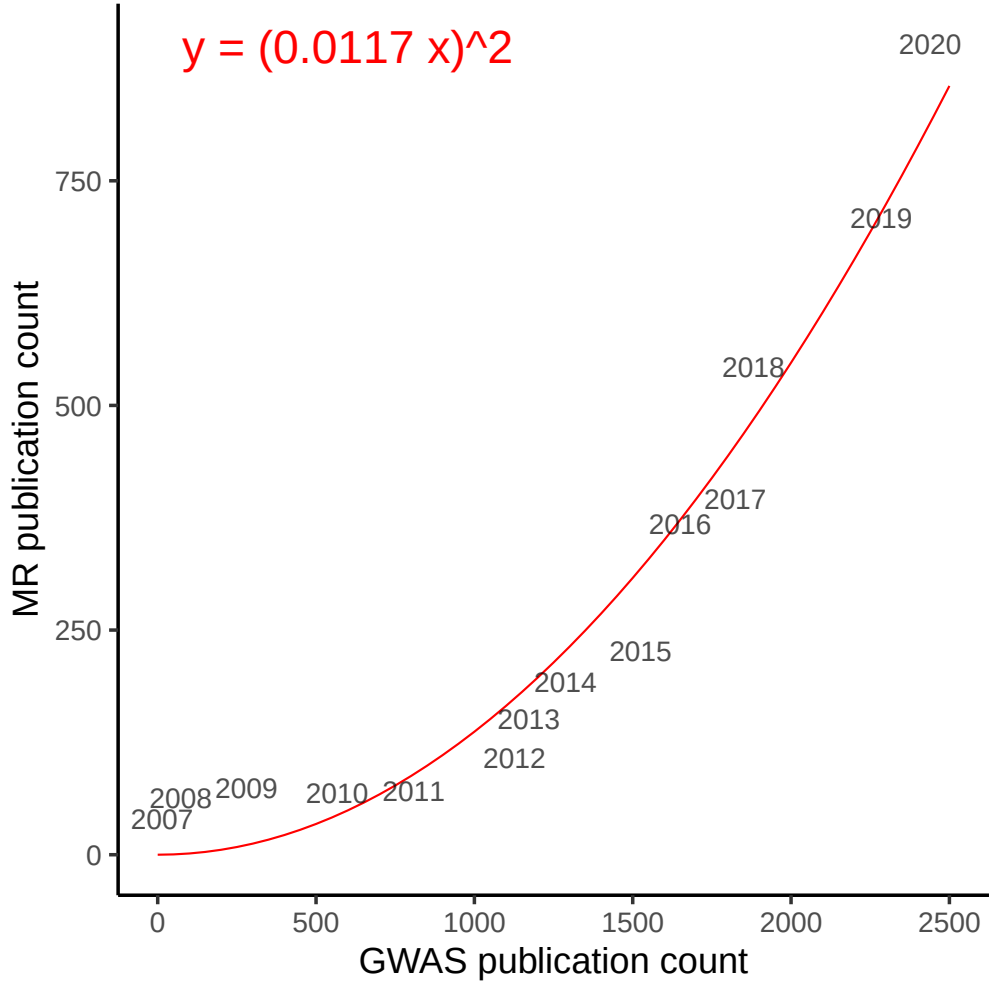
More (not so old) ideas

- The current literature on (narrow sense) MR often attribute the idea to Katan (1986), who proposed to address reverse causation in the hypothesised effect of low serum cholesterol on cancer risk via polymorphisms in the *APOE* gene.
- The same reasoning is applied to study the effectiveness of bone marrow transplantation in treating leukaemia by Gray and Wheatley (1991). They also coined the term "Mendelian randomization".
- MR becomes more widely known after the seminal lecture and article by Davey Smith and Ebrahim (2003).
- Later, it is recognized that the proposal amounts to an instrumental variable analysis (Thomas and Conti 2004; Didelez and Sheehan 2007).

More (not so old) ideas

- The current literature on (narrow sense) MR often attribute the idea to Katan (1986), who proposed to address reverse causation in the hypothesised effect of low serum cholesterol on cancer risk via polymorphisms in the *APOE* gene.
- The same reasoning is applied to study the effectiveness of bone marrow transplantation in treating leukaemia by Gray and Wheatley (1991). They also coined the term "Mendelian randomization".
- MR becomes more widely known after the seminal lecture and article by Davey Smith and Ebrahim (2003).
- Later, it is recognized that the proposal amounts to an instrumental variable analysis (Thomas and Conti 2004; Didelez and Sheehan 2007).
- A great talk by George Davey Smith on where MR came from: `https://www.youtube.com/watch?v=Ai5Vf74xVmQ`.

Surging popularity of MR



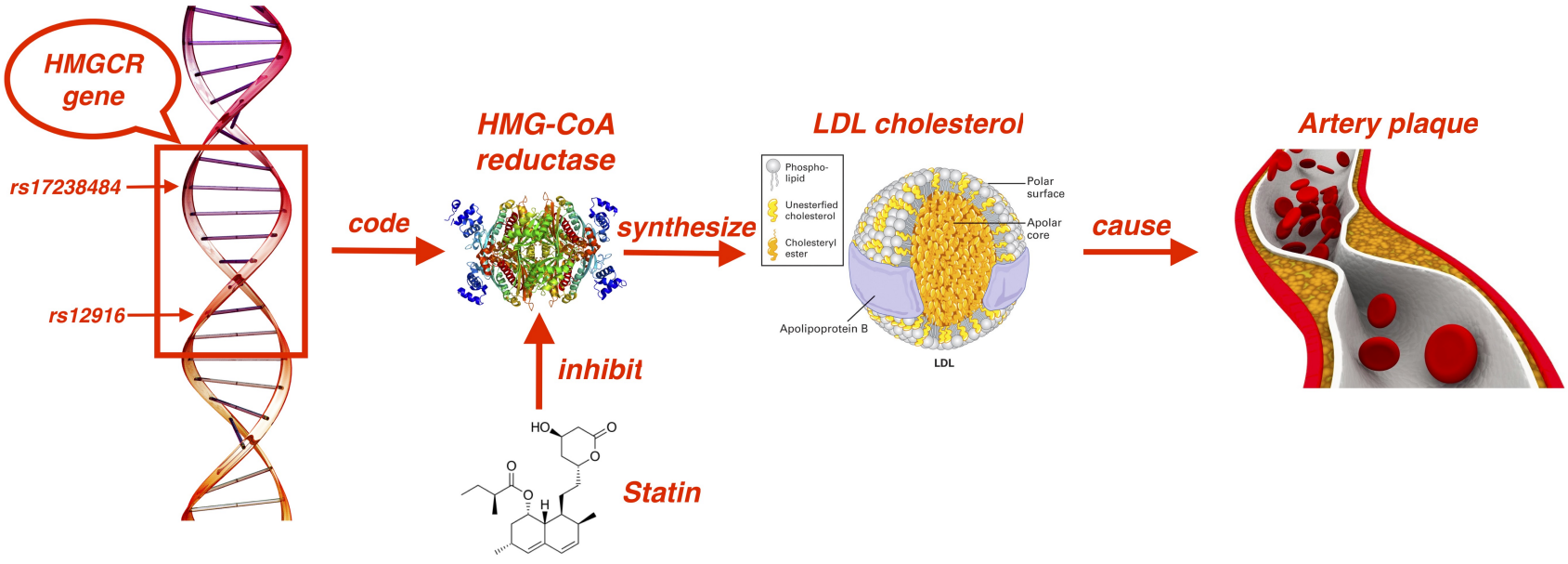
- Fueled by the availability of GWAS datasets.[1]

[1] Data are obtained from Web of Science (`https://www.webofknowledge.com/`).

Outline

Example: Causal effect of the "bad" cholesterol

A well understood pathway of heart disease

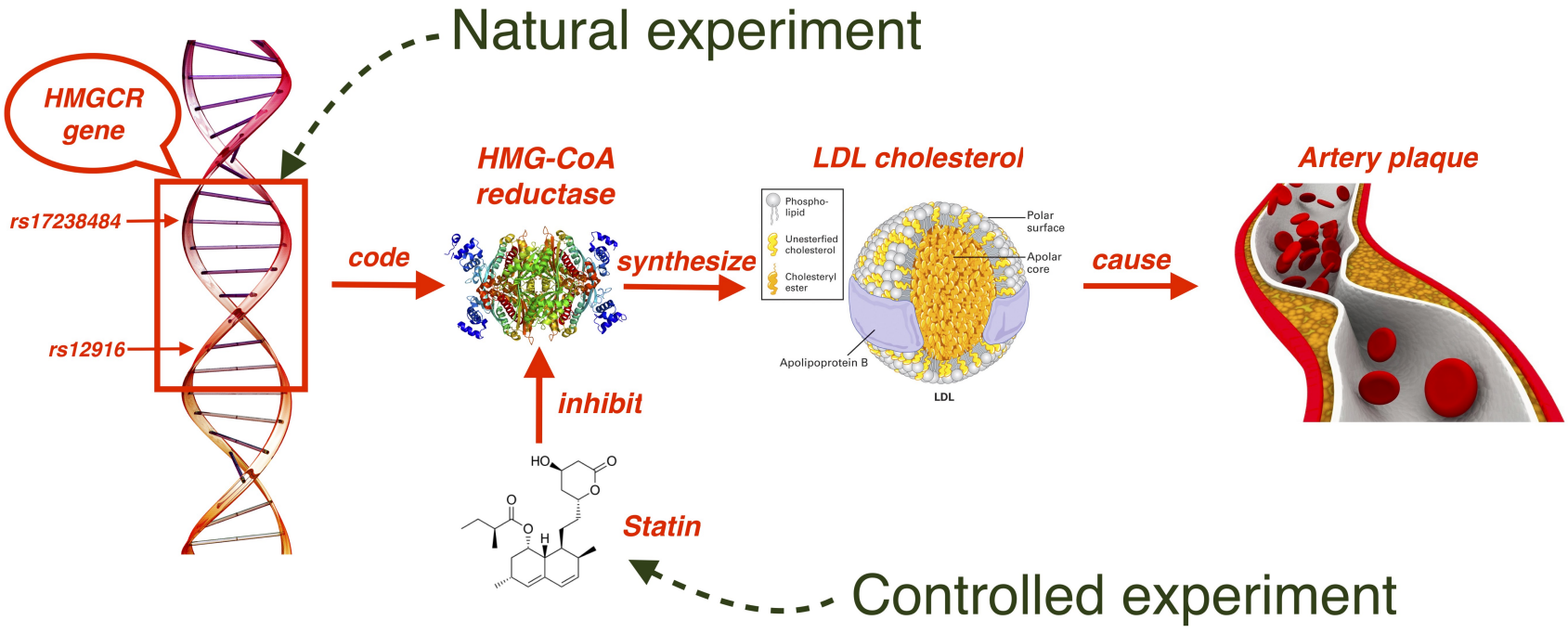


Basic idea

People who inherited certain alleles of *rs17238484* and *rs12916* have **naturally** higher concentration of LDL cholesterol.

Example: Causal effect of the "bad" cholesterol

A well understood pathway of heart disease

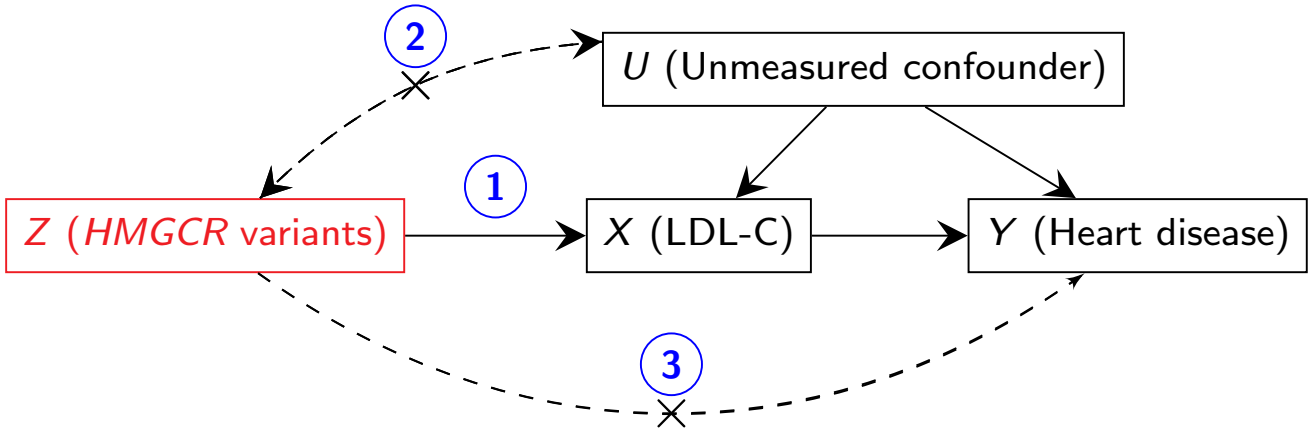


Basic idea

People who inherited certain alleles of *rs17238484* and *rs12916* have **naturally** higher concentration of LDL cholesterol.

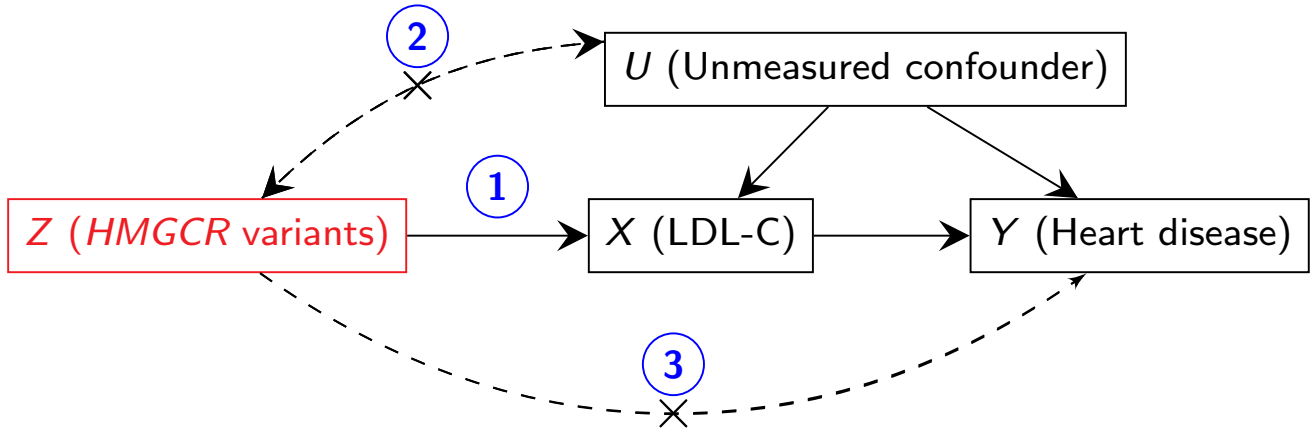
When do genetic instruments give correct answers?

The IV diagram



When do genetic instruments give correct answers?

The IV diagram

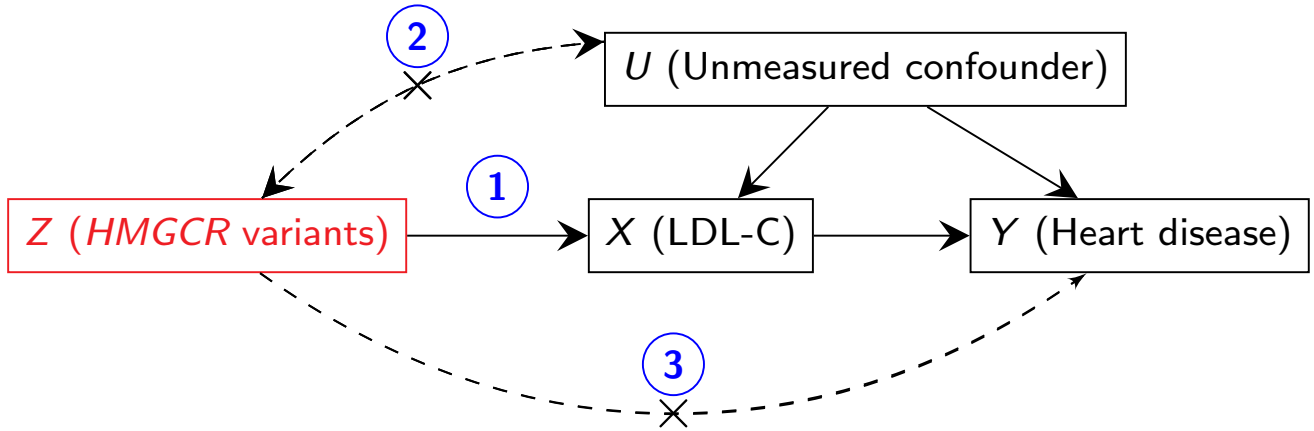


Must assume 3 core IV assumptions $\Longrightarrow$ Partial identification

1. **Relevance:** $Z \not\perp\!\!\!\perp X$.
2. **Exogeneity (natural experiment):** $Z \perp\!\!\!\perp U$.
3. **Exclusion restriction:** Z has no direct effect on Y.

When do genetic instruments give correct answers?

The IV diagram



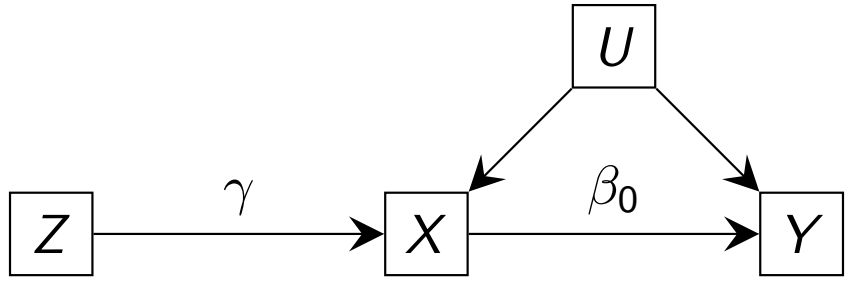
Must assume 3 core IV assumptions $\Longrightarrow$ Partial identification

1. **Relevance:** $Z \not\perp\!\!\!\perp X$.
2. **Exogeneity (natural experiment):** $Z \perp\!\!\!\perp U$.
3. **Exclusion restriction:** Z has no direct effect on Y.

Plus 1 extra assumption $\Longrightarrow$ Point identification

Could be linearity, monotonicity (Angrist, Imbens & Rubin, 1996), or homogeneity (Hernán & Robins, 2006; Wang & Tchetgen Tchetgen, 2018).

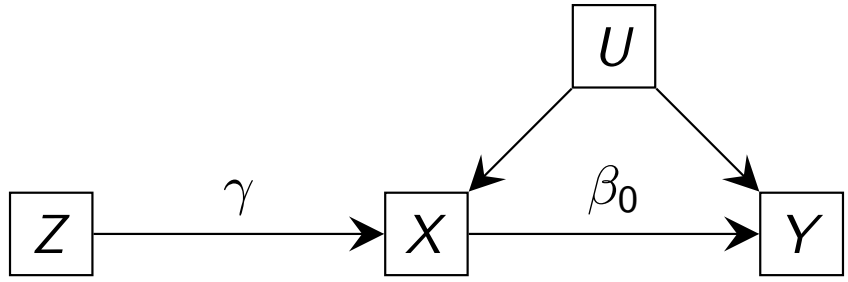
Basic idea: division



The Wald estimator

$$\text{Causal effect of } X \text{ on } Y\ (\beta_0) = \frac{\textbf{Causal effect of } \boldsymbol{Z} \textbf{ on } \boldsymbol{Y}\ (\Gamma = \gamma \cdot \beta_0)}{\textbf{Causal effect of } \boldsymbol{Z} \textbf{ on } \boldsymbol{X}\ (\gamma)}.$$

Basic idea: division



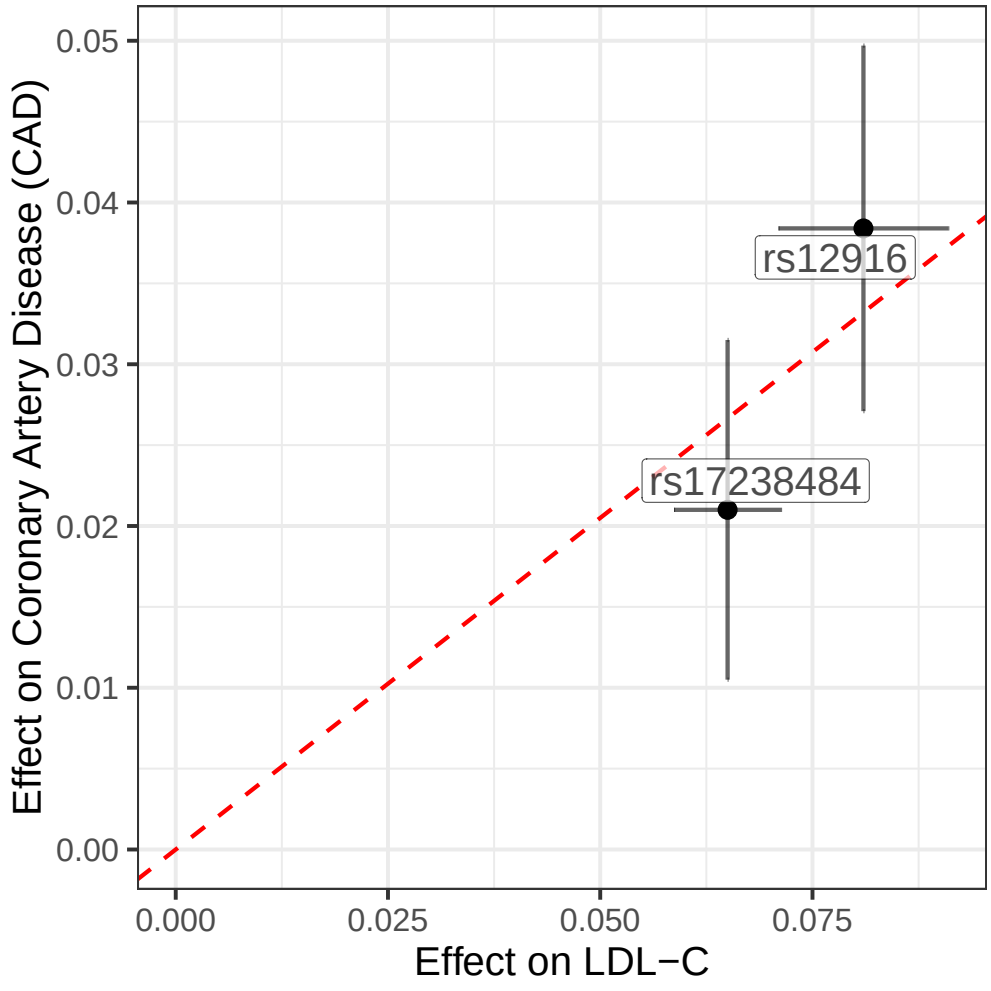
The Wald estimator

$$\text{Causal effect of } X \text{ on } Y\ (\beta_0) = \frac{\textbf{Causal effect of } \boldsymbol{Z} \textbf{ on } \boldsymbol{Y}\ (\Gamma = \gamma \cdot \beta_0)}{\textbf{Causal effect of } \boldsymbol{Z} \textbf{ on } \boldsymbol{X}\ (\gamma)}.$$

Heuristic: Linear structural equation model

$$\begin{aligned} X &= \gamma Z_j + \eta_X U + E_X, \\ Y &= \beta_0 X + \eta_Y U + E_Y \\ &= (\beta_0 \gamma) Z + \underbrace{f(U, E_X, E_Y)}_{\text{independent of } Z} \end{aligned}$$

Example: Causal effect of LDL-cholesterol



Division $\overset{\text{in statistics}}{=\!=\!=}$ Regression with no intercept.

Main challenge of Mendelian randomization

Violation of exclusion restriction due to pleiotropy (multiple functions of genes)

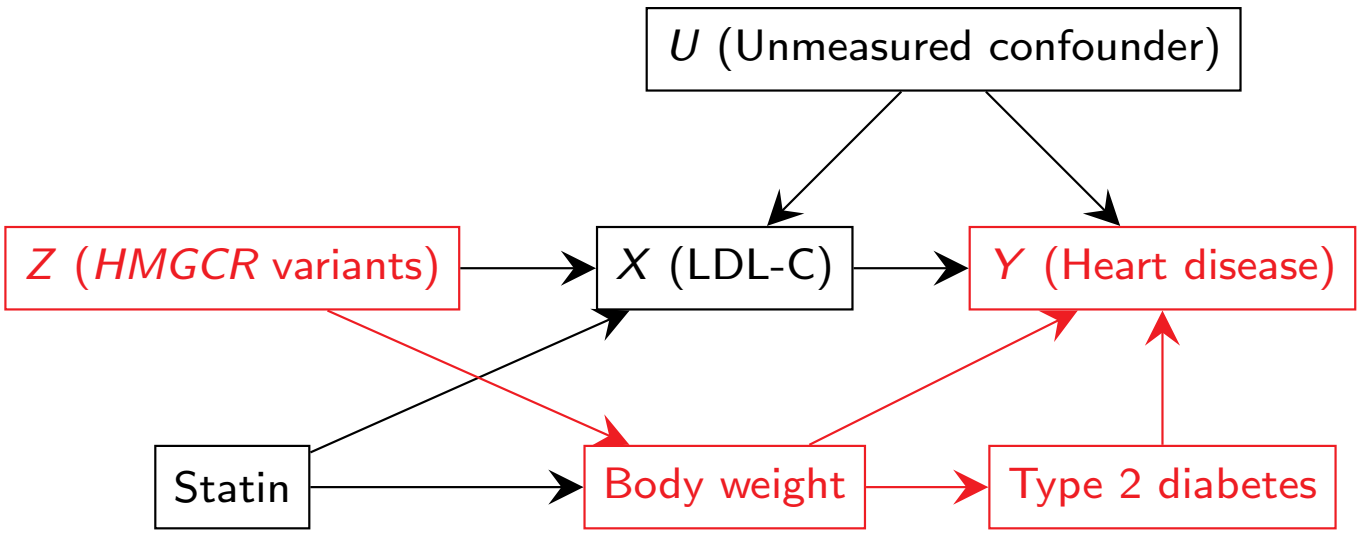
[2] Swerdlow, D. I., et al. "HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials." *Lancet* (2015).

[3] Boyle, E. et al. (2017). "An expanded view of complex traits: from polygenic to omnigenic". *Cell* 169, p1177–1186.

Main challenge of Mendelian randomization

Violation of exclusion restriction due to pleiotropy (multiple functions of genes)

Example: *HMGCR* is associated with body weight[2]



- Recent studies show that pleiotropy is indeed wide-spread.[3]

[2] Swerdlow, D. I., et al. "HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials." *Lancet* (2015).

[3] Boyle, E. et al. (2017). "An expanded view of complex traits: from polygenic to omnigenic". *Cell* 169, p1177–1186.

Two recent ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Two recent ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Today's Answer: Find many rusty-if-not-broken calipers!!

Two recent ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Today's Answer: Find many rusty-if-not-broken calipers!!

Question 2: When is that enough?

Two recent ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Today's Answer: Find many rusty-if-not-broken calipers!!

Question 2: When is that enough?

1. $< 50\%$ of the calipers are broken (Kang et al., 2016); or
2. Rusty readings are balanced around the truth (Bowden et al., 2015).

Two recent ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Today's Answer: Find many rusty-if-not-broken calipers!!

Question 2: When is that enough?

1. $< 50\%$ of the calipers are broken (Kang et al., 2016); or
2. Rusty readings are balanced around the truth (Bowden et al., 2015).

Remaining issues

1. Both situations are common in MR.
2. Need to deal with many weak instruments.

Outline

3-sample summary-data MR

Instrumental variables $\boldsymbol{Z}_{1:p}$: Independent SNPs.

Exposure variable X: Body mass index (BMI).

Outcome variable Y: Systolic blood pressure (SBP).

Data preprocessing (non-overlapping 3 GWAS)

Name	Selection GWAS	Exposure GWAS	Outcome GWAS
Dataset	`BMI-FEM`	`BMI-MAL`	`SBP-UKBB`
Source	GIANT (female)	GIANT (male)	UK BioBank
Sample size	171977	152893	317754
GWAS	$\mathrm{lm}(X \sim Z_j)$	$\mathrm{lm}(X \sim Z_j)$	$\mathrm{lm}(Y \sim Z_j)$
Coefficient	**Used for selection**	$\hat{\gamma}_j$	$\hat{\Gamma}_j$
Std. Err.		σ_{Xj}	σ_{Yj}

Step 1 Use `BMI-FEM` to **select significant and independent SNPs** (p-value $\leq p_{\mathrm{sel}} = 5 \times 10^{-8}$, $p = 25$).

Step 2 Use `BMI-MAL` to obtain $(\hat{\gamma}_j, \sigma_{Xj})_{j=1}^{p}$.

Step 3 Use `SBP-UKBB` to obtain $(\hat{\Gamma}_j, \sigma_{Yj})_{j=1}^{p}$.

3-sample summary-data Mendelian randomization

Instrumental variables $\boldsymbol{Z}_{1:p}$: SNPs.

Exposure variable X: Body mass index (BMI).

Outcome variable Y: Systolic blood pressure (SBP).

Data preprocessing (non-overlapping 3 GWAS)

Name	Selection GWAS	Exposure GWAS	Outcome GWAS
Dataset	`BMI-FEM`	`BMI-MAL`	`SBP-UKBB`
Source	GIANT (female)	GIANT (male)	UK BioBank
Sample size	171977	152893	317754
GWAS	$\mathrm{lm}(X \sim Z_j)$	$\mathrm{lm}(X \sim Z_j)$	$\mathrm{lm}(Y \sim Z_j)$
Coefficient	Used for selection	$\hat{\gamma}_j$	$\hat{\Gamma}_j$
Std. Err.		$\boldsymbol{\sigma_{Xj} = 1}$	$\boldsymbol{\sigma_{Yj} = 1}$

Step 1 Use `BMI-FEM` to select significant and independent SNPs (p-value $\leq p_{\mathrm{sel}} = 5 \times 10^{-8}$, $p = 25$).

Step 2 Use `BMI-MAL` to obtain $(\hat{\gamma}_j, \boldsymbol{\sigma_{Xj} = 1})_{j=1}^{p}$.

Step 3 Use `SBP-UKBB` to obtain $(\hat{\Gamma}_j, \boldsymbol{\sigma_{Yj} = 1})_{j=1}^{p}$.

Assumption 1

Measurement error model

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\Gamma} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \gamma \\ \Gamma \end{pmatrix}, \; \boldsymbol{I}_{2p} \right).$$

Pre-processing warrants Assumption 1

Name	Selection GWAS	Exposure GWAS	Outcome GWAS
GWAS	$\mathrm{Im}(X \sim Z_j)$	$\mathrm{Im}(X \sim Z_j)$	$\mathrm{Im}(Y \sim Z_j)$
Coefficient	Used for selection	$\hat{\gamma}_j$	$\hat{\Gamma}_j$
Std. Err.		$\boldsymbol{\sigma_{Xj} = 1}$	$\boldsymbol{\sigma_{Yj} = 1}$

- Large sample size $\Rightarrow$ CLT.
- (Approximate) Independence due to
  1. Non-overlapping samples (in all three GWAS).
  2. Independent SNPs.

Assumption 2

Linking the genetic associations

The causal effect β_0 satisfies $\boldsymbol{\Gamma \approx \beta_0 \gamma}$. This contains two claims:

1. The relationship is **approximately linear**.
2. The slope β_0 has a **causal interpretation**.

We will consider 3 versions of Assumption 2 below.

Assumptions 1 & 2.1 $\Longrightarrow$ Profile score (PS)

Assumption 2.1 (All accurate calipers)

The linear relation $\Gamma_j = \beta_0 \gamma_j$ is true for every j.

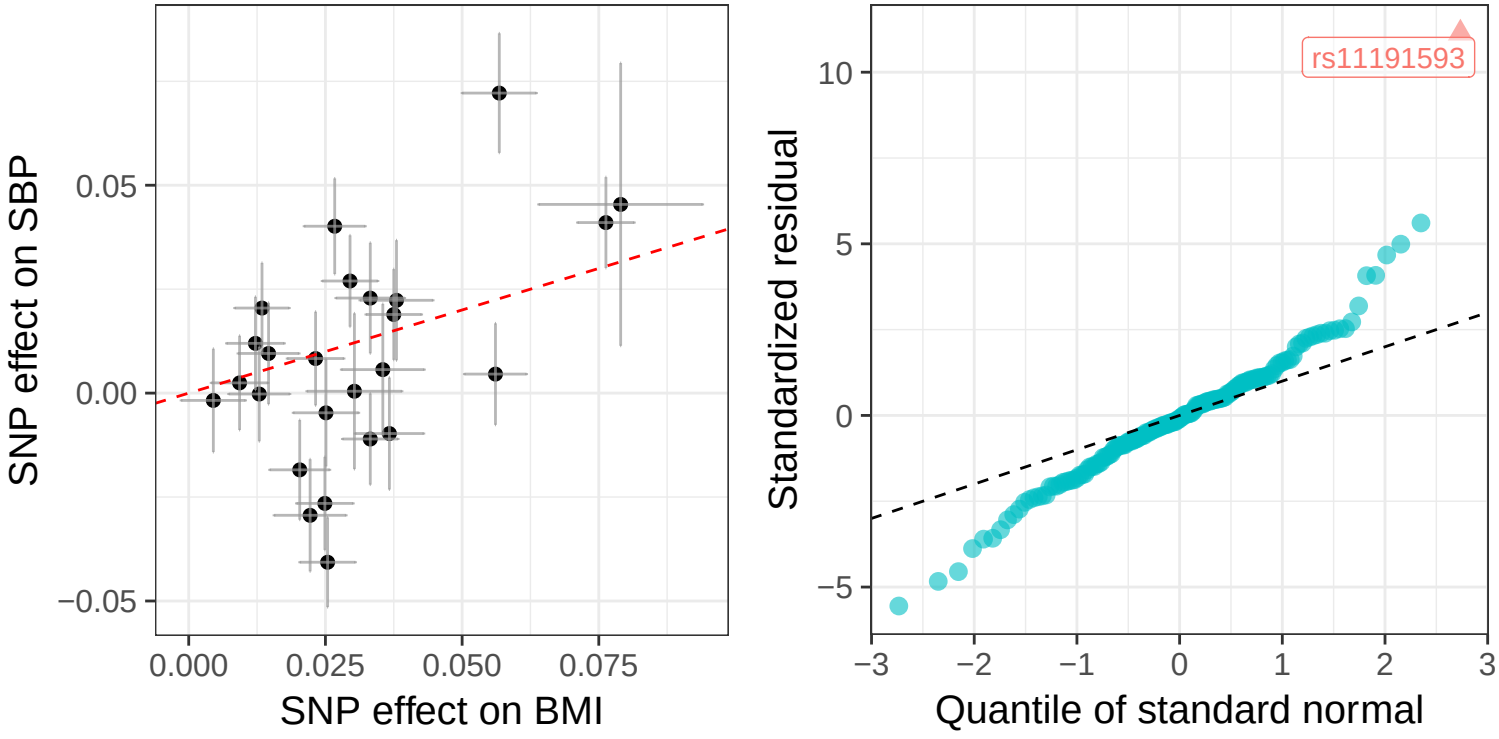
- Log-likelihood of the data:

$$l(\beta, \gamma_1, \ldots, \gamma_p) = -\frac{1}{2}\Big[\sum_{j=1}^{p}(\hat{\gamma}_j - \gamma_j)^2 + \sum_{j=1}^{p}(\hat{\Gamma}_j - \gamma_j\beta)^2\Big].$$

- **Profile likelihood:** $\boldsymbol{l(\beta) = \max_{\gamma} l(\beta, \gamma) = -\frac{1}{2}\sum_{j=1}^{p}\frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{1+\beta^2}}$.
- This extends the **limited information maximum likelihood (LIML)** (Anderson & Rubin, 1949) to the **two-sample summary-data setting**.
- Can prove consistency and asymptotic normality when $\|\gamma\|^2 \to \infty$ (instruments are collectively strong).

Diagnostic plots show clear overdispersion

BMI-SBP Example (continued)



- Left ($p = 25$, $p_{\text{sel}} < 5 \cdot 10^{-8}$): Scatter-plot of GWAS summary data.
- Right ($p = 160$, $p_{\text{sel}} < 10^{-4}$): Q-Q plot of standardized residual:

$$t_j(\hat{\beta}) = \frac{\hat{\Gamma}_j - \hat{\beta}\hat{\gamma}_j}{\sqrt{1 + \hat{\beta}^2}}.$$

Why did Assumption 2.1 fail? $\Longrightarrow$ Assumption 2.2

Heuristic: Linear structural equation model (with invalid IVs)

$$X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X U + E_X,$$

$$Y = \beta_0 X + \sum_{j=1}^{p} \alpha_j Z_j + \eta_Y U + E_Y$$

$$= \sum_{j=1}^{p} \underbrace{(\beta_0 \gamma_j + \alpha_j)}_{\Gamma_j} Z_j \quad + \quad \underbrace{f(U, E_X, E_Y)}_{\text{independent of } Z}$$

Why did Assumption 2.1 fail? $\Longrightarrow$ Assumption 2.2

Heuristic: Linear structural equation model (with invalid IVs)

$$X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X U + E_X,$$

$$Y = \beta_0 X + \sum_{j=1}^{p} \alpha_j Z_j + \eta_Y U + E_Y$$

$$= \sum_{j=1}^{p} \underbrace{(\beta_0 \gamma_j + \alpha_j)}_{\Gamma_j} Z_j \quad + \quad \underbrace{f(U, E_X, E_Y)}_{\text{independent of } Z}$$

Assumption 2.2 (Random rusty calipers)

Assume $\alpha_j = \Gamma_j - \beta_0 \gamma_j$ **is independent of** γ_j and $\alpha_j \overset{i.i.d.}{\sim} \mathrm{N}(0, \tau_0^2)$.

- Independence is crucial but non-verifiable.
- First occurred in Bowden et al. (2015) with a neat acronym—InSIDE (Instrument Strength Independent of Direct Effect).

A Neyman-Scott problem

MLE is not consistent under Assumption 2.2

- The profile likelihood under Assumption 2.2 is given by

$$I(\beta, \tau^2) = -\frac{1}{2}\sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{1 + \beta^2 + \tau^2} + \log(1+\tau^2),$$

- Easy to verify

$$\mathbb{E}\Big[\frac{\partial}{\partial\beta} I(\beta_0, \tau_0^2)\Big] = 0.$$

- But **the other score function is biased**:

$$\frac{\partial}{\partial\tau^2} I(\beta, \tau^2) = \frac{1}{2}\sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{\left(1 + \beta^2 + \tau^2\right)^2} - \frac{1}{1+\tau^2}.$$

Assumptions 1 & 2.2 $\Longrightarrow$ Adjusted profile score (APS)

- We take the approach of McCullagh & Tibshirani (1990) to adjust the profile score

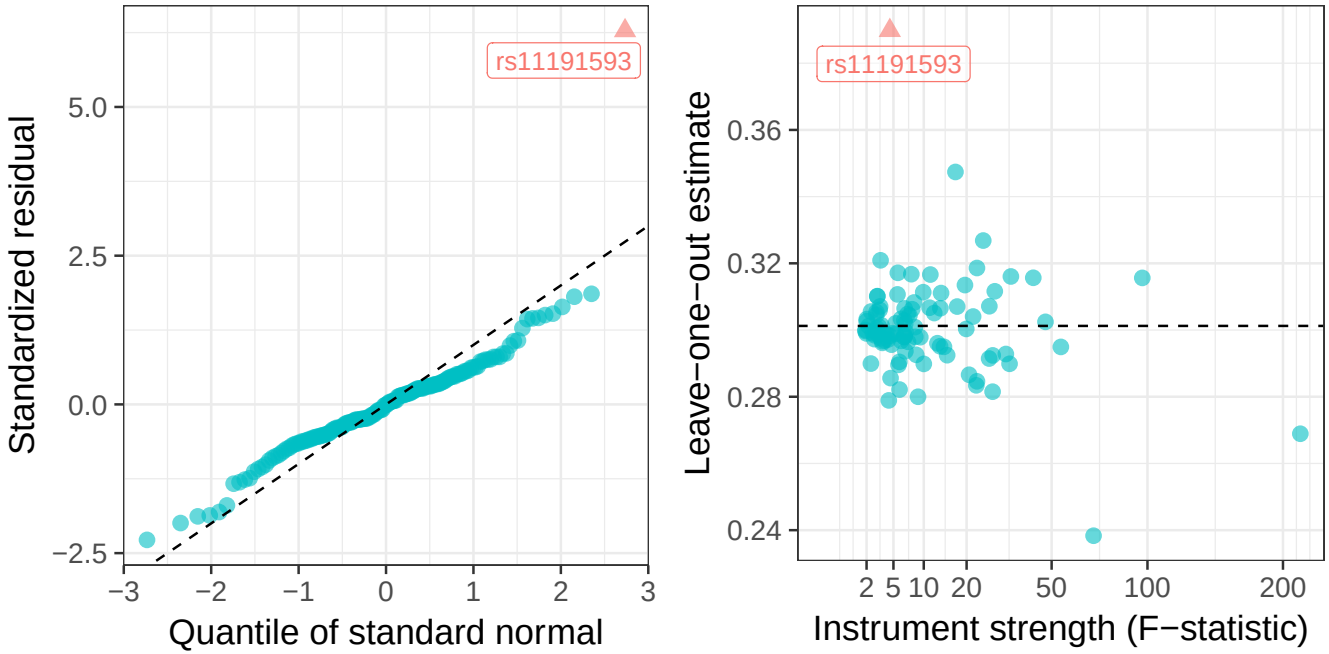
$$\psi_1(\beta, \tau^2) = -\frac{\partial}{\partial \beta} l(\beta, \tau^2),$$

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

- Under reasonable assumptions, can show any nontrivial (finite) solution is consistent and asymptotic normal.

Diagnostic plots show influential outlier

- Same 160 SNPs ($p_{\text{sel}} < 10^{-4}$).



Left: Q-Q plot of std. residuals; Right: Influence of a single SNP.

- **A clear outlier: rs11191593, with high influence.**
- A GWAS catalog search: *rs11191593* is strongly associated with immature red blood cell count.[4]
- Slightly underdispersed (probably because β is underestimated).

[4] Astle, W. et al. (2016). "The allelic landscape of human blood cell trait variation and links to common complex disease." Cell 167: 1415-1429.

Assumptions 1 & 2.3 $\Longrightarrow$ RAPS

Assumption 2.3 (Random rusty calipers **& a few broken**)

Most $\alpha_j \sim \mathrm{N}(0, \tau_0^2)$, but **a small number of $|\alpha_j|$ might be very large**.

Assumptions 1 & 2.3 $\Longrightarrow$ RAPS

Assumption 2.3 (Random rusty calipers **& a few broken**)

Most $\alpha_j \sim \mathrm{N}(0, \tau_0^2)$, but **a small number of $|\alpha_j|$ might be very large**.

Robust adjusted profile score (RAPS)

- Define standardized residual: $t_j(\beta, \tau^2) = \dfrac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{1 + \beta^2 + \tau^2}}$.
- For some **robust loss ρ (let $\psi = \rho'$)**, the RAPS equations are

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

$$\psi_2^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \frac{t_j \cdot \psi(t_j) - \mathbb{E}[T\psi(T)]}{1 + \beta^2 + \tau^2}, \text{ for } T \sim \mathrm{N}(0, 1).$$

Assumptions 1 & 2.3 $\Longrightarrow$ RAPS

Assumption 2.3 (Random rusty calipers **& a few broken**)

Most $\alpha_j \sim \mathrm{N}(0, \tau_0^2)$, but **a small number of $|\alpha_j|$ might be very large**.

Robust adjusted profile score (RAPS)

- Define standardized residual: $t_j(\beta, \tau^2) = \dfrac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{1 + \beta^2 + \tau^2}}$.
- For some **robust loss ρ (let $\psi = \rho'$)**, the RAPS equations are

$$\psi_1^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \boldsymbol{\psi(t_j)} \cdot \frac{\partial}{\partial\beta} t_j,$$

$$\psi_2^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \frac{\boldsymbol{t_j \cdot \psi(t_j)} - \mathbb{E}[T\psi(T)]}{1 + \beta^2 + \tau^2}, \text{ for } T \sim \mathrm{N}(0, 1).$$

- Reduces to APS when $\rho(t) = t^2/2$ so $\psi(t) = t$.

Assumptions 1 & 2.3 $\Longrightarrow$ RAPS

Assumption 2.3 (Random rusty calipers **& a few broken**)

Most $\alpha_j \sim \mathrm{N}(0, \tau_0^2)$, but **a small number of $|\alpha_j|$ might be very large**.

Robust adjusted profile score (RAPS)

- Define standardized residual: $t_j(\beta, \tau^2) = \dfrac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{1 + \beta^2 + \tau^2}}$.
- For some **robust loss ρ (let $\psi = \rho'$)**, the RAPS equations are

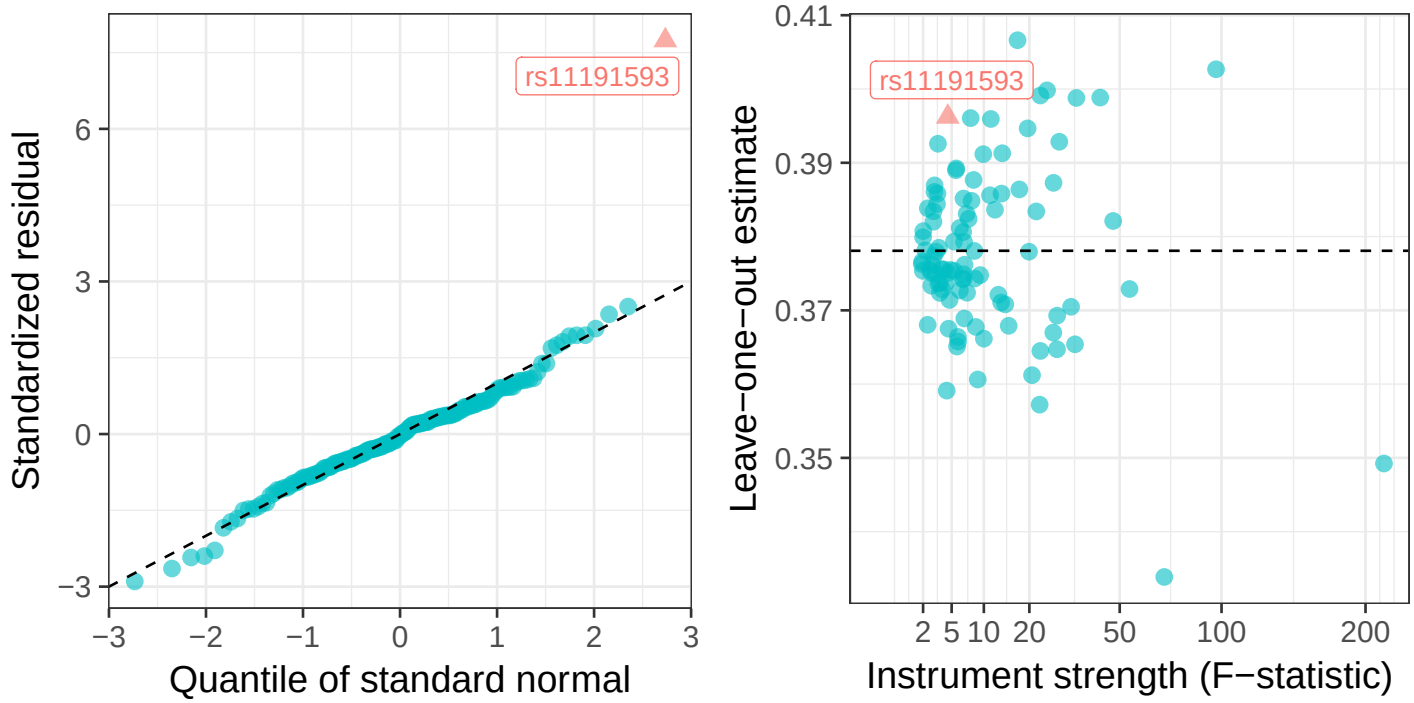
$$\psi_1^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \psi(t_j) \cdot \frac{\partial}{\partial \beta} t_j,$$

$$\psi_2^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \frac{t_j \cdot \psi(t_j) - \mathbb{E}[T\psi(T)]}{1 + \beta^2 + \tau^2}, \text{ for } T \sim \mathrm{N}(0, 1).$$

- Reduces to APS when $\rho(t) = t^2/2$ so $\psi(t) = t$.
- Can establish local identifiability and asymptotic normality.

Diagnostic plots show satisfactory fit

- Same 160 SNPs, now using RAPS with Huber's loss function.



- Influence of the **outlier** `rs11191593` is limited.

More details about MR.RAPS can be found in our paper.[5]

[5] Zhao, Q. et al. (2020). "Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score." *Annals of Statistics*, 48(3):1742-1769.

Outline

Extensions

- Improve statistical efficiency with many weak instruments.[6]
  - Idea due to Lindsay (1985): Solve the following equation

$$\sum_{j=1}^{p} \begin{pmatrix} \textbf{Estimated quality} \\ \textbf{of instrument } j \end{pmatrix} \cdot \begin{pmatrix} \textbf{Estimated error} \\ \textbf{of instrument } j \end{pmatrix} = 0.$$

  - Quality of instrument is estimated by empirical Bayes.

---

[6] Zhao, Q. et al. (2019). "Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization". *International Journal of Epidemiology*, 48(5):1478-1492.

[7] Wang, J. et al. (2020). "Causal Inference for Heritable Phenotypic Risk Factors Using Heterogeneous Genetic Instruments." bioRxiv:2020.05.06.077982.

[8] Iong, D. et al. (2020). "A Latent Mixture Model for Heterogeneous Causal Mechanisms in Mendelian Randomization." arXiv:2007.06476.

Extensions

- Improve statistical efficiency with many weak instruments.[6]
  - Idea due to Lindsay (1985): Solve the following equation

$$\sum_{j=1}^{p} \begin{pmatrix} \textbf{Estimated quality} \\ \textbf{of instrument } j \end{pmatrix} \cdot \begin{pmatrix} \textbf{Estimated error} \\ \textbf{of instrument } j \end{pmatrix} = 0.$$

  - Quality of instrument is estimated by empirical Bayes.
- Deal with multiple exposures, overlapping samples, determining causal direction.[7]

---

[6] Zhao, Q. et al. (2019). "Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization". *International Journal of Epidemiology*, 48(5):1478-1492.

[7] Wang, J. et al. (2020). "Causal Inference for Heritable Phenotypic Risk Factors Using Heterogeneous Genetic Instruments." bioRxiv:2020.05.06.077982.

[8] Iong, D. et al. (2020). "A Latent Mixture Model for Heterogeneous Causal Mechanisms in Mendelian Randomization." arXiv:2007.06476.

Extensions

- Improve statistical efficiency with many weak instruments.[6]
  - Idea due to Lindsay (1985): Solve the following equation

$$\sum_{j=1}^{p} \begin{pmatrix} \textbf{Estimated quality} \\ \textbf{of instrument } \boldsymbol{j} \end{pmatrix} \cdot \begin{pmatrix} \textbf{Estimated error} \\ \textbf{of instrument } \boldsymbol{j} \end{pmatrix} = 0.$$

  - Quality of instrument is estimated by empirical Bayes.
- Deal with multiple exposures, overlapping samples, determining causal direction.[7]
- Discover mechanistic heterogeneity.[8]
  - Idea: Instruments can be **clustered based on** $\beta_j = \Gamma_j / \gamma_j$. Each cluster corresponds to a distinct biological pathway.

---

[6] Zhao, Q. et al. (2019). "Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization". *International Journal of Epidemiology*, 48(5):1478-1492.

[7] Wang, J. et al. (2020). "Causal Inference for Heritable Phenotypic Risk Factors Using Heterogeneous Genetic Instruments." bioRxiv:2020.05.06.077982.

[8] Iong, D. et al. (2020). "A Latent Mixture Model for Heterogeneous Causal Mechanisms in Mendelian Randomization." arXiv:2007.06476.

Extensions

- Improve statistical efficiency with many weak instruments.[6]
  - Idea due to Lindsay (1985): Solve the following equation

$$\sum_{j=1}^{p} \begin{pmatrix} \textbf{Estimated quality} \\ \textbf{of instrument } \boldsymbol{j} \end{pmatrix} \cdot \begin{pmatrix} \textbf{Estimated error} \\ \textbf{of instrument } \boldsymbol{j} \end{pmatrix} = 0.$$

  - Quality of instrument is estimated by empirical Bayes.
- Deal with multiple exposures, overlapping samples, determining causal direction.[7]
- Discover mechanistic heterogeneity.[8]
  - Idea: Instruments can be **clustered based on** $\beta_j = \Gamma_j/\gamma_j$. Each cluster corresponds to a distinct biological pathway.
- More information:
  `http://www.statslab.cam.ac.uk/~qz280/project/iv-mr/`.

---

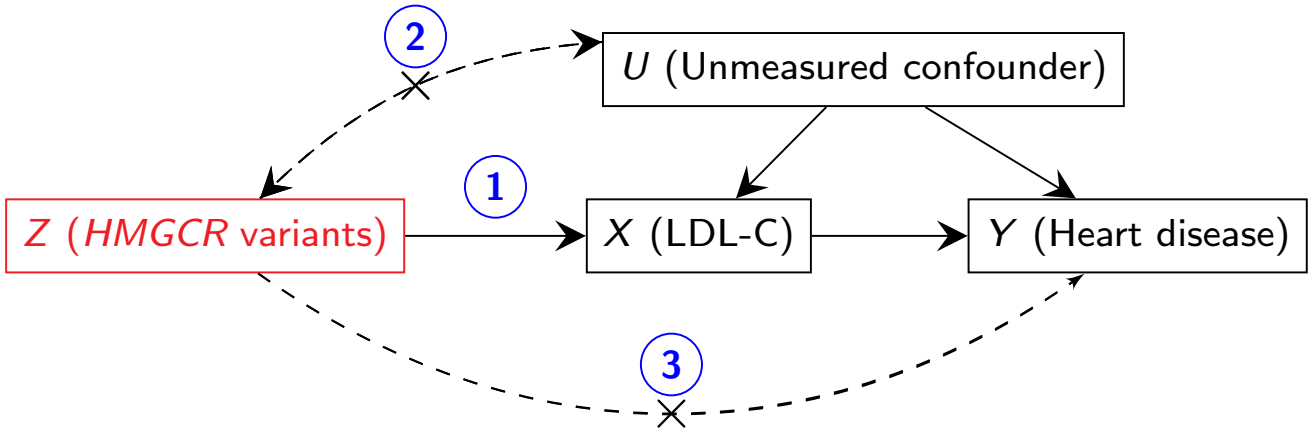
[6] Zhao, Q. et al. (2019). "Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization". *International Journal of Epidemiology*, 48(5):1478-1492.

[7] Wang, J. et al. (2020). "Causal Inference for Heritable Phenotypic Risk Factors Using Heterogeneous Genetic Instruments." bioRxiv:2020.05.06.077982.

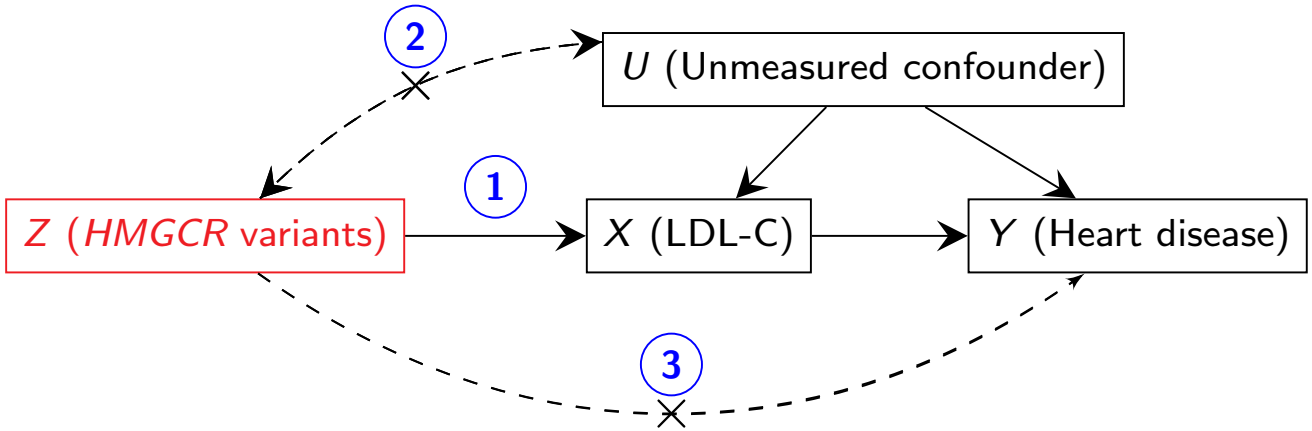
[8] Iong, D. et al. (2020). "A Latent Mixture Model for Heterogeneous Causal Mechanisms in Mendelian Randomization." arXiv:2007.06476.

Outline

Are genes trully randomized?



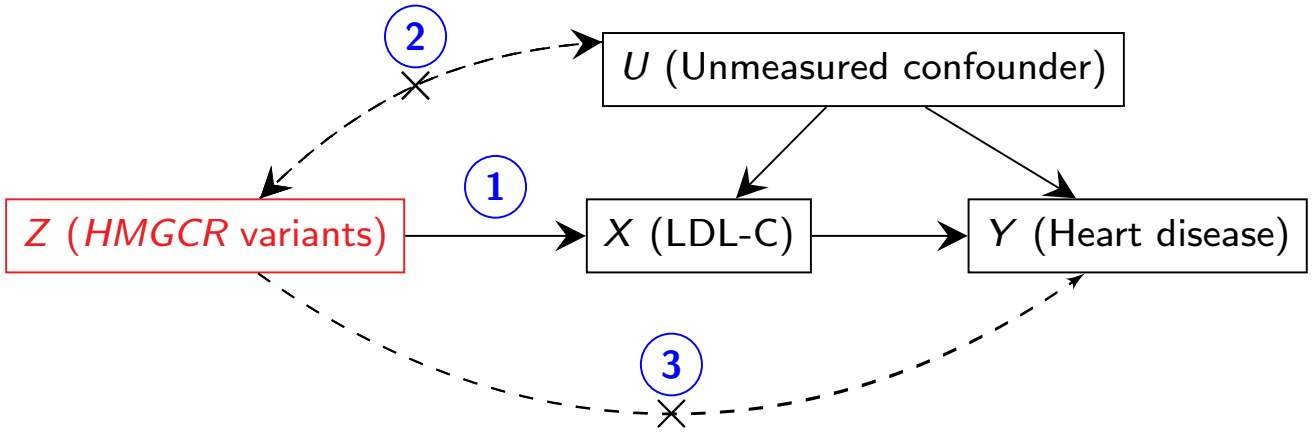
Are genes trully randomized?



Recall the core IV assumptions

1. **Relevance:** $Z \not\perp\!\!\!\perp X$.
2. **Exogeneity (natural experiment):** $Z \perp\!\!\!\perp U$.
3. **Exclusion restriction:** Z has no direct effect on Y.

Are genes trully randomized?

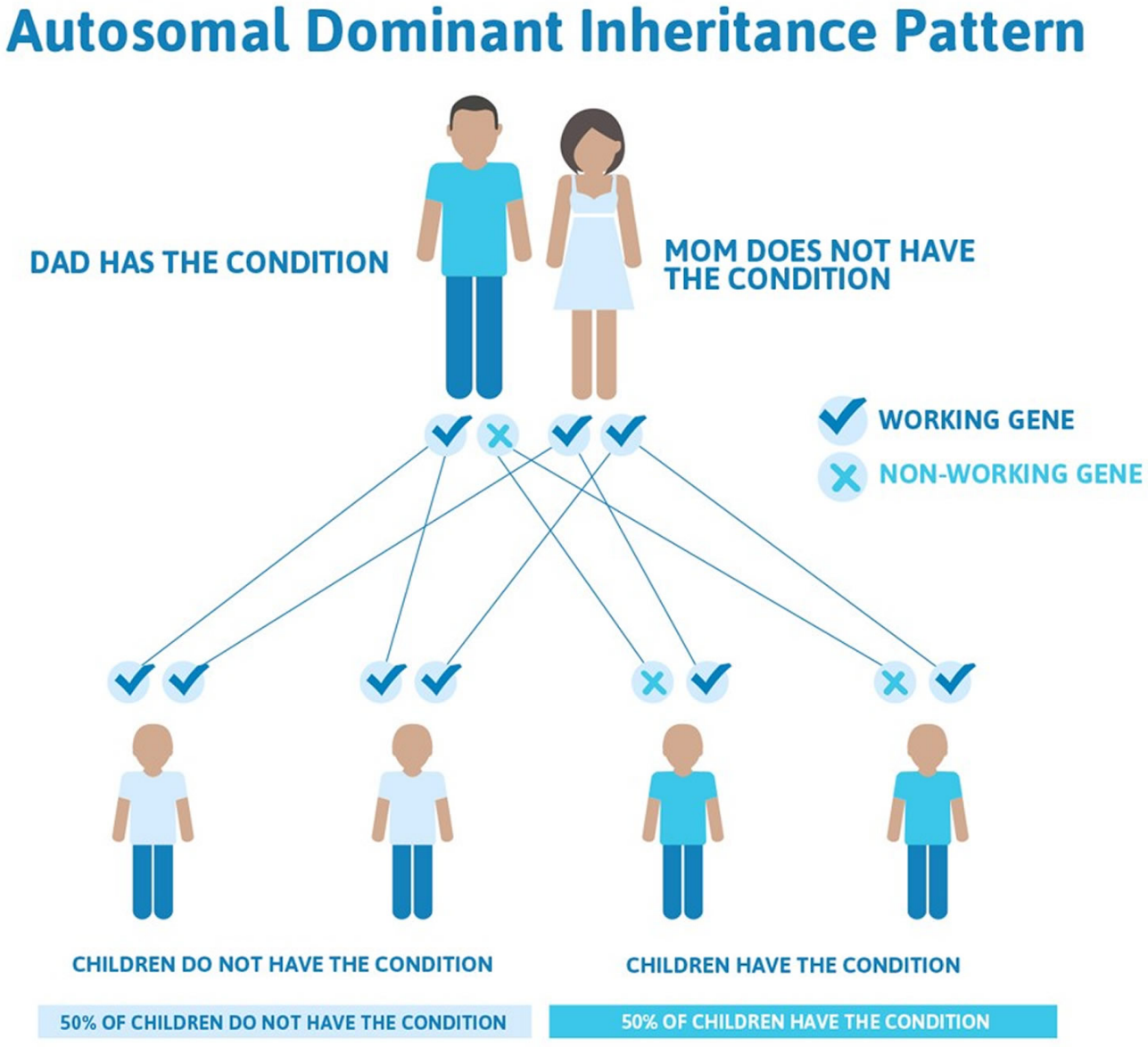


Recall the core IV assumptions

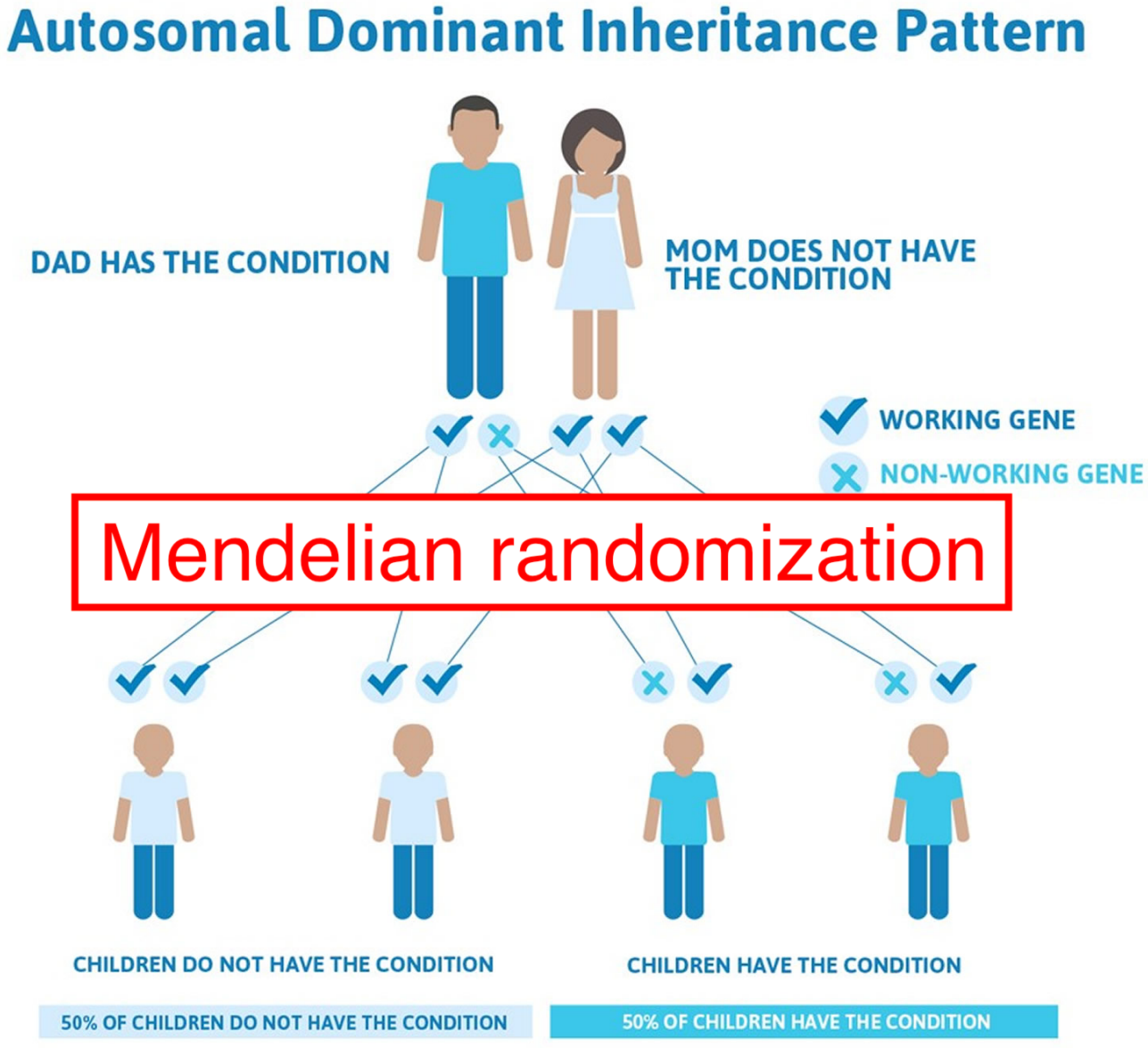
1. **Relevance:** $Z \not\perp\!\!\!\perp X$.
2. **Exogeneity (natural experiment):** $Z \perp\!\!\!\perp U$.
3. **Exclusion restriction:** Z has no direct effect on Y.

Genes are Mendelian randomized, but GWAS sampling is not!

Recall: Heredity as a natural experiment



Recall: Heredity as a natural experiment



Within-family MR

- Davey Smith and Ebrahim (2003): MR is best justified in parent-offspring design.
- But this has not been widely used due to lack of data.

Within-family MR

- Davey Smith and Ebrahim (2003): MR is best justified in parent-offspring design.
- But this has not been widely used due to lack of data.

Model-based

- Brumpton et al. (2020): dynastic effect, assortative mating, and population stratification can strongly bias MR.
- Add family fixed effects to the linear structural equations.

Within-family MR

- Davey Smith and Ebrahim (2003): MR is best justified in parent-offspring design.
- But this has not been widely used due to lack of data.

Model-based

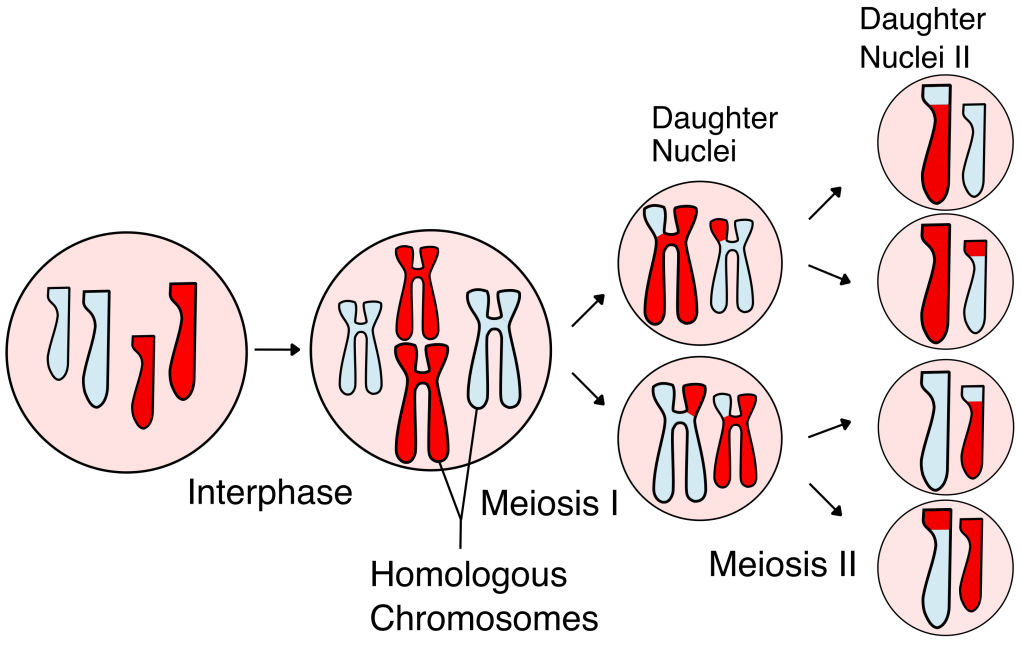
- Brumpton et al. (2020): dynastic effect, assortative mating, and population stratification can strongly bias MR.
- Add family fixed effects to the linear structural equations.

Almost exact inference

- Base inference exactly on the randomness in inheritance.
- Ideas are drawn from:
  1. Randomization inference for experiments (Fisher) and observational data (Rubin, Rosenbaum).
  2. Randomization tests to find causal variants (Spielman, McGinnis, & Ewens, 1993; Bates et al., 2020).

Mendelian randomization: Two stages

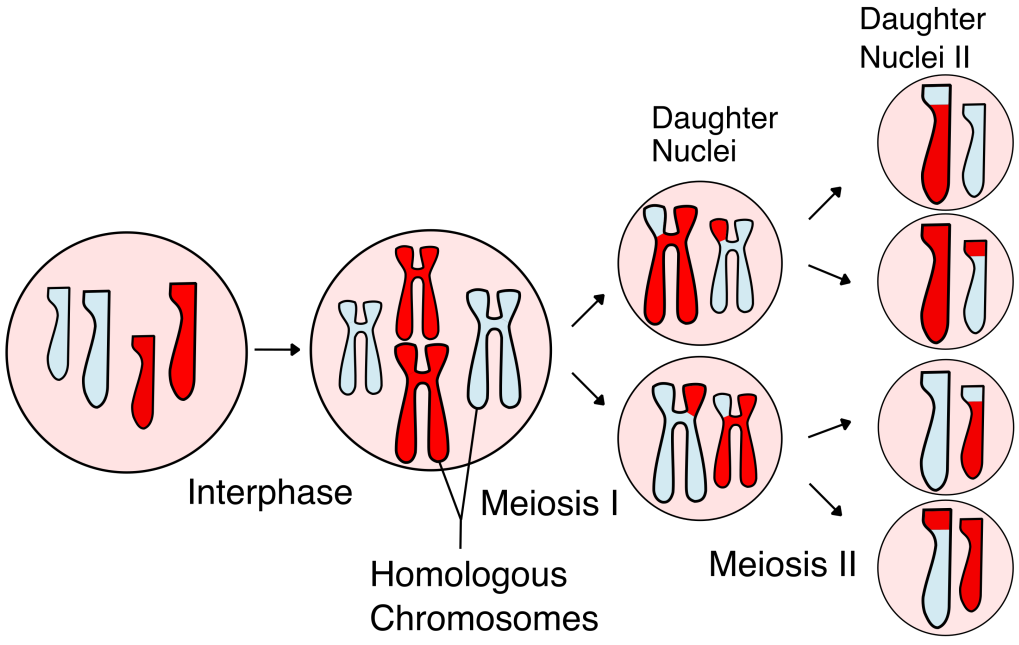
Genetic recombination in meiosis



- Meiosis is a special type of cell division to produce gametes.

Mendelian randomization: Two stages

Genetic recombination in meiosis



- Meiosis is a special type of cell division to produce gametes.

Fertilization

- Fusion of gametes (sperm and egg cell) is completely at random.
- However, mating is usually at random.

Genetic trio studies

Data: Genotypes and phenotypes of mother, father, and offspring.

Genetic trio studies

Data: Genotypes and phenotypes of mother, father, and offspring.

Notation for genetic data

- $M/F/Z$: mother/father/offspring.
- Superscript f/m: Haplotypes inherited from father/mother.
- So $M_j^f \in \{0, 1\}$ is mother's haplotype at locus j inherited from her father.
- No superscript means genotypes: $Z_j = Z_j^f + Z_j^m \in \{0, 1, 2\}$.

Genetic trio studies

Data: Genotypes and phenotypes of mother, father, and offspring.

Notation for genetic data

- $M/F/Z$: mother/father/offspring.
- Superscript f/m: Haplotypes inherited from father/mother.
- So $M_j^f \in \{0,1\}$ is mother's haplotype at locus j inherited from her father.
- No superscript means genotypes: $Z_j = Z_j^f + Z_j^m \in \{0,1,2\}$.

Key ideas

- Spielman et al. (1993): Conditional on parental haplotypes.
- Bates. et al (2020): Use existing models for meiosis to obtain the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{F}^m, \boldsymbol{F}^f$:

$$Z_j^m = M_j^{U_j^m}, \; Z_j^f = F_j^{U_j^f}.$$

- Haldane (1919): Ancestry indicator $\boldsymbol{U}$ follows a Poisson process.

Illustration for within-family MR

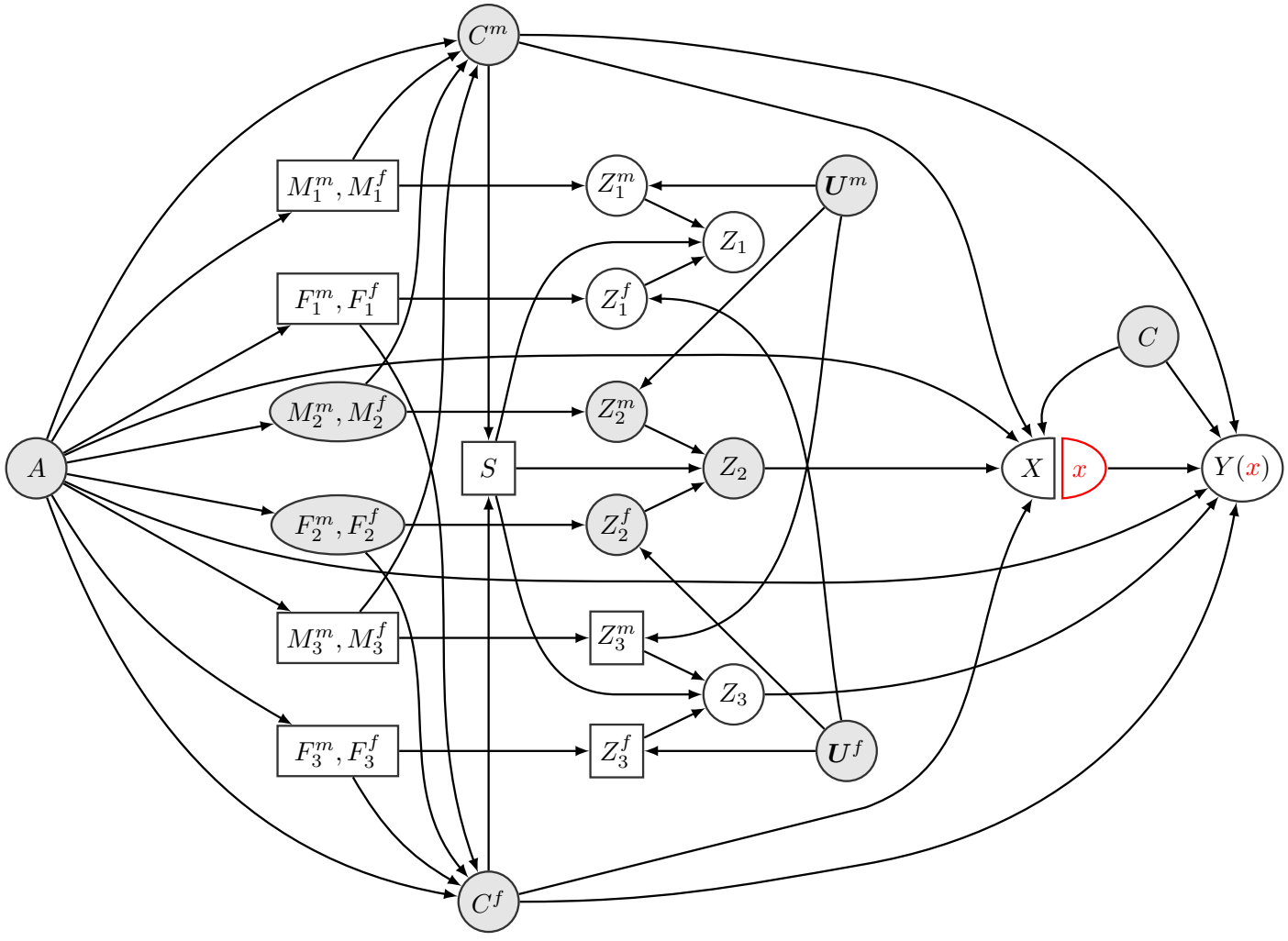


Illustration for within-family MR

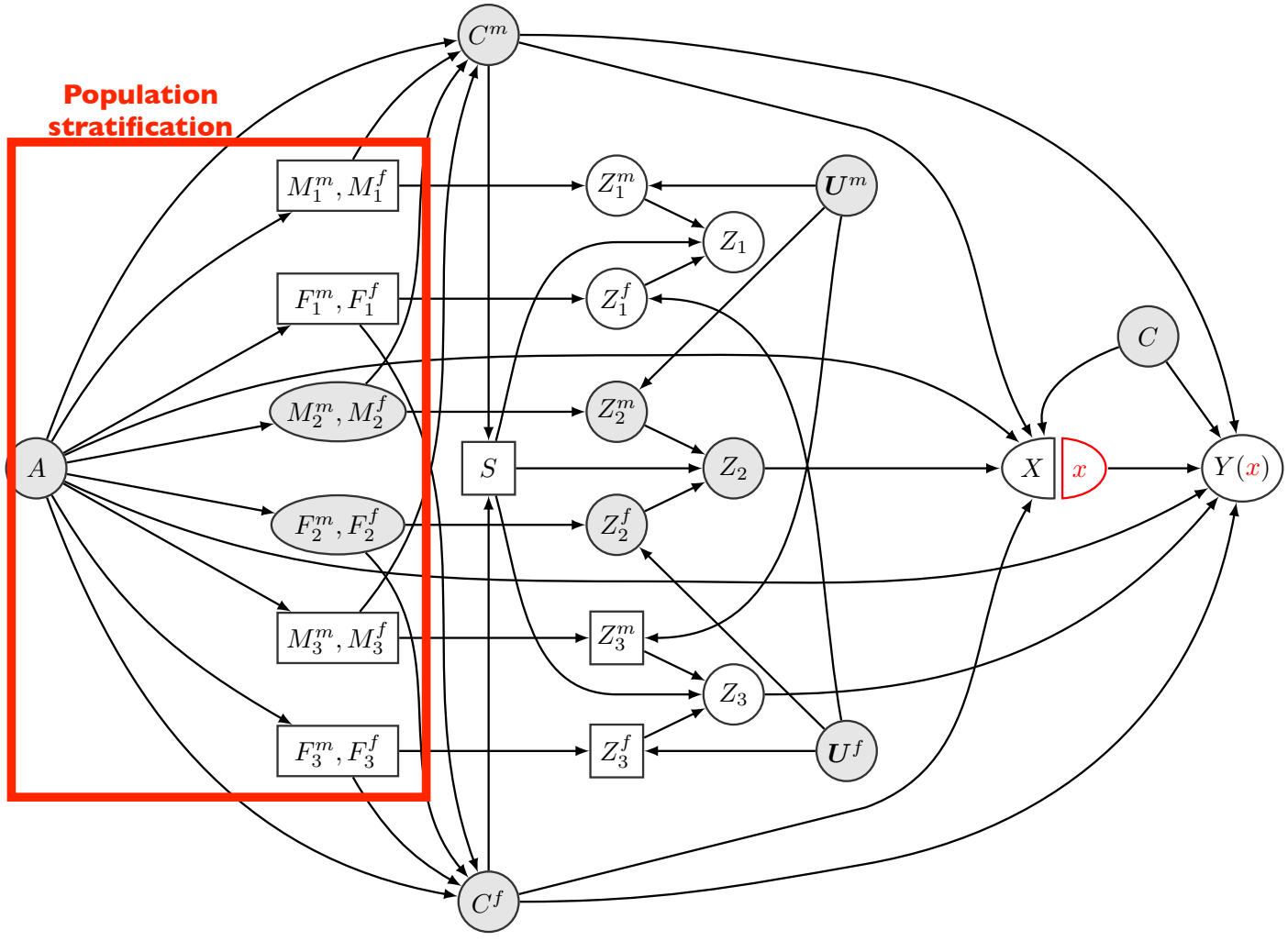


Illustration for within-family MR

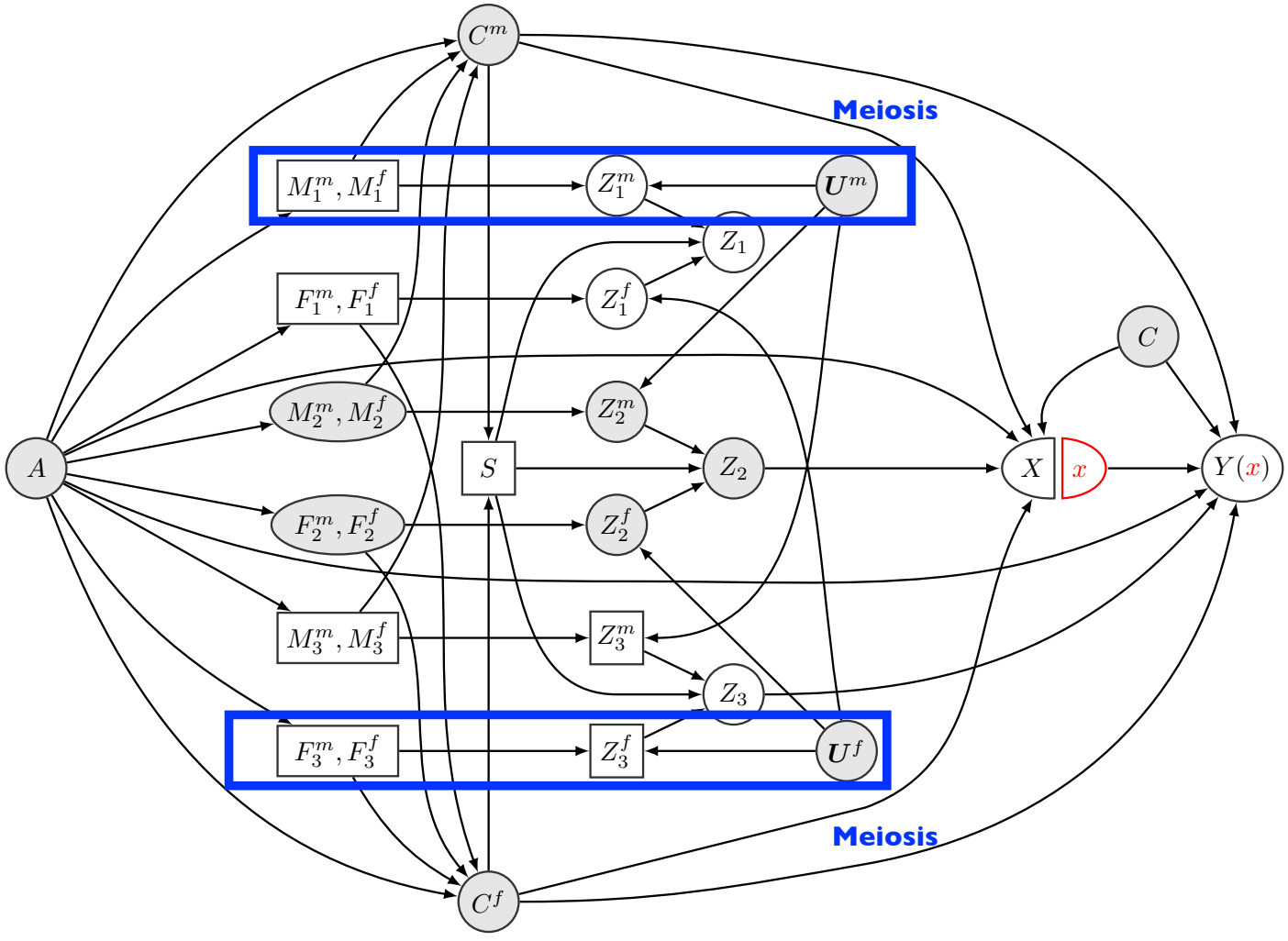


Illustration for within-family MR

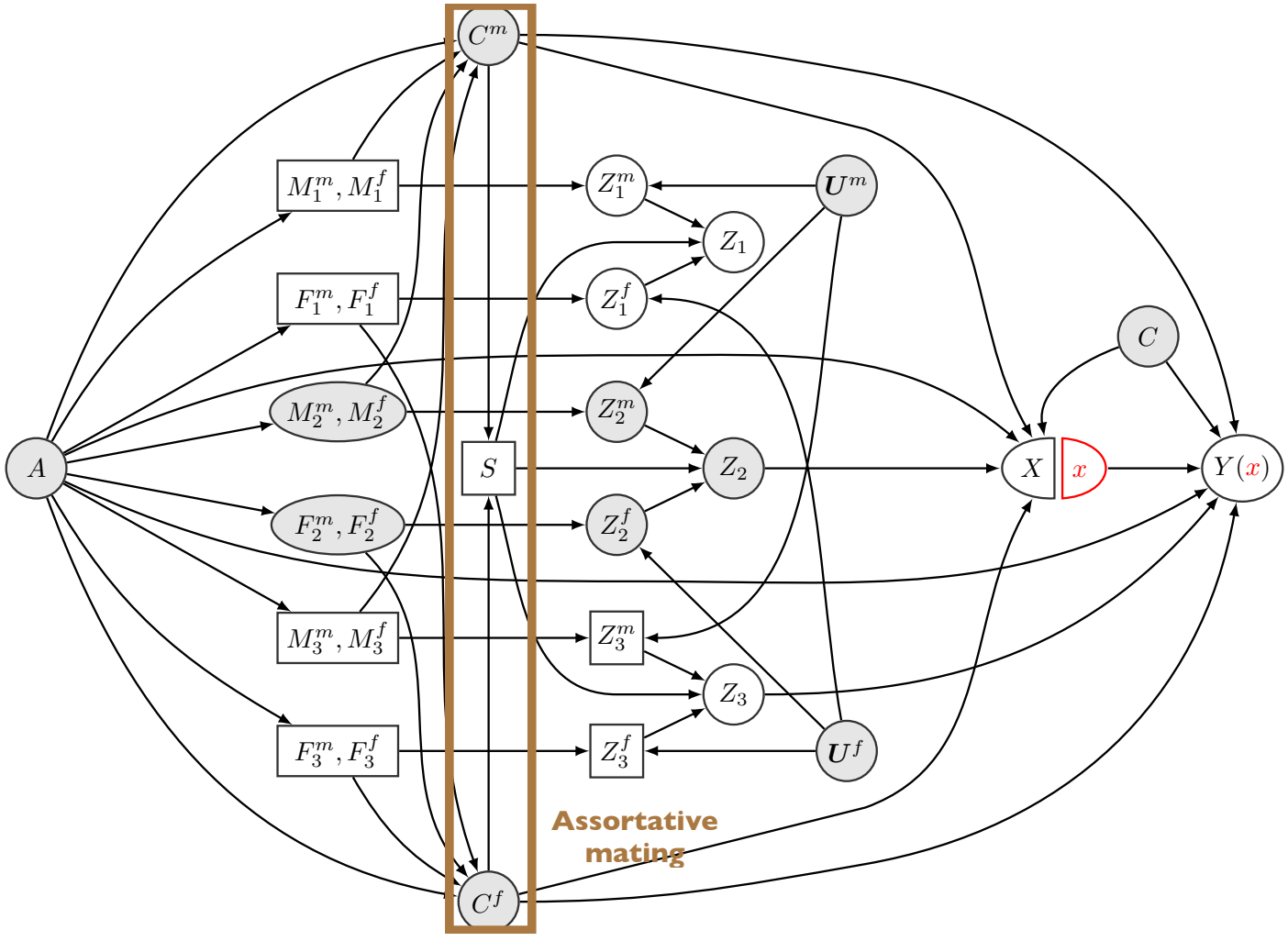


Illustration for within-family MR

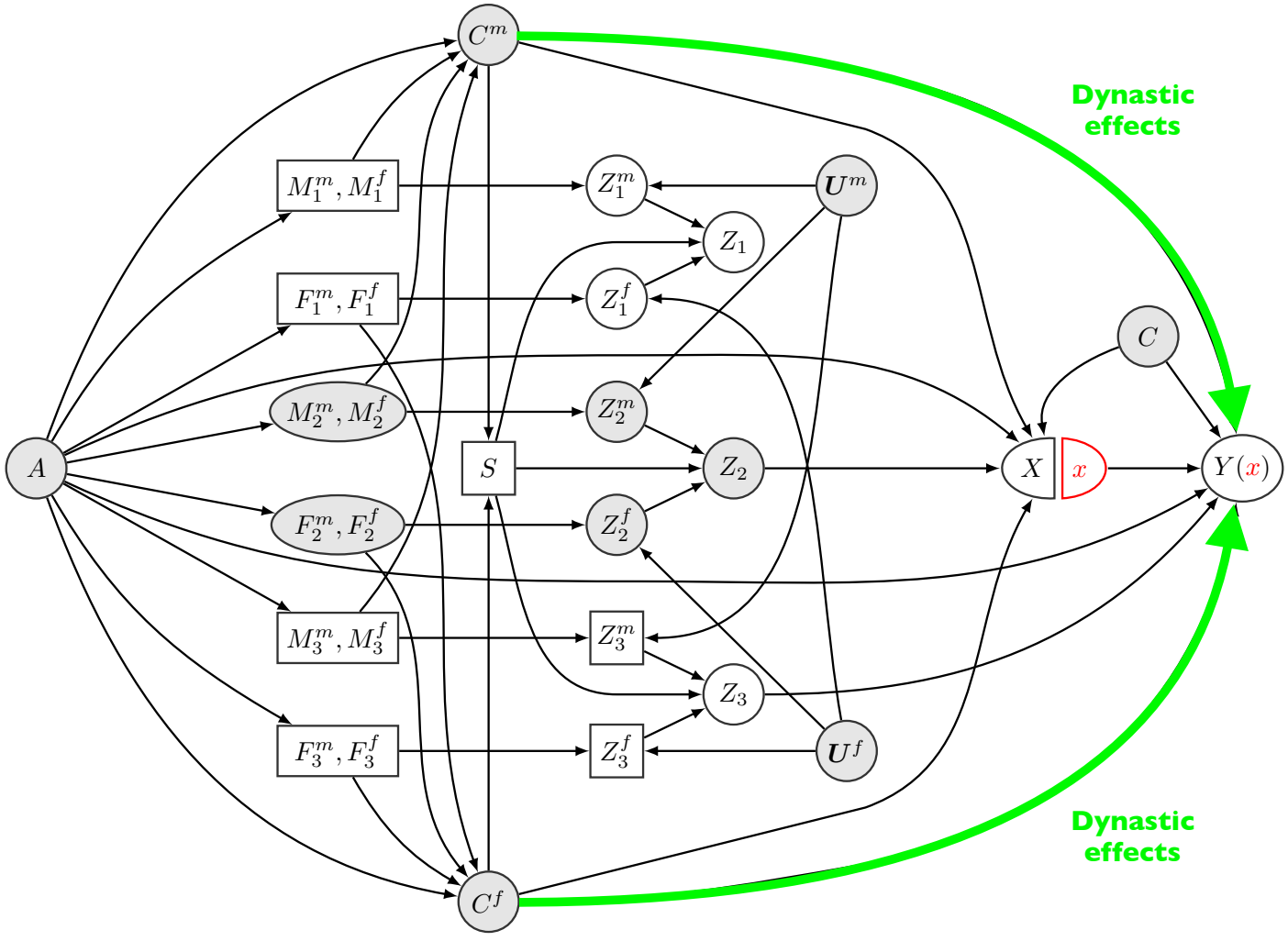


Illustration for within-family MR

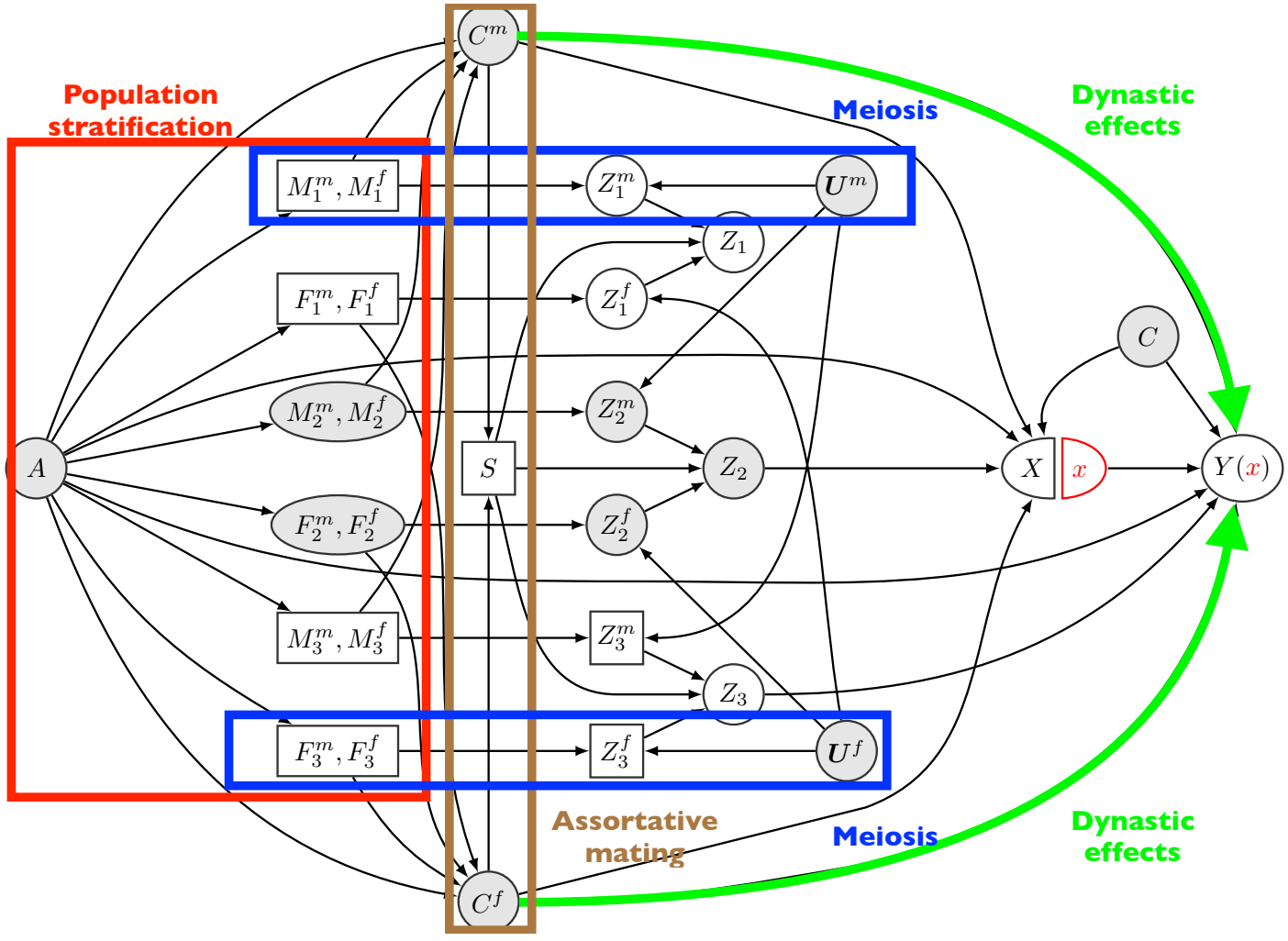
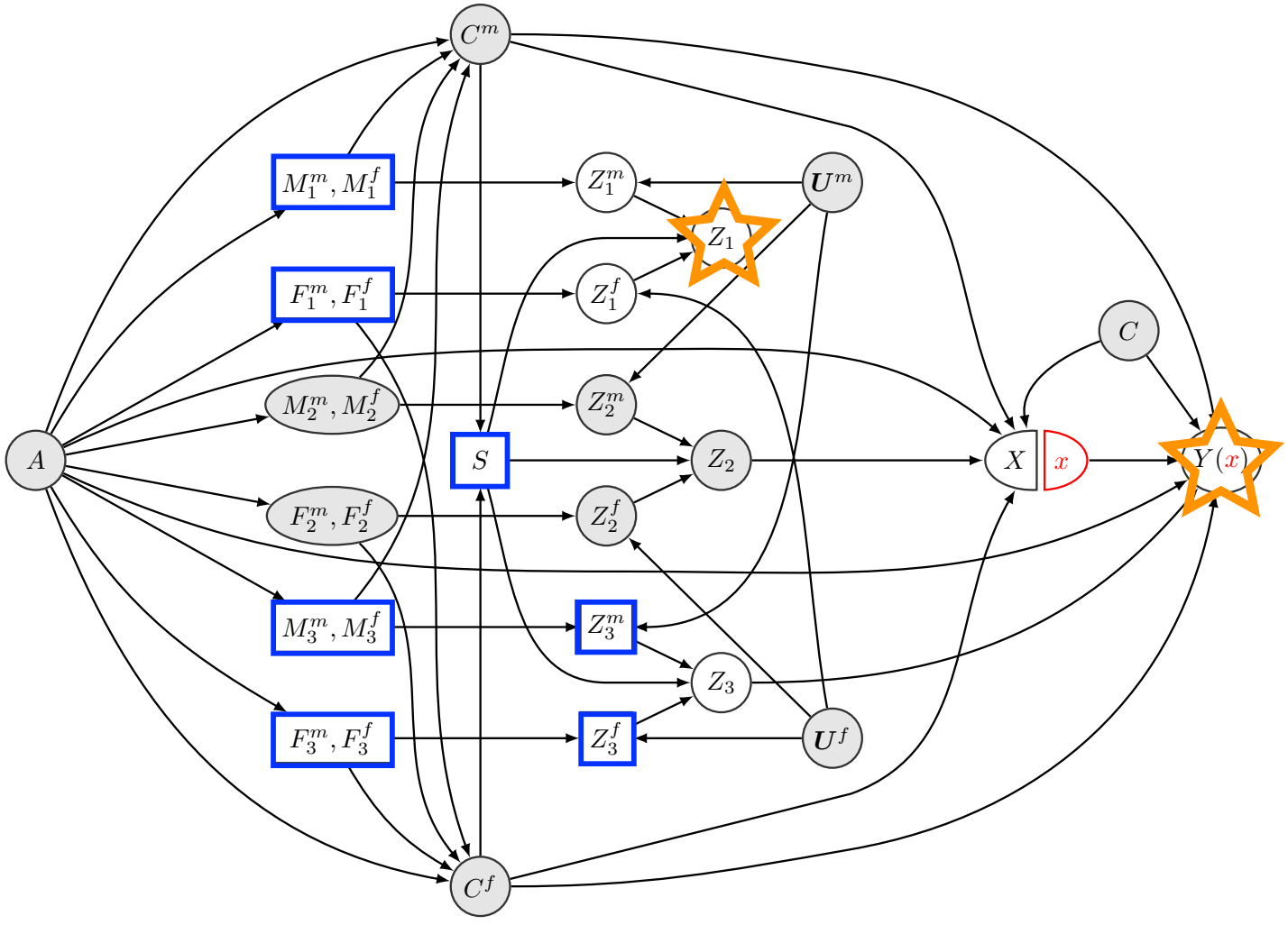
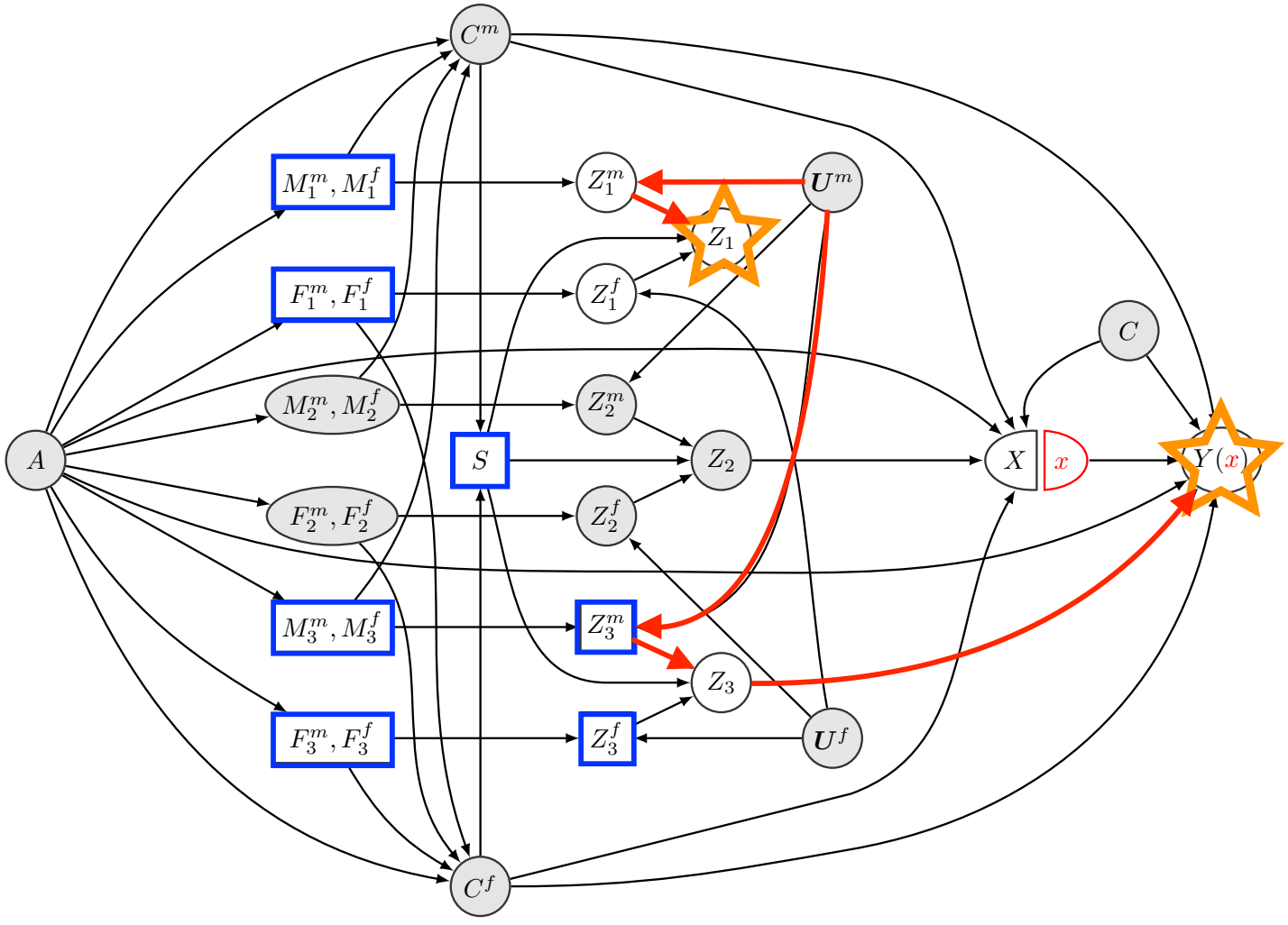


Illustration for within-family MR



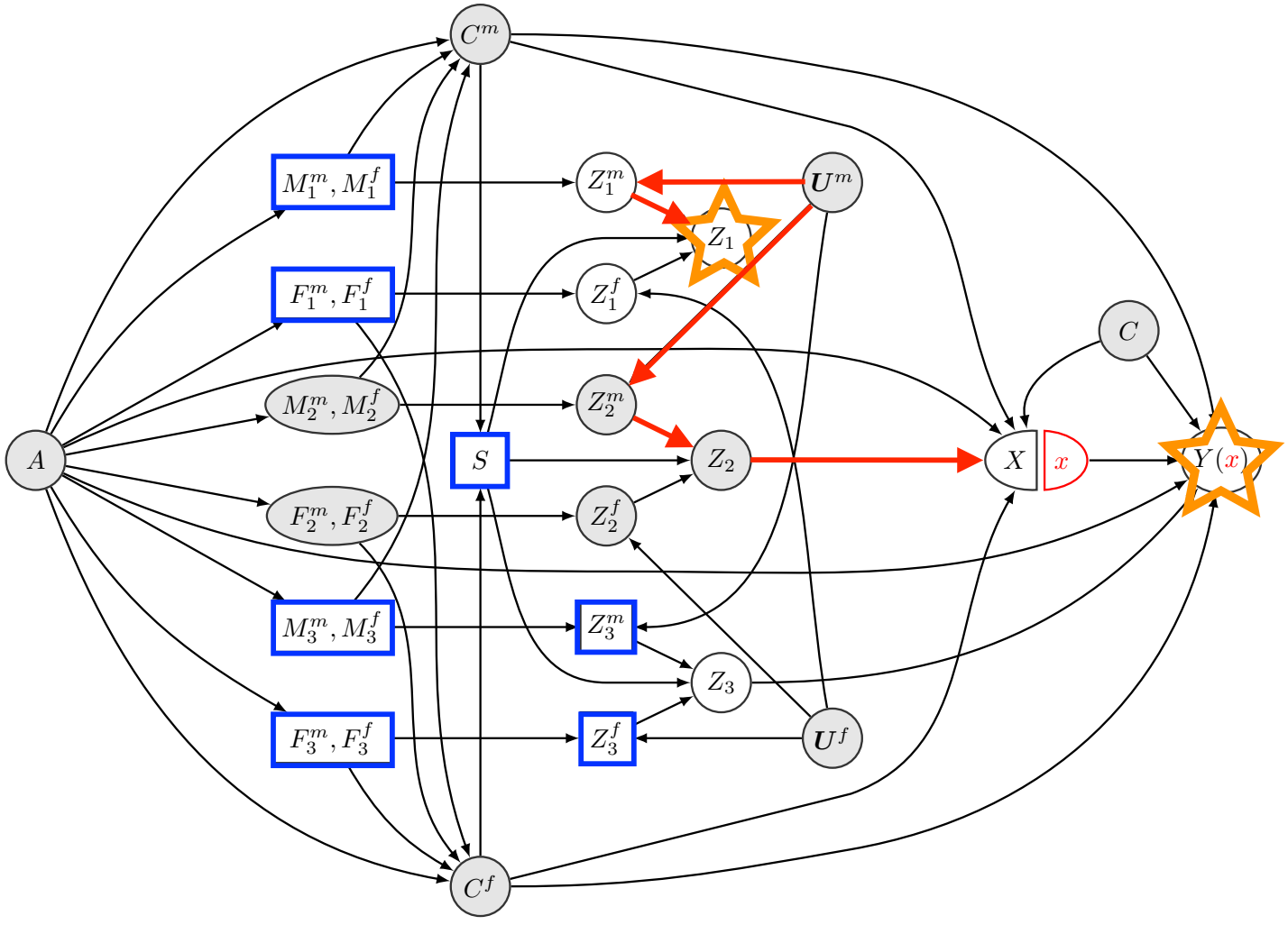
d-separation: $Z_1 \perp\!\!\!\perp Y(x) \mid (\boldsymbol{M}_1^{mf}, \boldsymbol{F}_1^{mf}, \boldsymbol{M}_3^{mf}, \boldsymbol{F}_3^{mf}, \boldsymbol{Z}_3^{mf})$.

Illustration for within-family MR



d-separation: $Z_1 \perp\!\!\!\perp Y(x) \mid (\boldsymbol{M}_1^{mf}, \boldsymbol{F}_1^{mf}, \boldsymbol{M}_3^{mf}, \boldsymbol{F}_3^{mf}, \boldsymbol{Z}_3^{mf})$.

Illustration for within-family MR



d-separation: $Z_1 \perp\!\!\!\perp Y(x) \mid (\boldsymbol{M}_1^{mf}, \boldsymbol{F}_1^{mf}, \boldsymbol{M}_3^{mf}, \boldsymbol{F}_3^{mf}, \boldsymbol{Z}_3^{mf})$.

Ongoing work

- More general results for sufficient adjustment sets.
- Simplification by the Markov structure on $\boldsymbol{U}$ (Haldane, 1919; Bates et al., 2020).
- Randomization test for the sharp null $H_0 : Y(1) - Y(0) = \beta$.
  - Key idea: Under H_0, $Y(0) = Y - \beta X$.
  - This is "almost exact" (exact if model on $\boldsymbol{U}$ is correct).
- Constructing powerful test statistics by incorporating the propensity score (Rosenbaum & Rubin, 1983).

Take-home messages

- Mendelian randomization dates back to the dawn of modern statistics and genetics.
- New life of an old idea:

  MR = Base causal inference on randomness in Mendelian inheritance.

- Challenges remain:
  1. Pleiotropy;
  2. Computation;
  3. Incomplete pedigrees.