Cross-fitted instrument: a blueprint for one-sample Mendelian Randomization

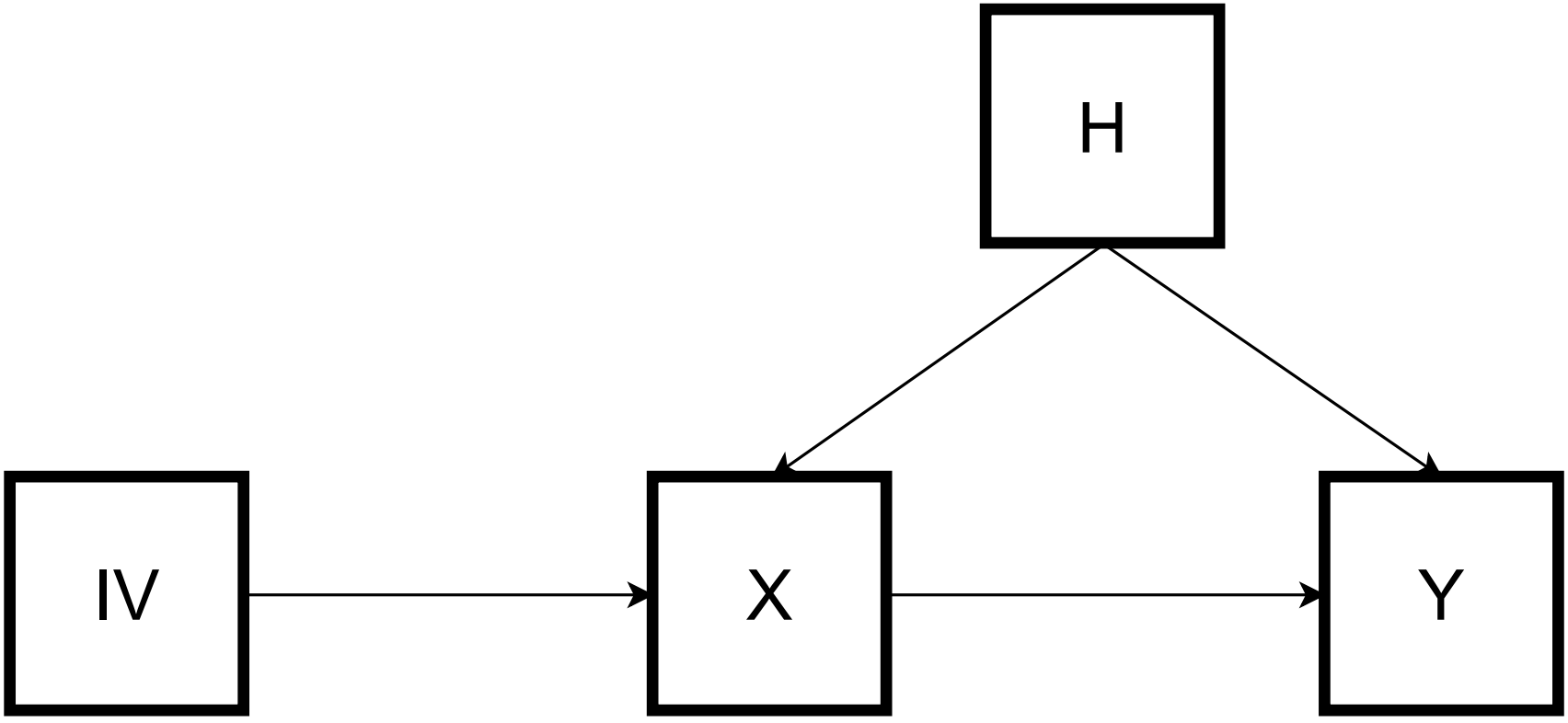
William R.P. Denault

Center for Fertility and Health: Norwegian Institute of Public Health

william.denault@gmail.com

July 2^{nd}, 2021

Instrumental variable



Mendelian Randomisation: using genotype as an instrument

- First idea in by Katan, APOUPOPROTEIN E ISOFORMS, SERUM CHOLESTEROL, AND CANCER, *Lancet*, 1986
- R Gray, K Wheatley, How to avoid bias when comparing bone marrow transplantation with chemotherapy, *Bone Marrow Transplantation*, 1991
- Davey Smith, Mendelian Randomization for Strengthening Causal Inference in Observational Studies: Application to Gene $\times$ Environment Interactions, *Perspectives on Psychological Science*, 2010

Random inheritance of genotype

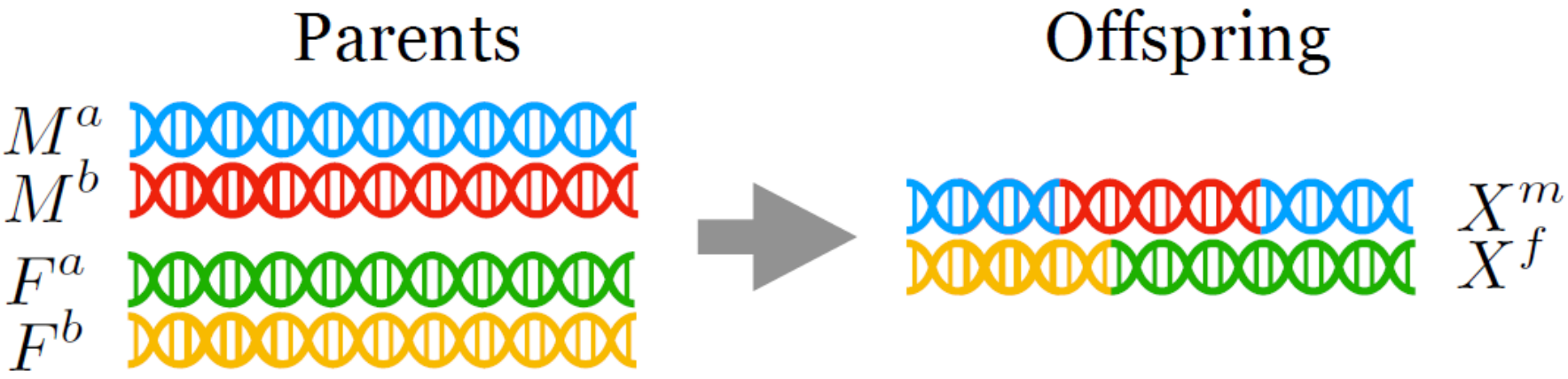
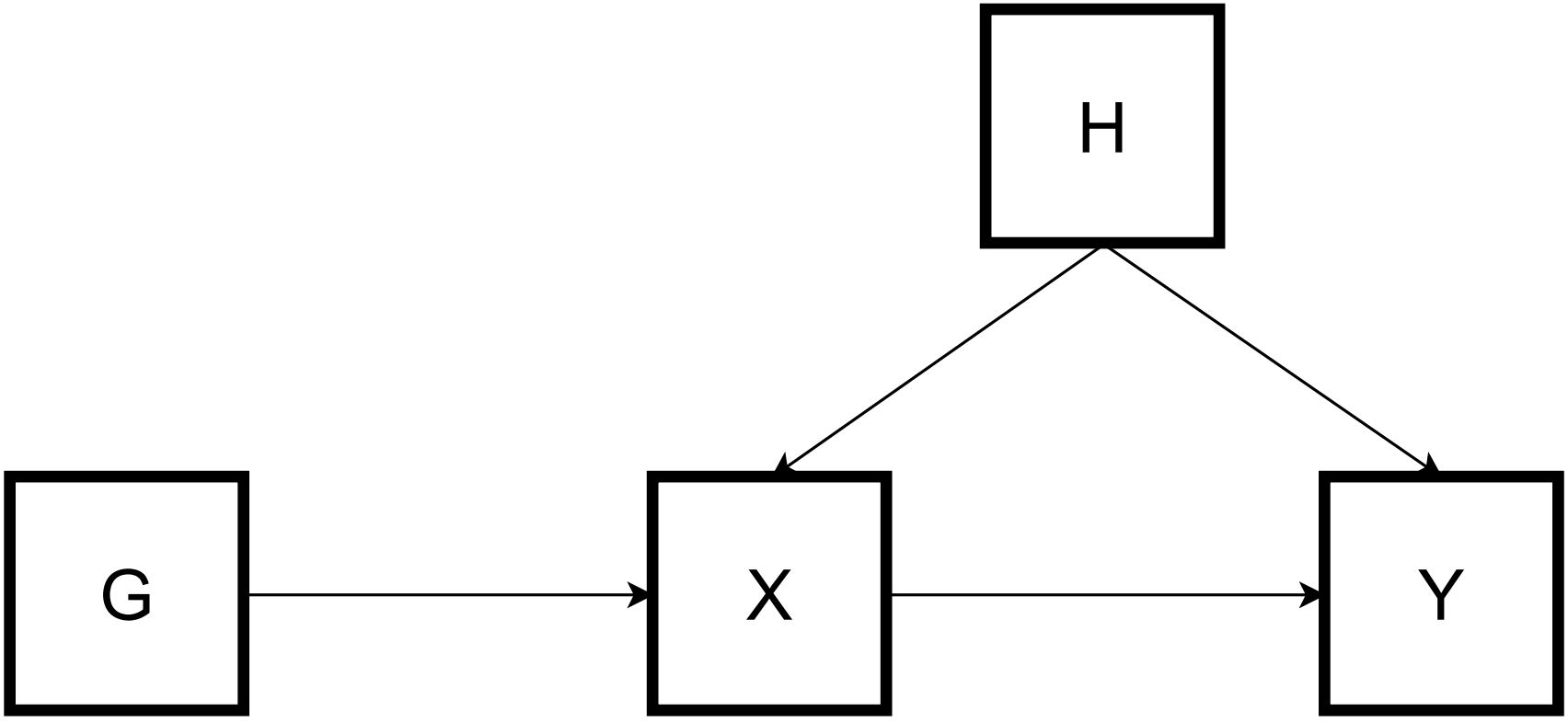
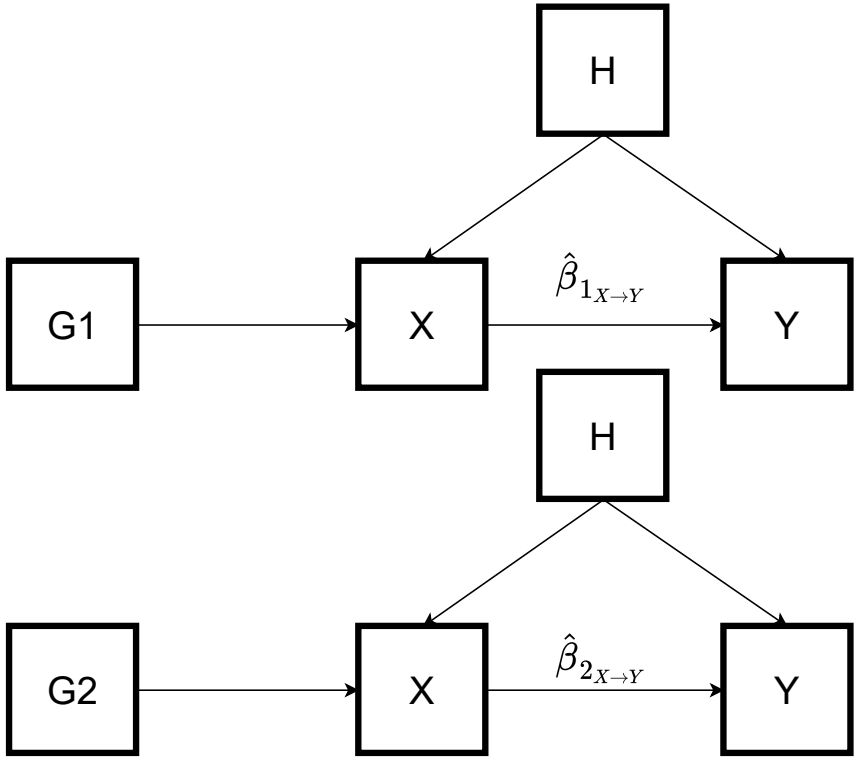


Figure from Bates *et al.* Causal Inference in Genetic Trio Studies, *PNAS*, 2021

Using genotype as an instrument

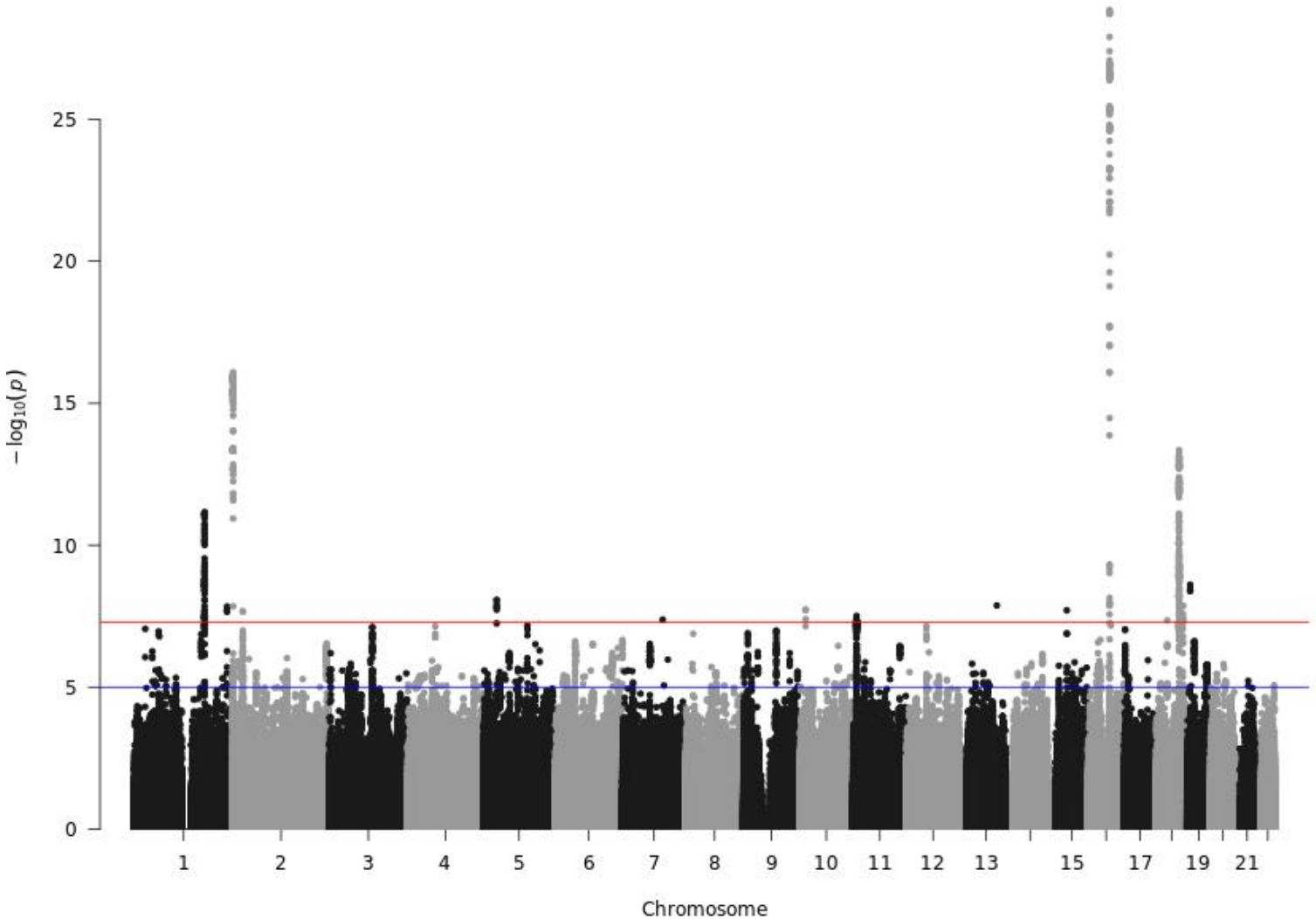


Using genotype as instruments

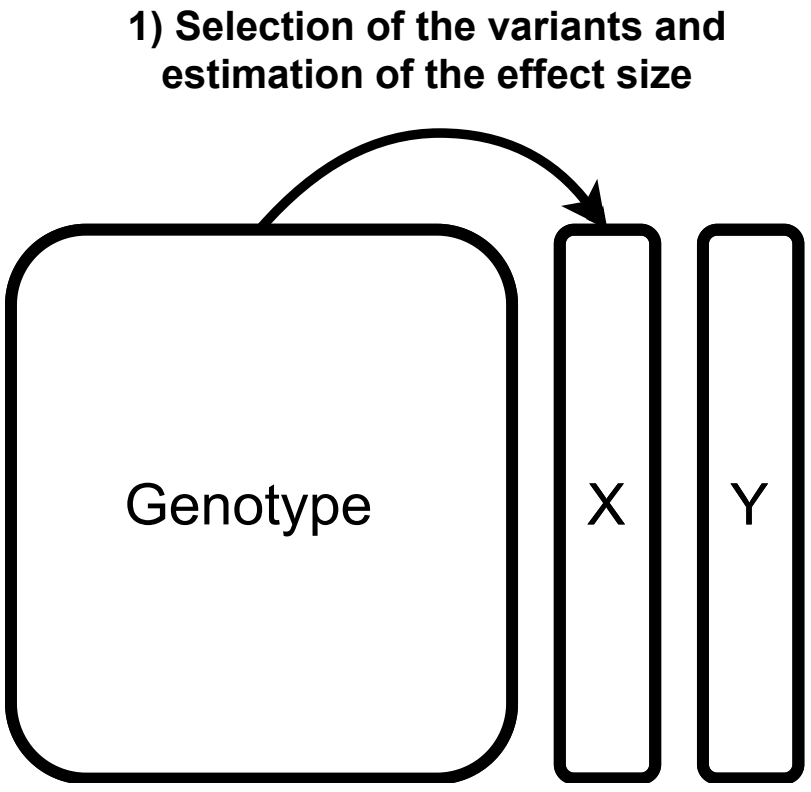


...

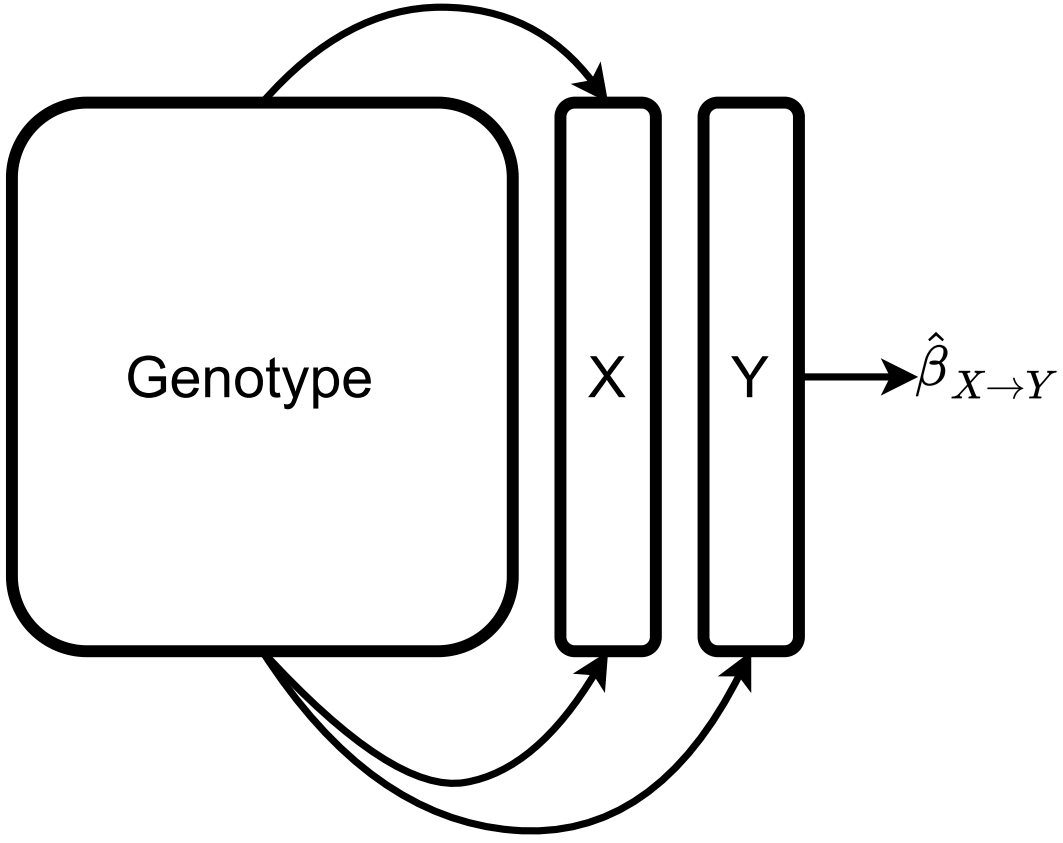
Selection of the instruments: GWAS



The basic strategy: one sample MR



The basic strategy: one sample MR



2) For each variant selected: two stage least square

Endogeneity/weak instrument bias

$$Y = \beta_{X \to Y} X + H + U, \qquad \mathbb{E}[U|\Pi, X] = 0 \tag{1}$$

$$X = Z\Pi + H + V, \qquad \mathbb{E}[V|X] = 0 \tag{2}$$

- $\beta_{X \to Y}$ is the effect of X on Y
- Π is the vector of regression coefficients for the instruments
- U and V are two correlated error terms
- H hidden confounder

Endogeneity/weak instrument bias

Nagar, The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* 1959

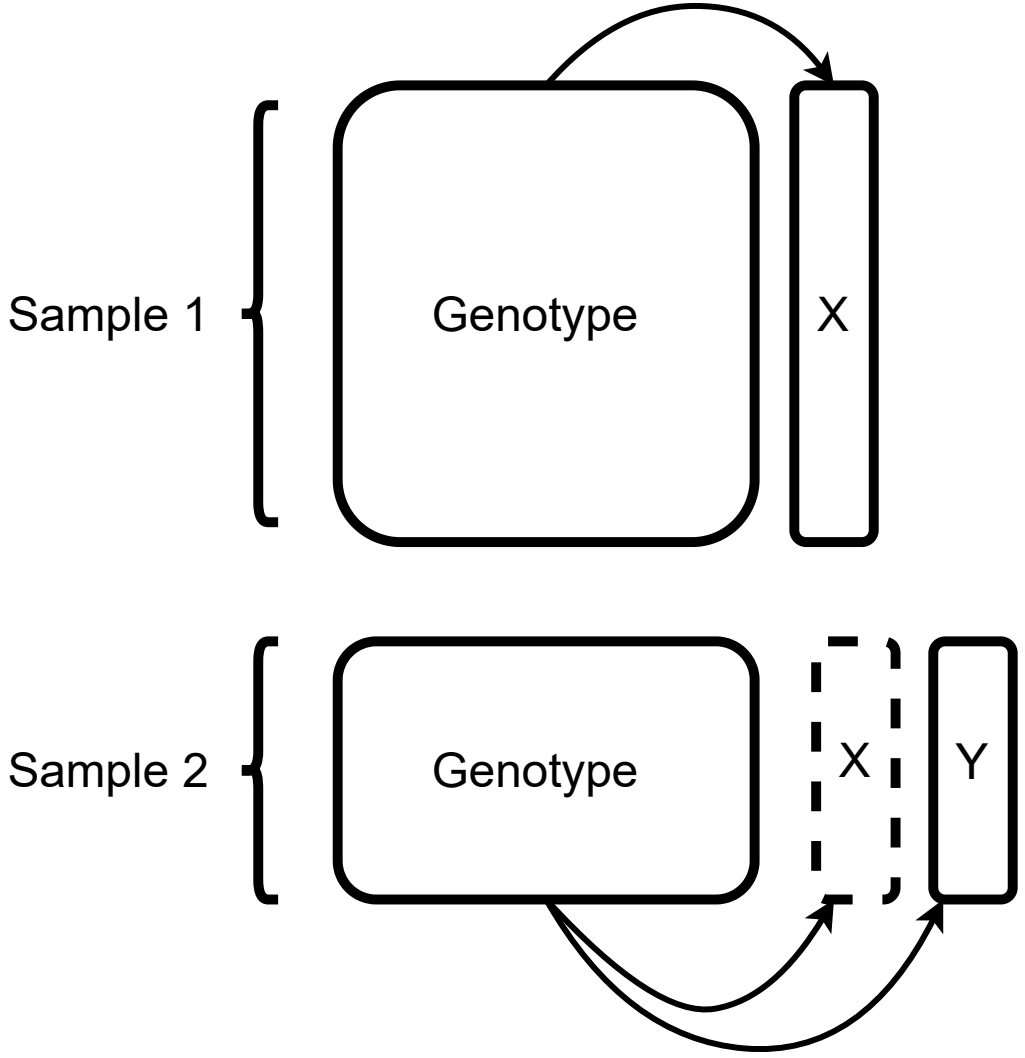
$$\text{Bias tsls} \approx \frac{\sigma_{U,V}}{\mathbb{E}(F)\sigma_U^2} \tag{3}$$

- $\sigma_{U,V}$ covariance of the error terms in the first- and second-stage regression models
- $F \approx$ strength of the instruments, sample size

Reducing the bias from tsls

1. Increase sample size
2. Find stronger instruments: SNPs have small effect size
3. Set $\sigma_{U,V}$ to 0

The two-sample MR



The two-sample MR: source of bias

Sample overlap

- Burgess and colleagues (2016) showed that if sample 1 and sample 2 are overlapping, endogeneity bias has to be expected. (Mounier and Kutalik, Correction for sample overlap, winner's curse and weak instrument bias in two-sample Mendelian Randomization, *BiorXiv*, March 28 2021)

Population heterogeneity

- The effect of a SNP can vary from a population to another (due to change in minor allele frequency). A SNP could be causal for the exposure (sample 1) but could be constant within another population.

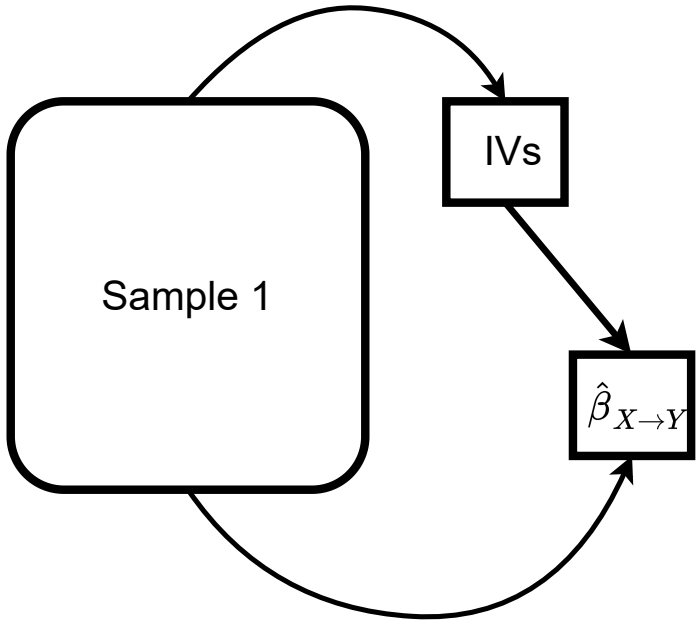
One sample MR

Pros

- Homogeneous population
- Fast

Cons

- Endogeneity bias/winner curse
- Overconfident confidence interval



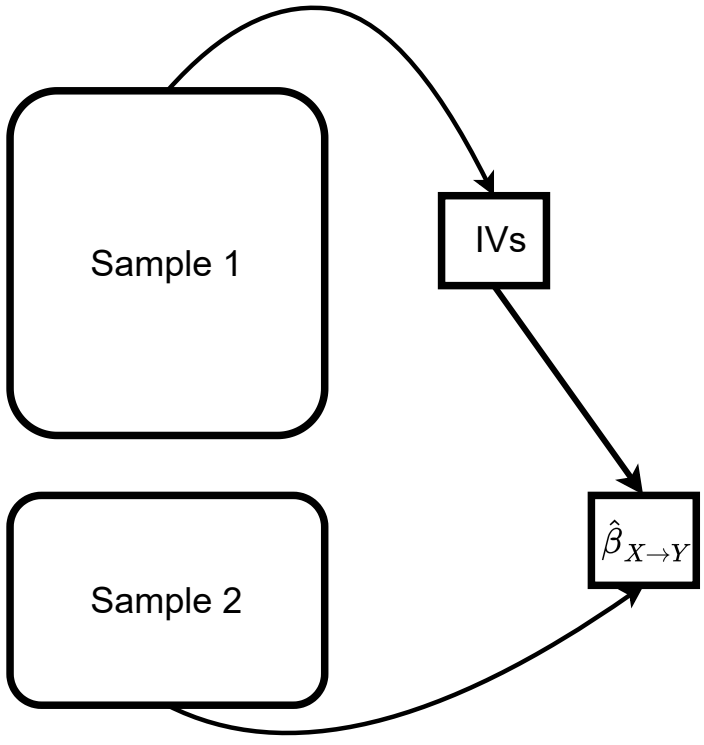
Two-sample MR

Pros

- Less prone to edogeneity bias
- Use of summary statistics available online

Cons

- Potentially unfeasible for rare or expensive phenotype
- Potentially slow
- Severe waste of data



Getting the best of both worlds

Endogeneity free one sample MR

- Propose an approach that use only one sample and that has no endogeneity bias/winner's curse
- We developed the concept of *cross-fitted instrument*/*cross-fitted instruments* (CFI/CFIs)

Build on

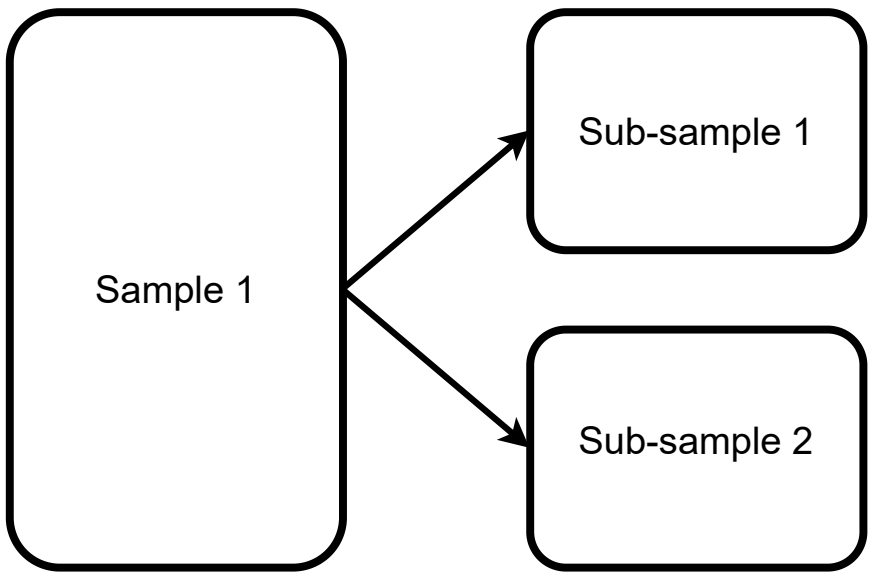
- Double Machine learning by Chernozhukov *et al.*, *The Econometrics Journal*, 2017
- Older approaches such as Split sample IV or Jackknifed IV from Angrist and Krueger, 1995 and Angrist, Krueger and Imbens, 1999

CFI: middle ground between Split sample IV and Jackknifed IV

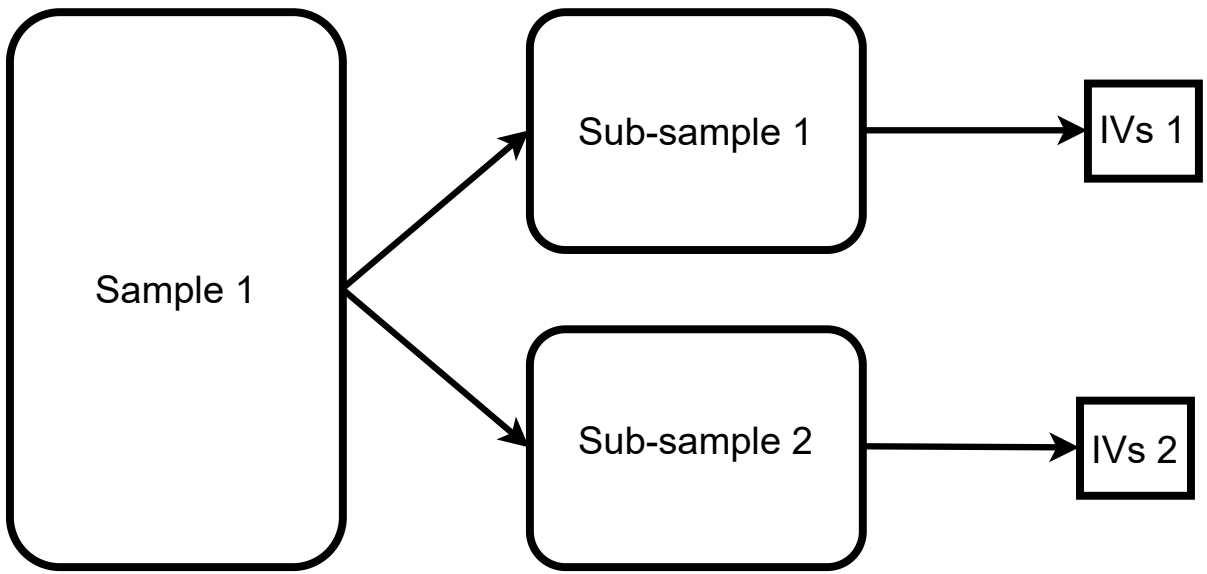
Cross-fitted instrument

- 2-fold cross-fitted instruments
- k-fold cross-fitted instrument/instruments
- CFMR1 and CFMR2

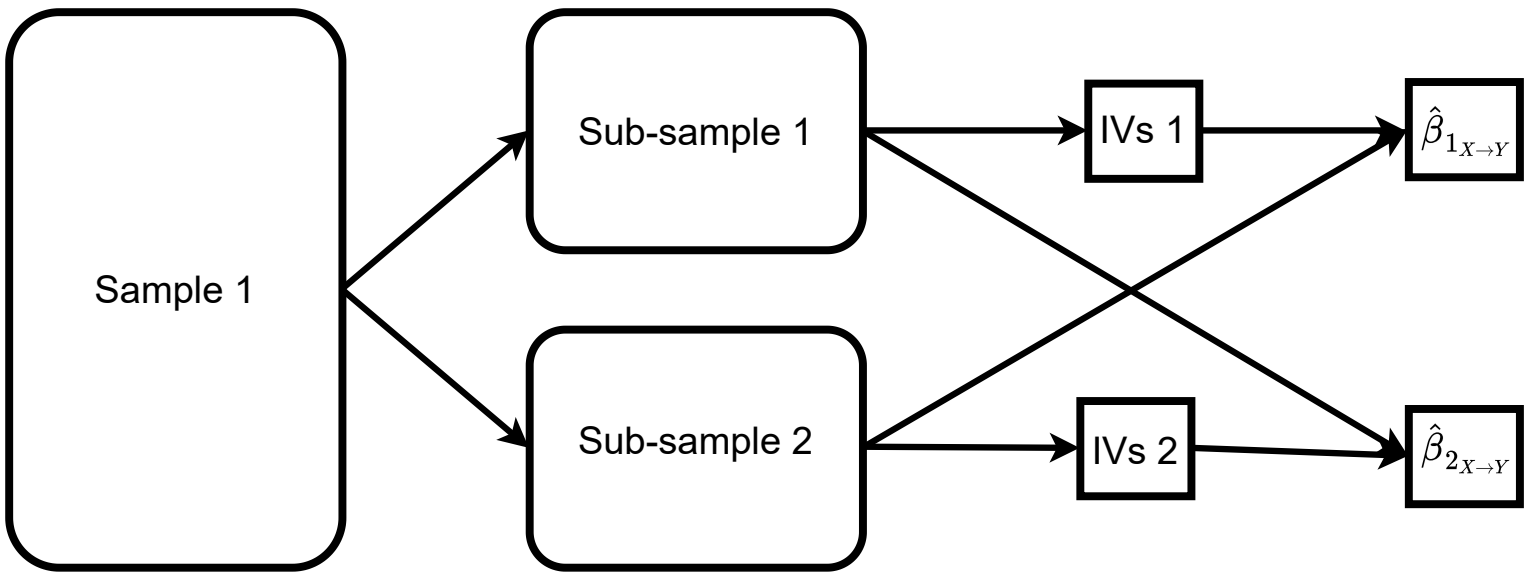
Sample splitting



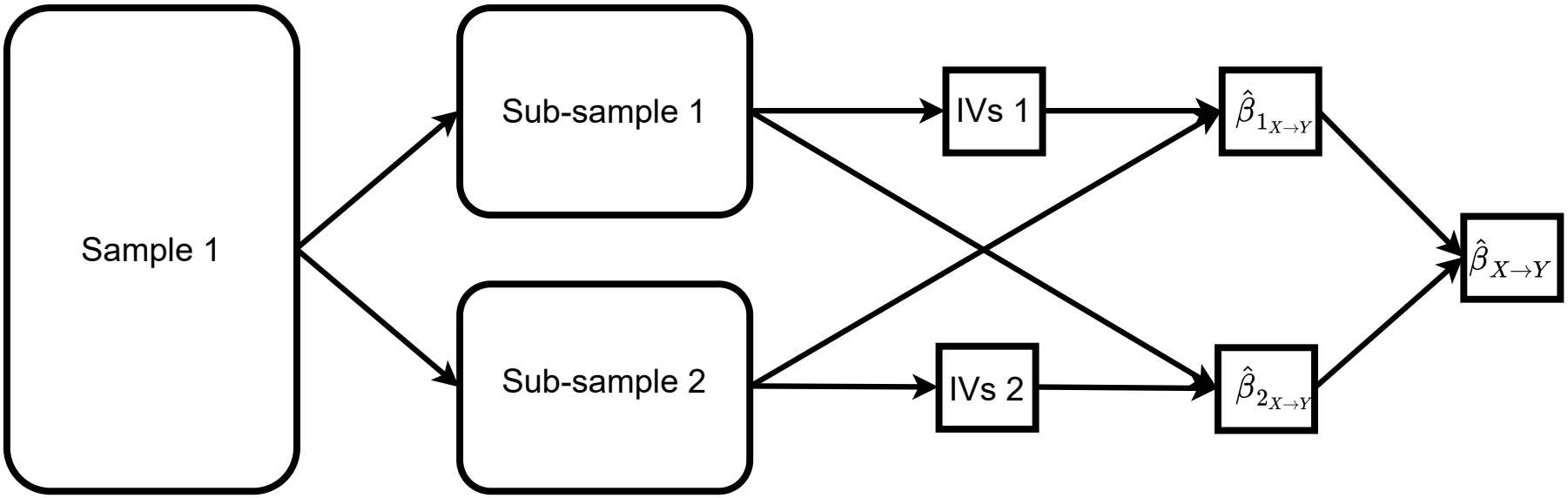
Selection of the instruments



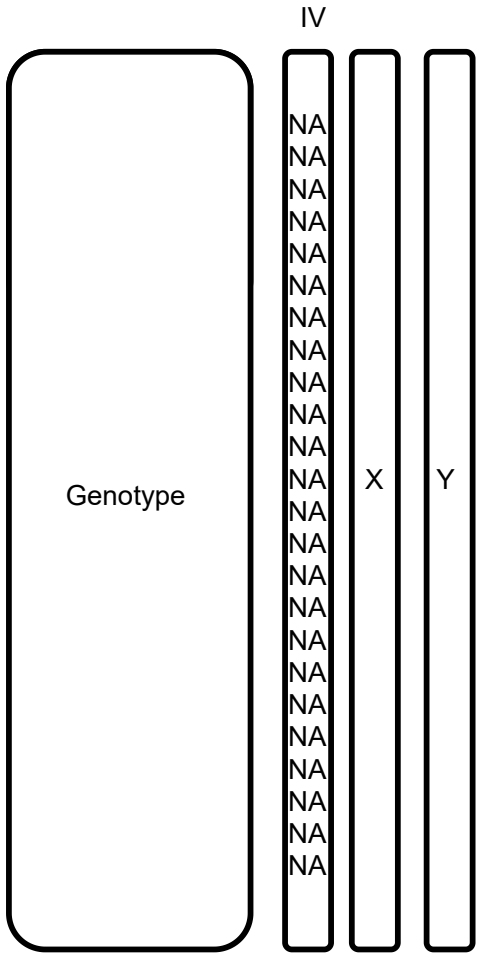
Two stage least squares

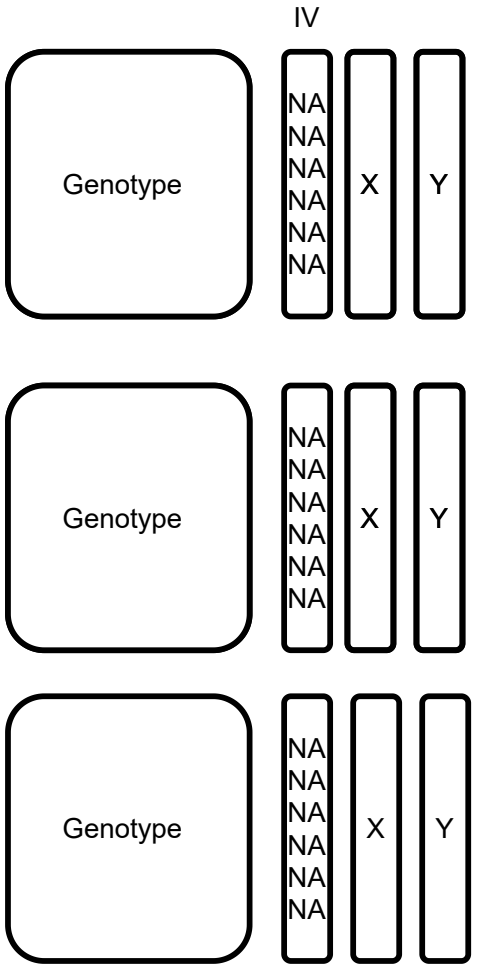


Average

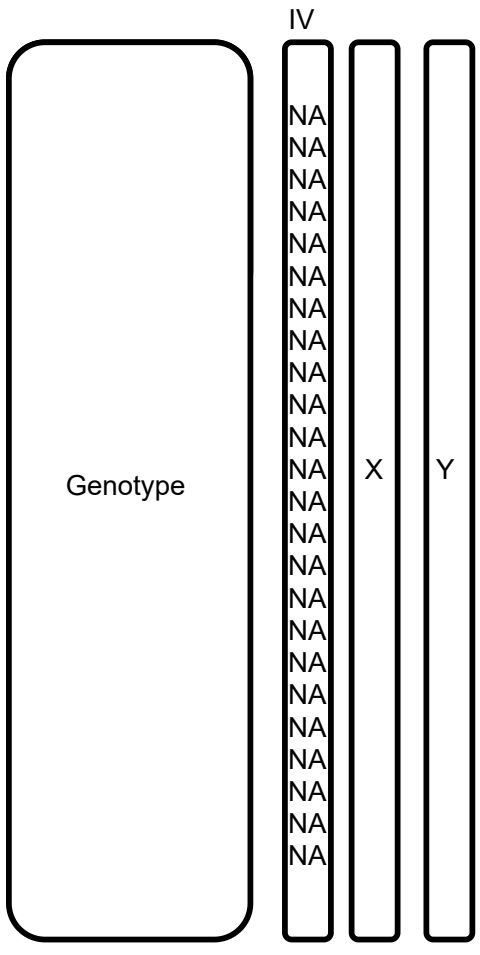

Sample 1
Sub-sample 1
Sub-sample 2
IVs 1
IVs 2
$\hat{\beta}_{1_{X\to Y}}$
$\hat{\beta}_{2_{X\to Y}}$
$\hat{\beta}_{X\to Y}$

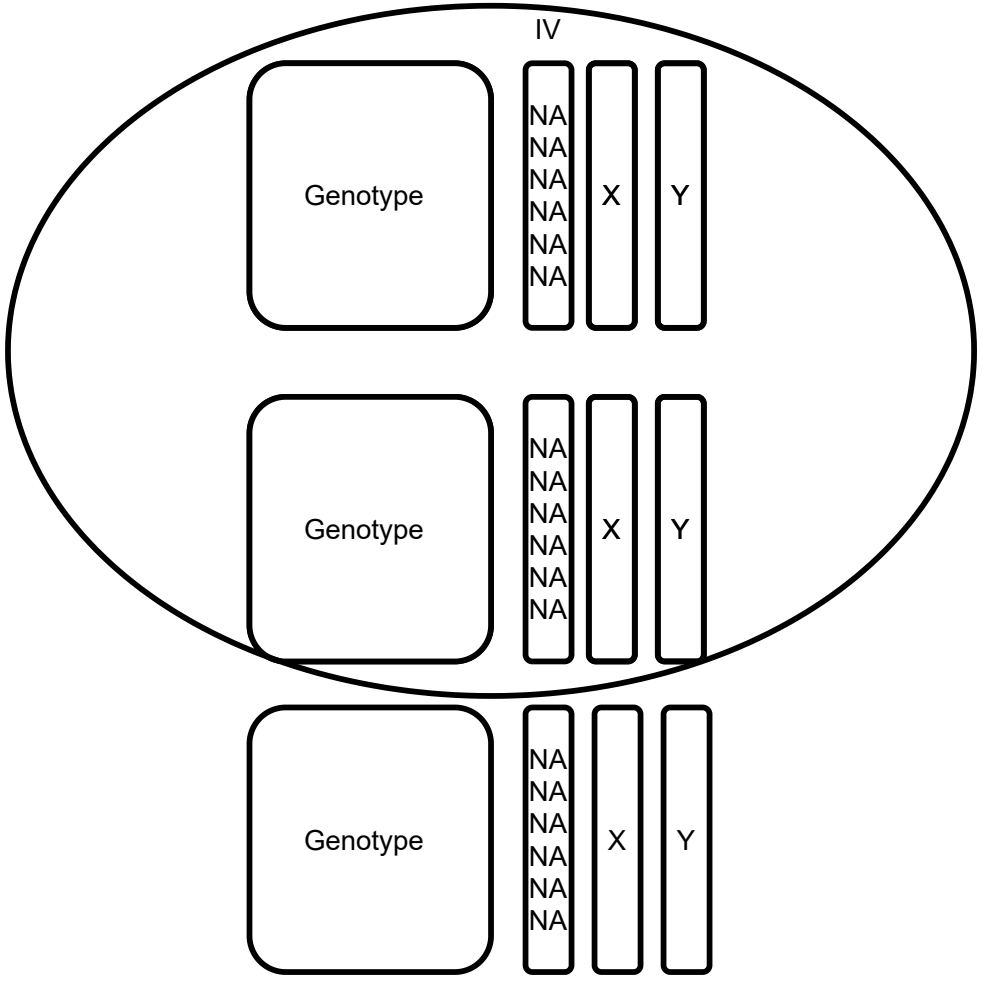
Sample splitting



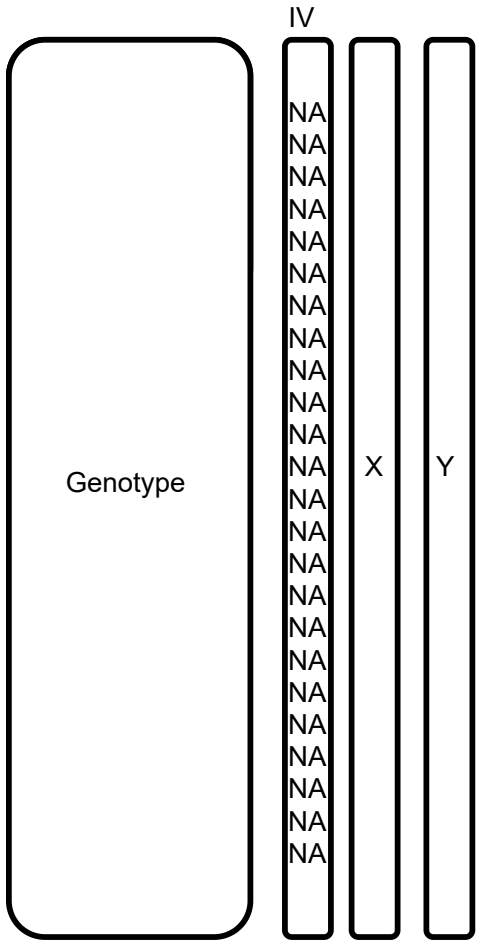


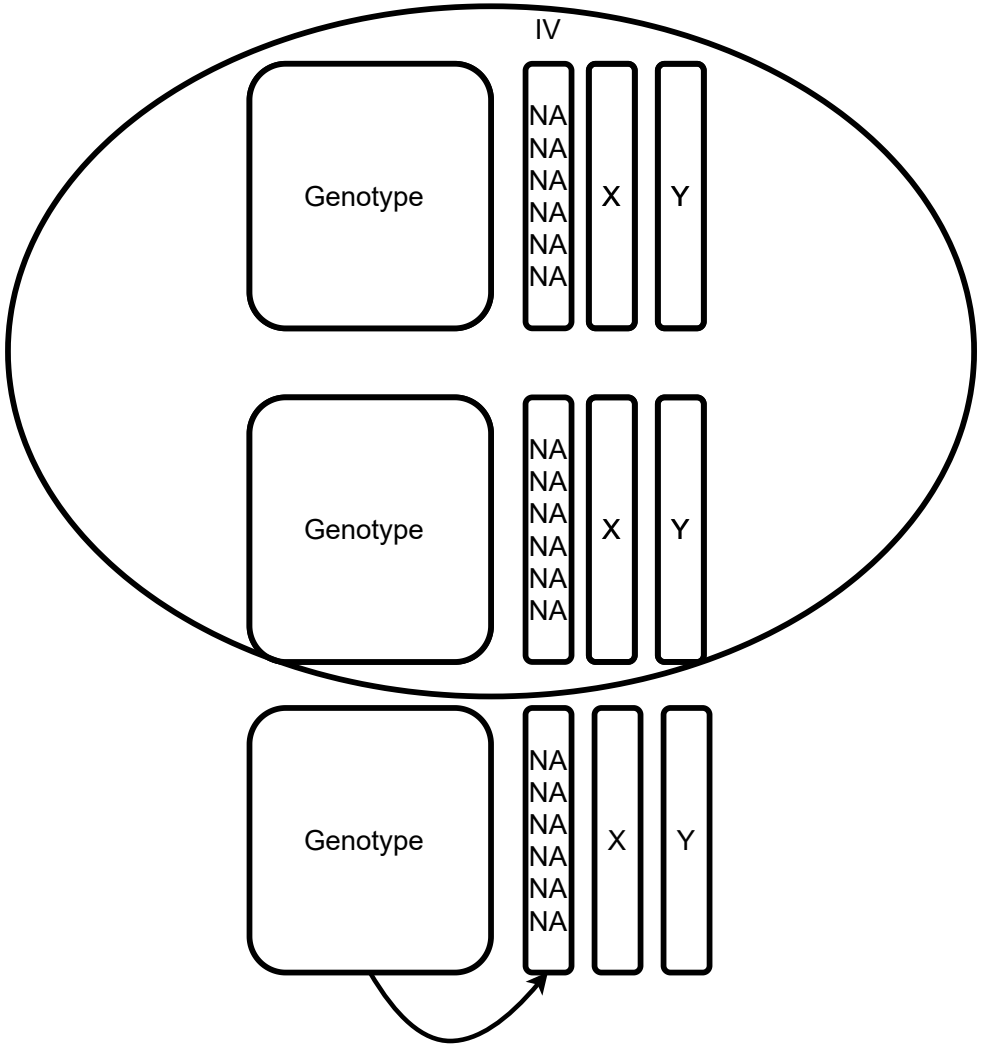
Select instruments using samples 1 and 2



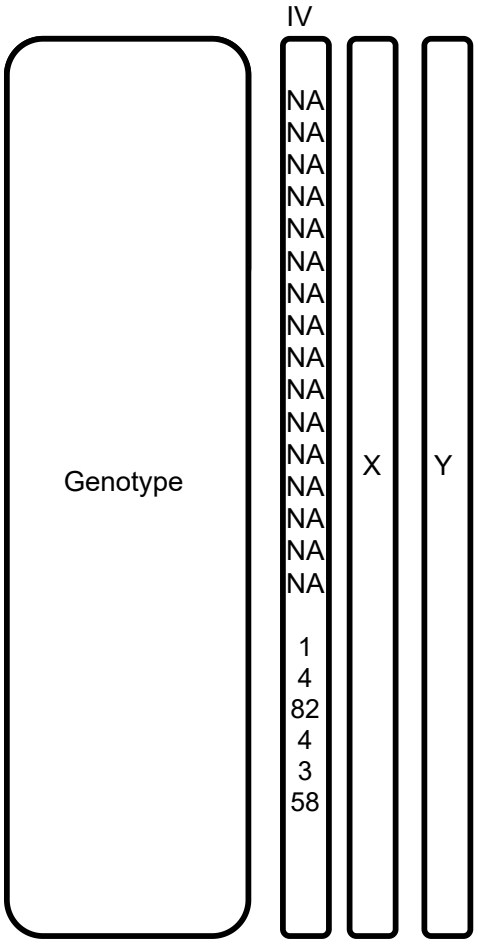


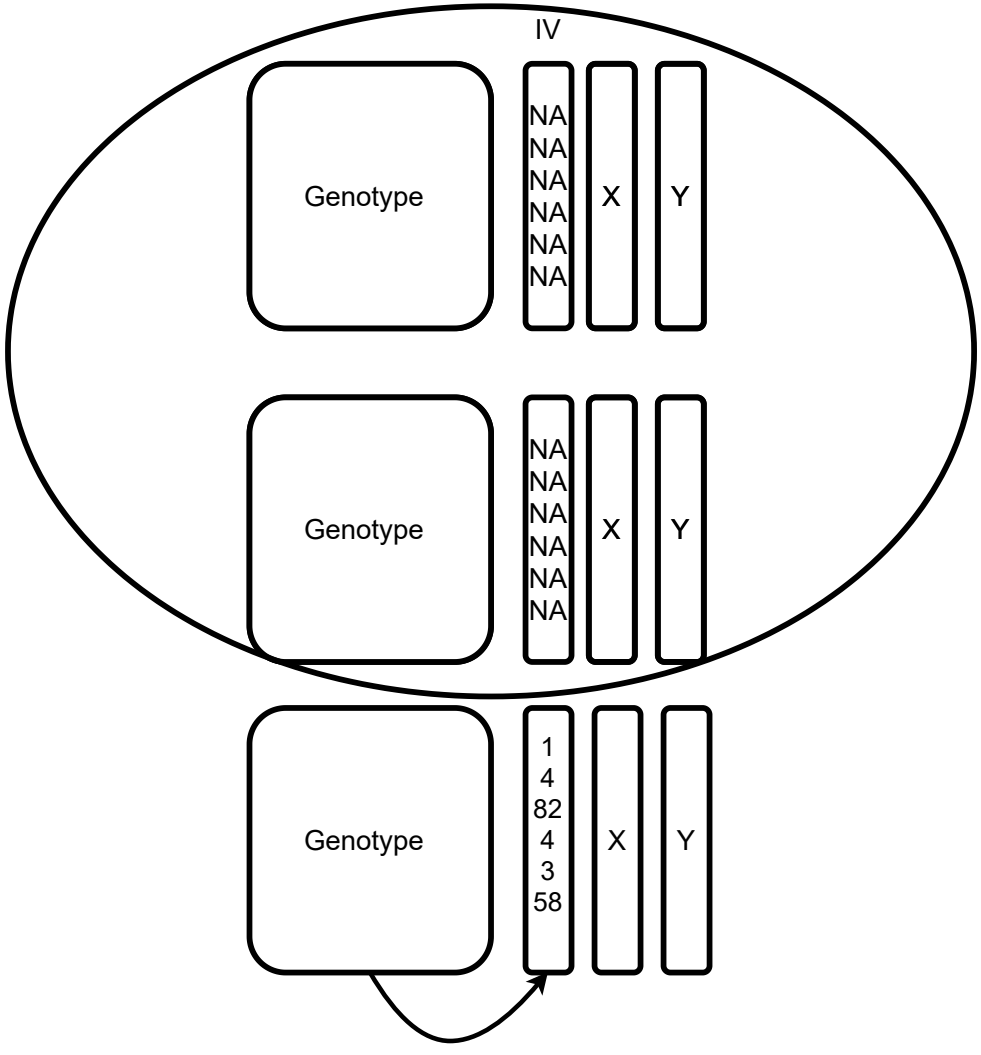
Predict X in sample 3 using estimates from samples 1 and 2



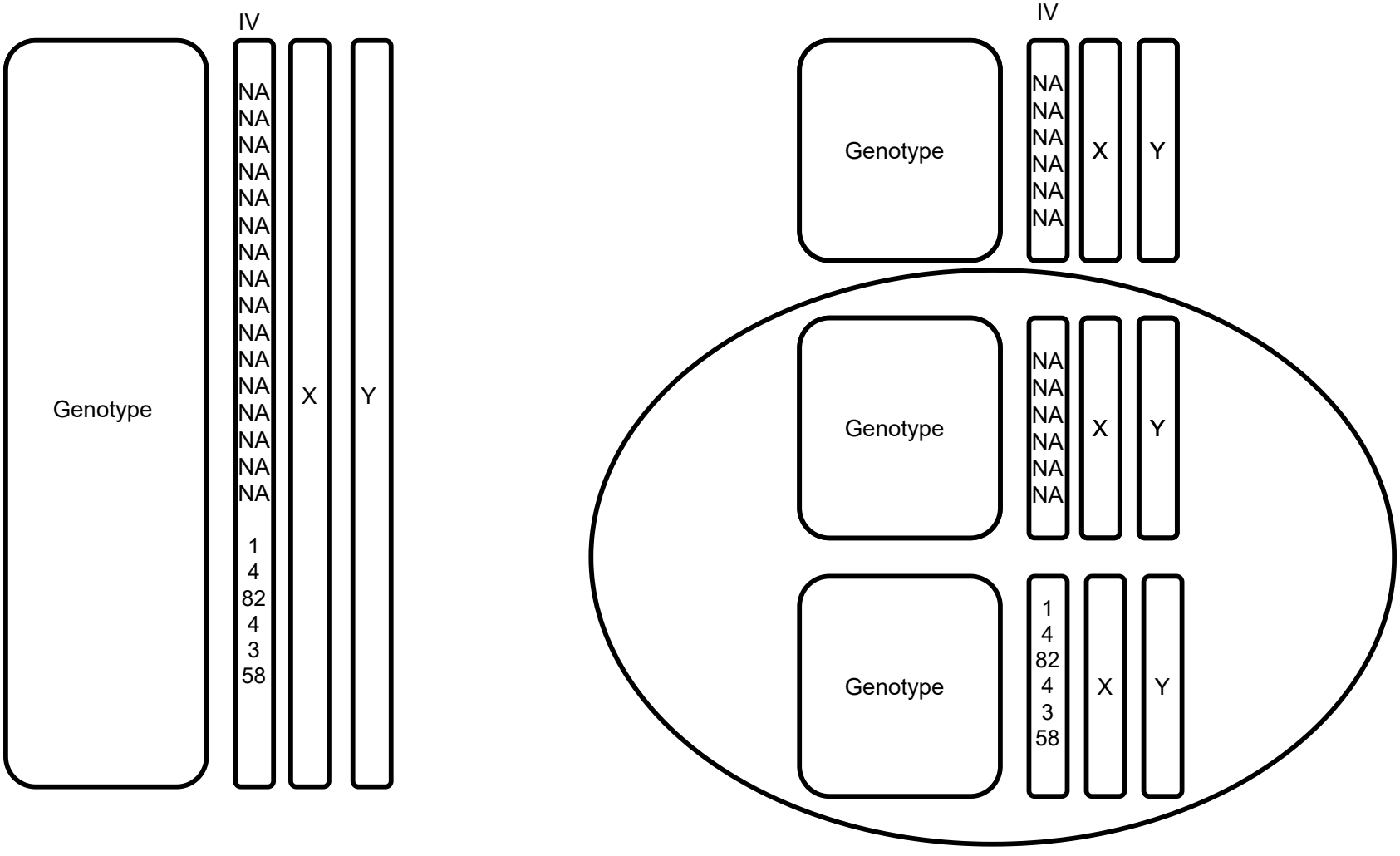


Write the IV vector

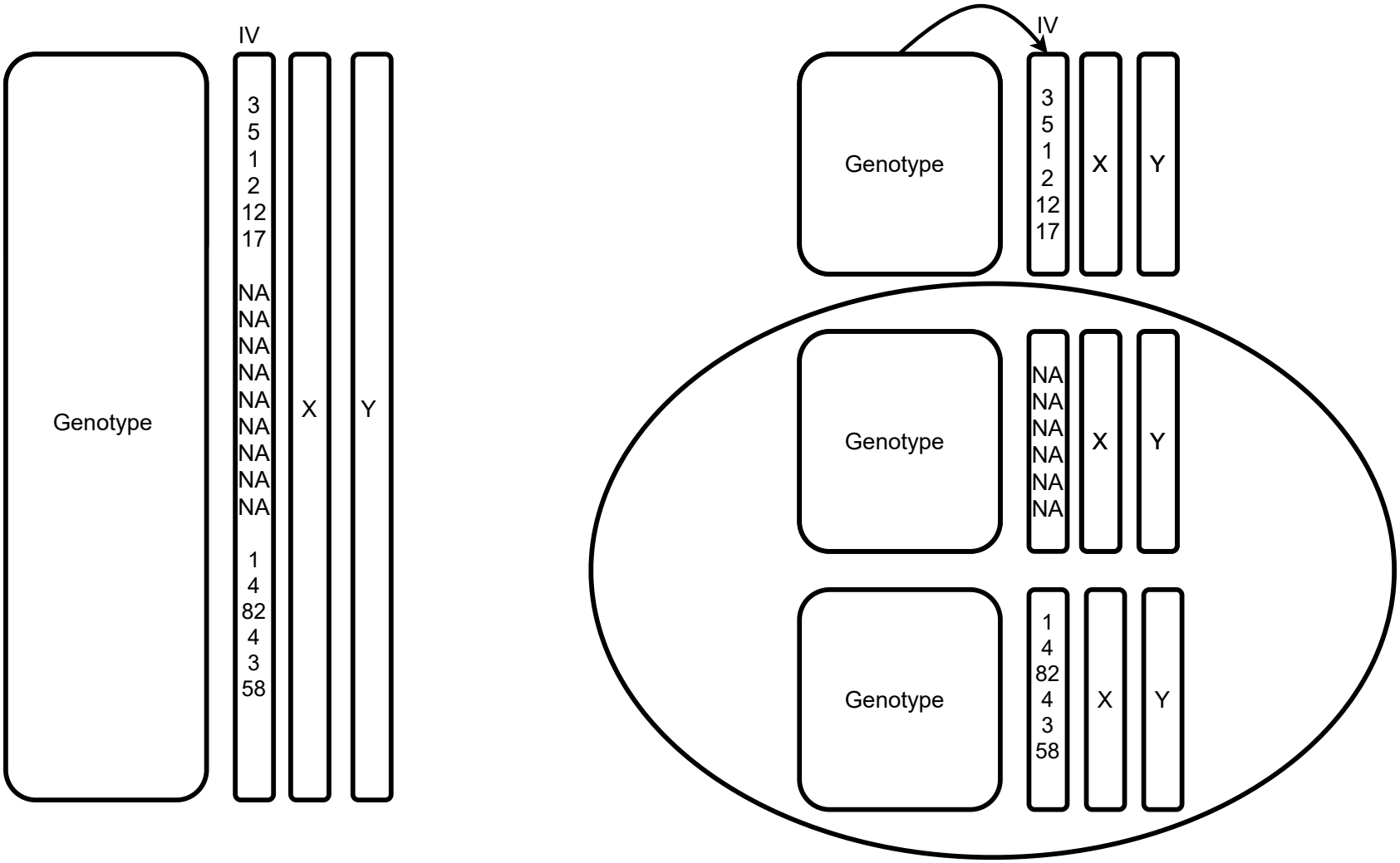




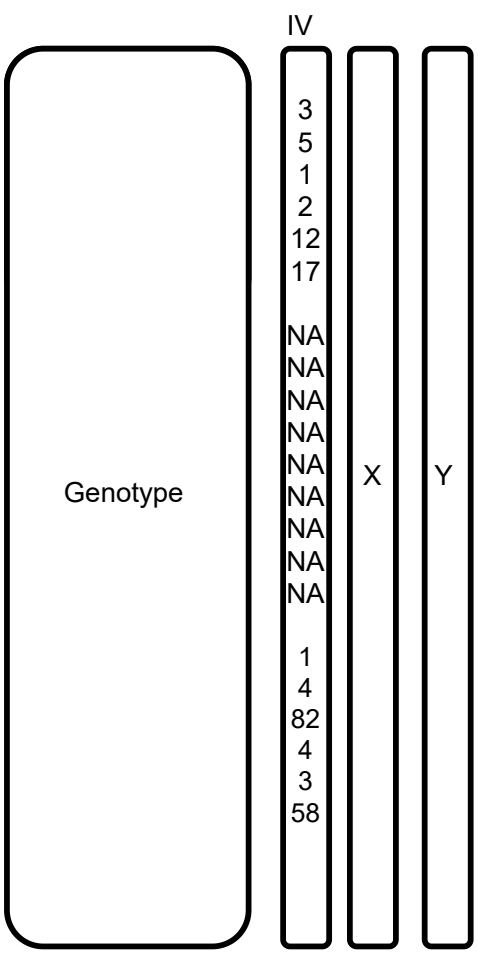
Select instruments using samples 2 and 3

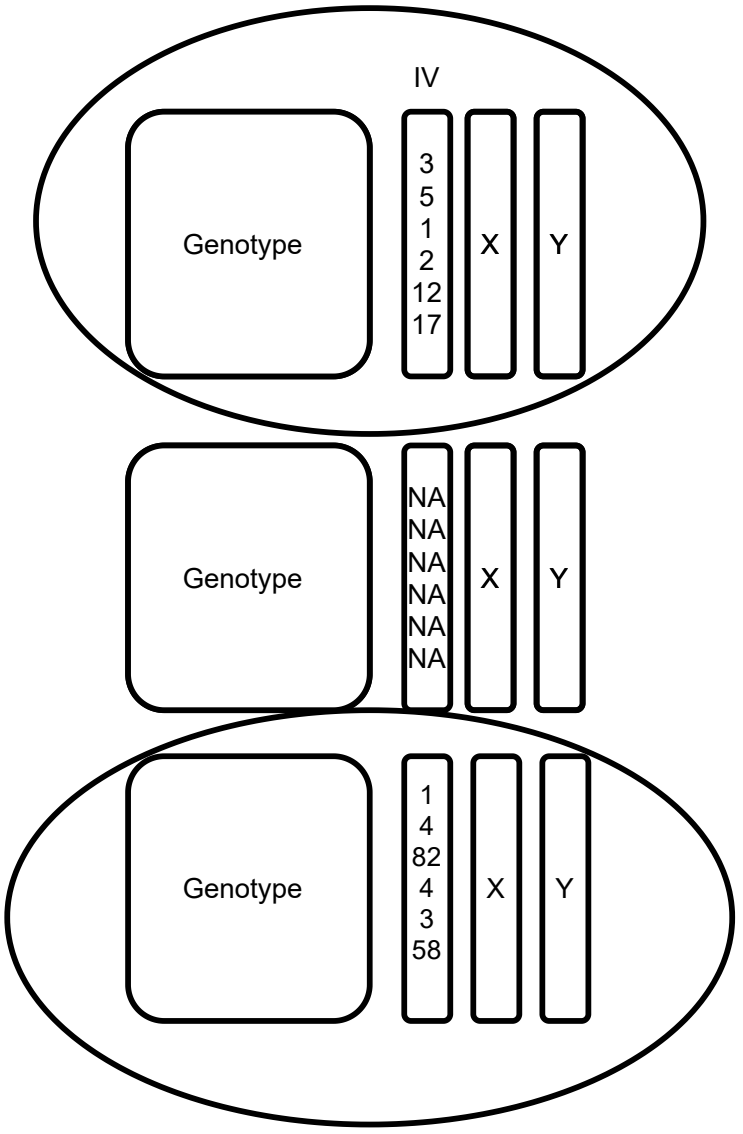


Predict X in sample 1 using estimates from samples 2 and 3

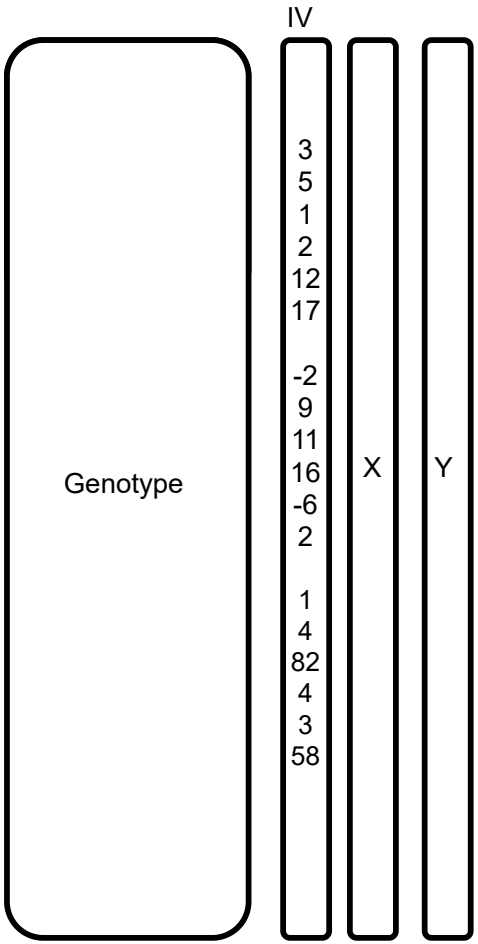


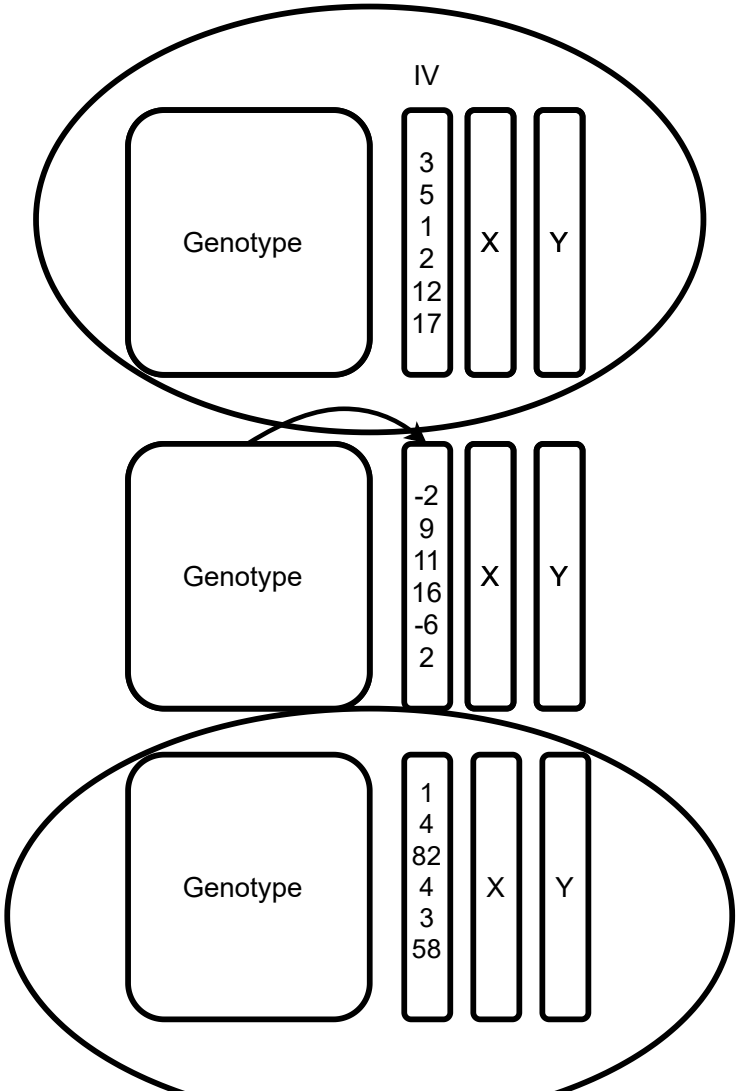
Select instruments using samples 1 and 3



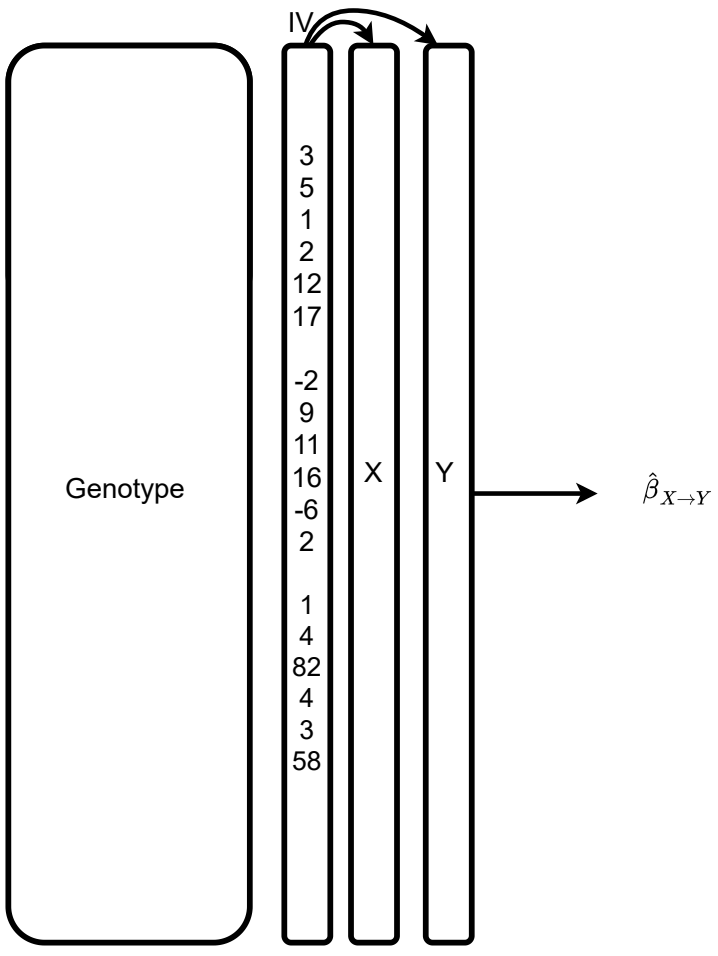


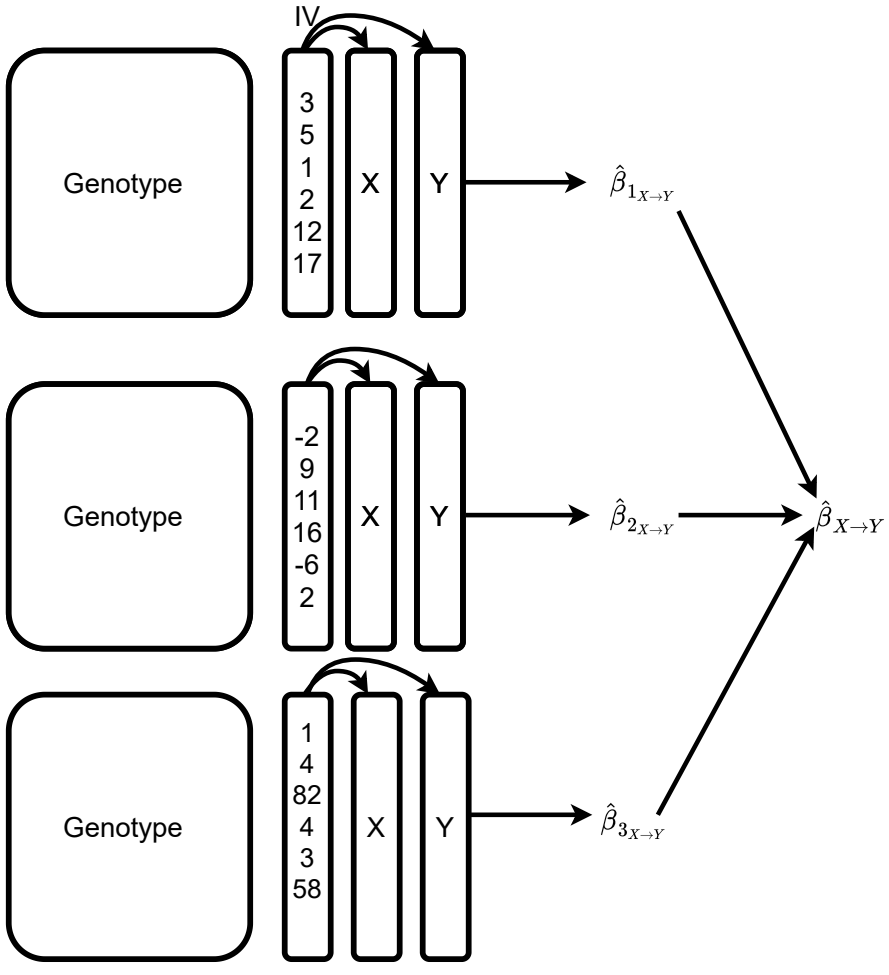
Predict X in sample 2 using estimates from samples 1 and 3





CFMR1 and CFMR2





Simulations and application

- Endogeneity bias
- power of CFMR vs two-sample MR
- Estimating the effect of pre-pregnancy maternal BMI on child birth weight

Bias in one sample MR

- We consider a set of 300 independent variants $(V_1, ..., V_{300})$
- $X = \sum_{l=1}^{5} \pi V_l + 40 \times h + v$
- $Y = 0.8X + h + u$
- h is a hidden confounder generated from a $N(0, 2)$ distribution, and v and u are two correlated error terms generated from a bivariate normal distribution.

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right]$$

We consider the following two scenarios:

1. where the variants explain 10% of the variance of X
2. where the variants explain 20% of the variance of X

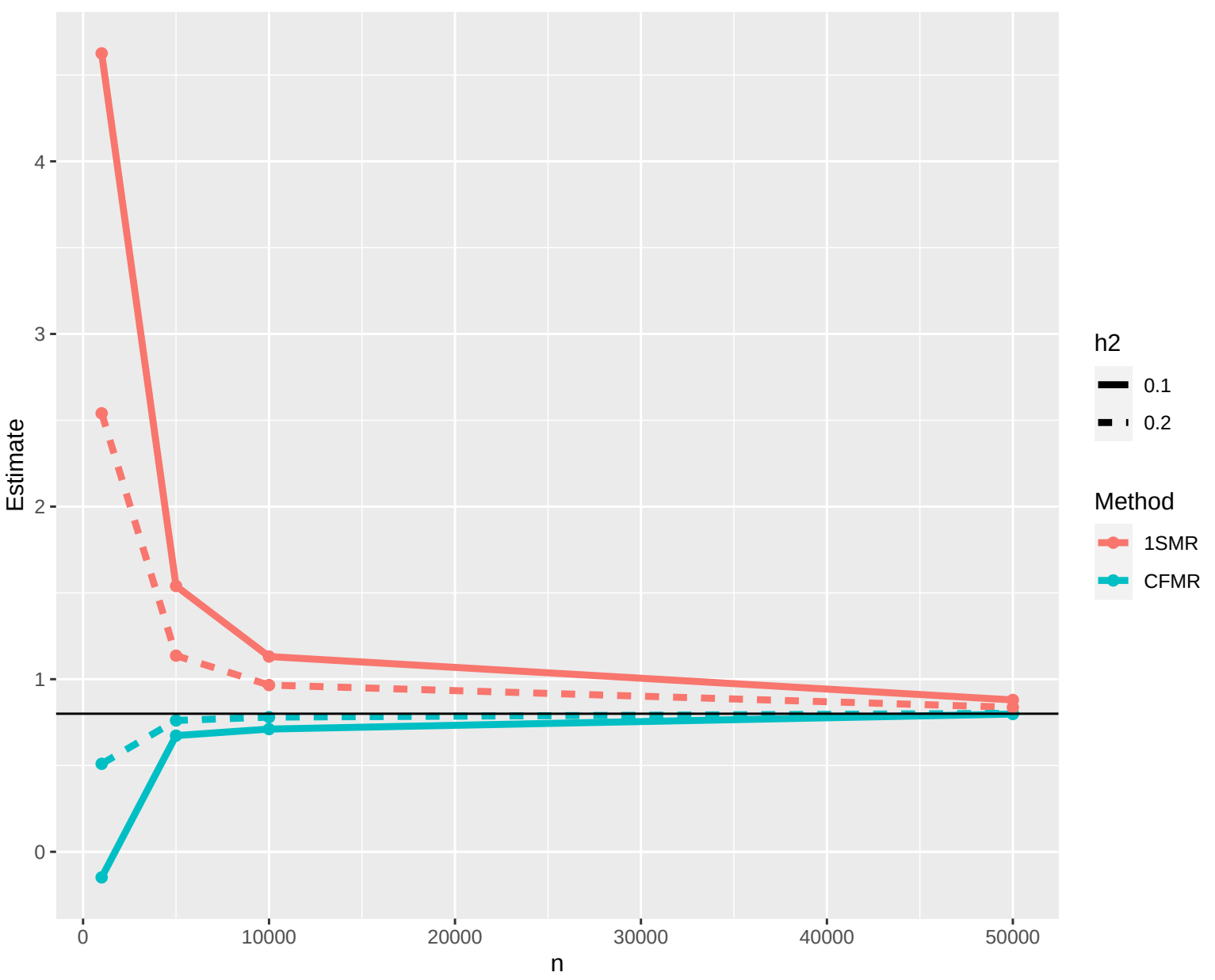
Bias in one sample MR

For each simulated dataset

- We applied 10-fold CFMR1 using a LASSO-based IV.
- We also build a predictor of X using LASSO on the entire dataset. We then used the prediction on the entire data as an instrument. We refer to 'one-sample MR estimates' when we estimate the effect of X on Y

Nota bene:

- In our manuscript we show that CFMR remains conservative even when using instrument that explain only 0.001% of the exposure variance.


Estimate
4
3
2
1
0
0
10000
20000
30000
40000
50000
n
h2
0.1
0.2
Method
1SMR
CFMR

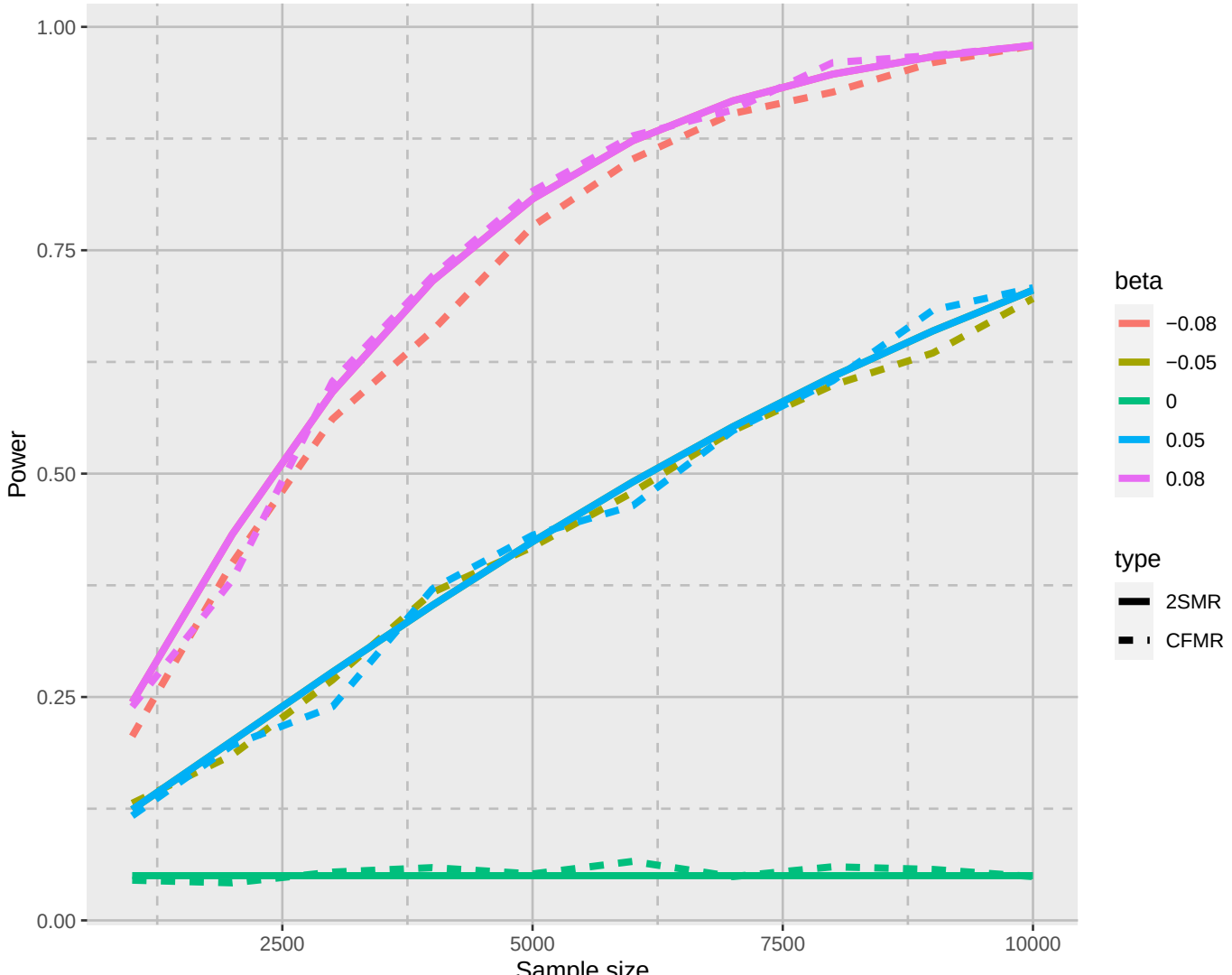
Power comparison

- We consider a set of 300 independent variants $(V_1, ..., V_{300})$
- $X = \sum_{l=1}^{5} \pi V_l + h + v$
- $Y = \theta_0 X + h + u$
- h is a hidden confounder generated from a $N(0, 2)$ distribution, and v and u are two correlated error terms generated from a bivariate normal distribution.

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right]$$

where the variants explain 20% of the variance of X Comparison with theoretical power of two-sample MR from Deng *et al.*, *Genetic Epidemiology*, 2020

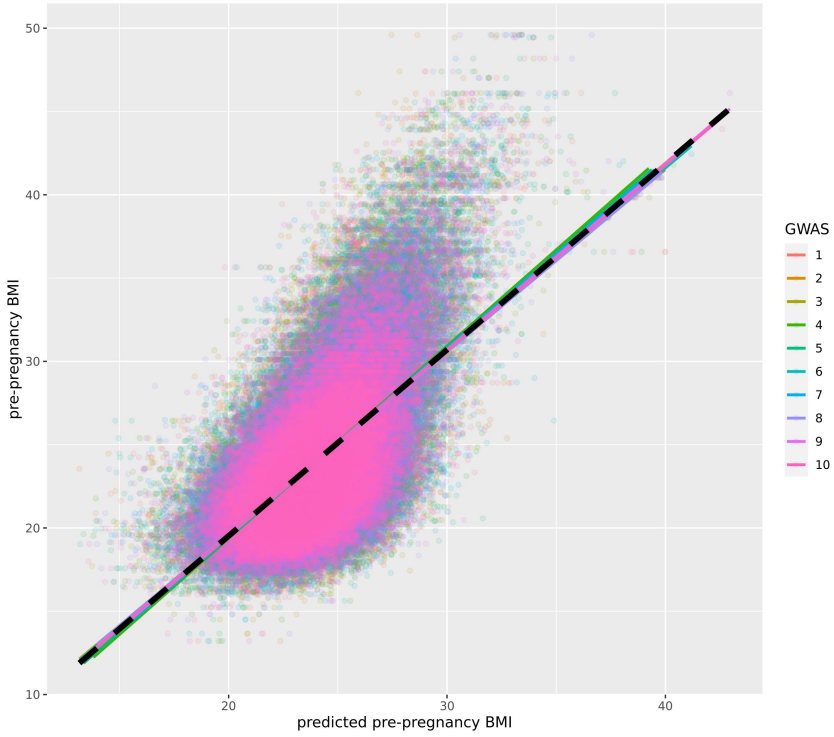
Power: thick lines from Deng *et al.*, *Genetic Epidemiology*, 2020

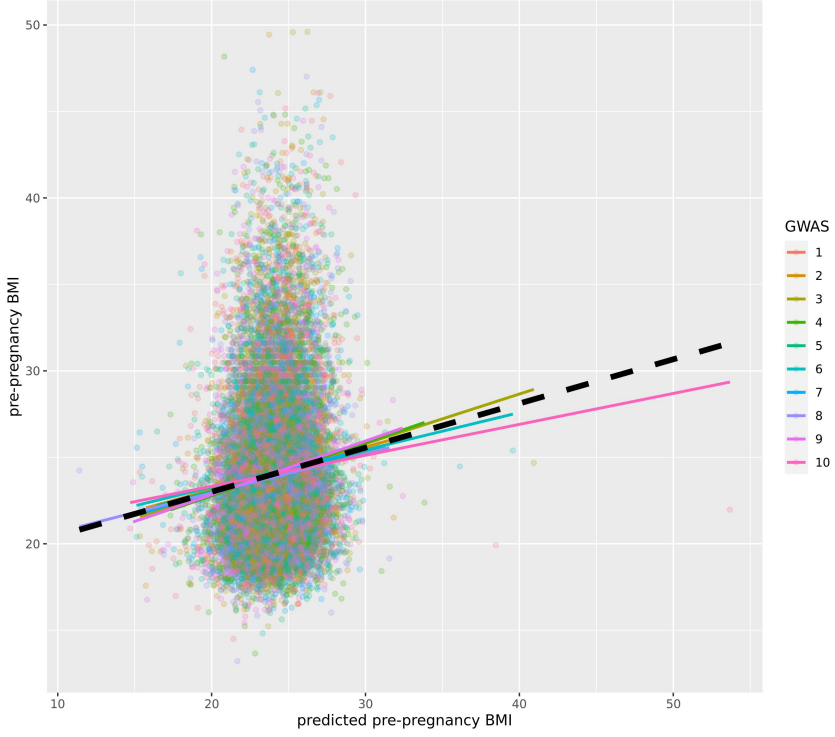


Estimating the effect of pre-pregnancy maternal BMI on childbirth weight

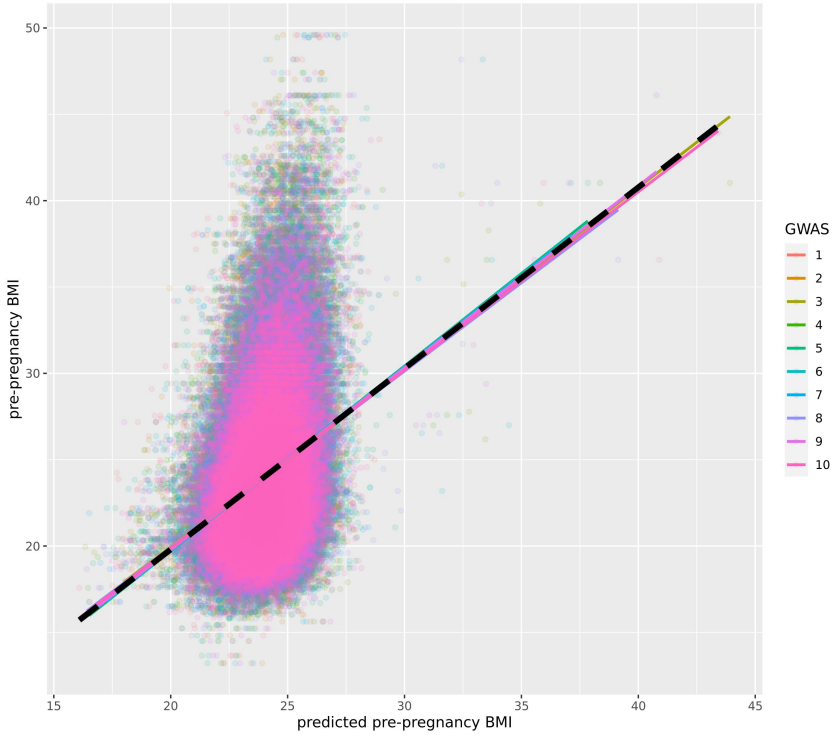
- We applied a 10 fold CFMR1 to a dataset comprising mother-child duos from the Norwegian Mother, father, and Child Cohort Study (MoBa), to re-examine the well-established effect of maternal pre-pregnancy BMI on offspring's birth weight (Tyrrel *et al., JAMA,* 2016).
- $26,896$ complete mother-child duos with genotype and phenotype data remained for the current analyses.
- 10 separate GWASes of pre-pregnancy BMI performed, with each GWAS encompassing $24,210$ randomly selected mothers.

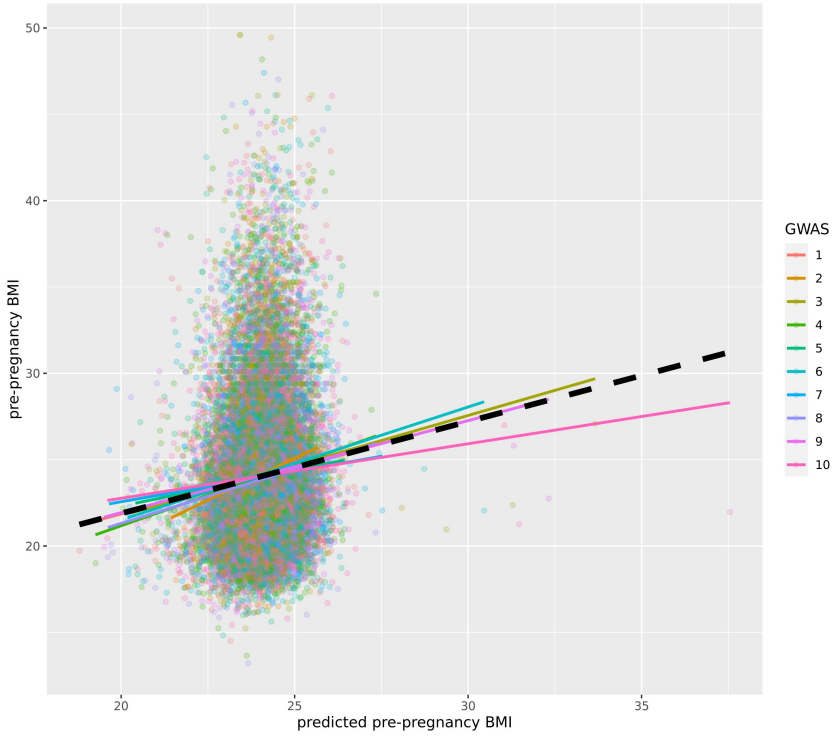
Polygenic score for maternal BMI (p-value 10^{-3})



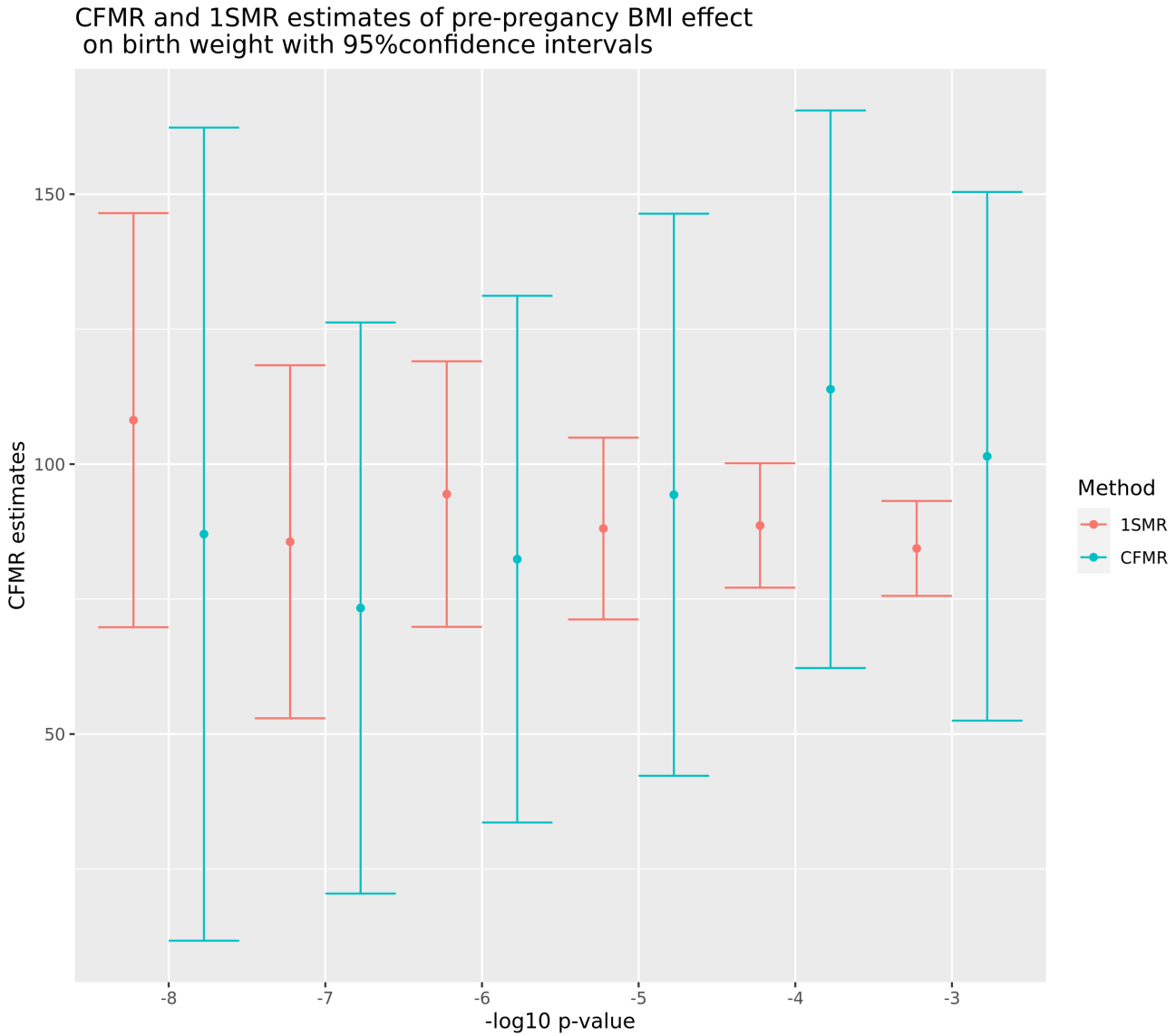


Polygenic score for maternal BMI (p-value 10^{-5})





CFMR estimates using different p-value threshold



Thank you for listening.

Joint work with:

- Jon Bohlin,
- Stephen Burgess,
- Christian Page,
- Astanand Jugessur

- Cross-fitted instrument: a blueprint for one-sample Mendelian Randomization, BiorXiv, 2021 (under review)
- https://github.com/william-denault/CFMR