# When GWAS gets too large for MR

*Gibran Hemani*

*2017-04-22*

One utility of genome wide association studies (GWAS) is in providing instruments and summary data to perform Mendelian randomization (MR), to make causal inference between phenotypes. As GWAS sample sizes continue to grow more significant genetic variants are discovered, and ostensibly this is advantageous for MR. With more variants, statistical power improves and causal associations can be more easily distinguished from reverse causation or horizontal pleiotropy.

In MR, valid instruments are required to influence the outcome only through the exposure. Several methods have been developed to minimise bias in MR when this is violated in one particular way: if the instrument influences the outcome through some pathway other than the exposure. This phenomenon is known as horizontal pleiotropy, and can arise if the genetic variant influences a confounder, which in turn influences both the exposure and the outcome.

Larger sample sizes might bring some problems to MR. First, as power improves, genetic variants that influence confounders are more likely to be identified as robust associations for an exposure. Second, if a researcher erroneously hypothesises that the outcome is causal for the exposure then SNPs that influence the exposure might be uncovered as significant associations for the outcome. If all significant hits for some trait are used as instruments in MR naively, this could include a mixture of valid instruments, SNPs that influence confounders, and SNPs that influence the hypothesised outcome (Figure 1).

Here we devise a novel framework to attempt to minimise this problem, and evaluate the performances of several recent methods for MR under the emerging reality of extremely large GWASs.

## Methods

### The model

Assume trait $X$ has a causal effect, $\beta_{xy}$, on trait $Y$. A trait that directly influences $X$ but not $Y$ is thought of as a precursor to $X$, $A$. Similarly, a trait that influences $Y$ directly but not $X$ is thought of as a precursor to $Y$, $B$. Finally, a trait that influences both $X$ and $Y$ is a confounder, $U$. Each of $X$, $Y$, $A$, $B$ and $U$ can have SNPs that influence them directly (Figure 1).

### Simulations

Simulations are performed to obtain GWAS summary data, to mimic the data used in two-sample Mendelian randomisation (2SMR).

**Variables**: A system is created such that $X$ has a causal influence on $Y$, there are five $U$ traits, five $A$ traits and five $B$ traits. Each trait has between five and fifty instruments.

**Effect sizes**: SNP effects are sampled from a normal distribution and the variance explained by all instruments on a trait is sampled uniformly between 1% and 10%. The effects of $A$ on $X$, $X$ on $Y$, $B$ on $Y$, $U$ on $X$ and $U$ on $Y$ are each sampled independently, with random signs and where the variance explained is sampled from a uniform distribution between 1$ and 10%. Null models are also generated where the effect of $X$ on $Y$ is 0.

**Samples**: Once effect sizes for all SNP-trait and trait-trait relationships have been sampled, 15 populations are generated, each with a sample size of between 20000 and 100000.

**Summary data**: Each population is assigned a trait, and GWAS is performed on that trait.

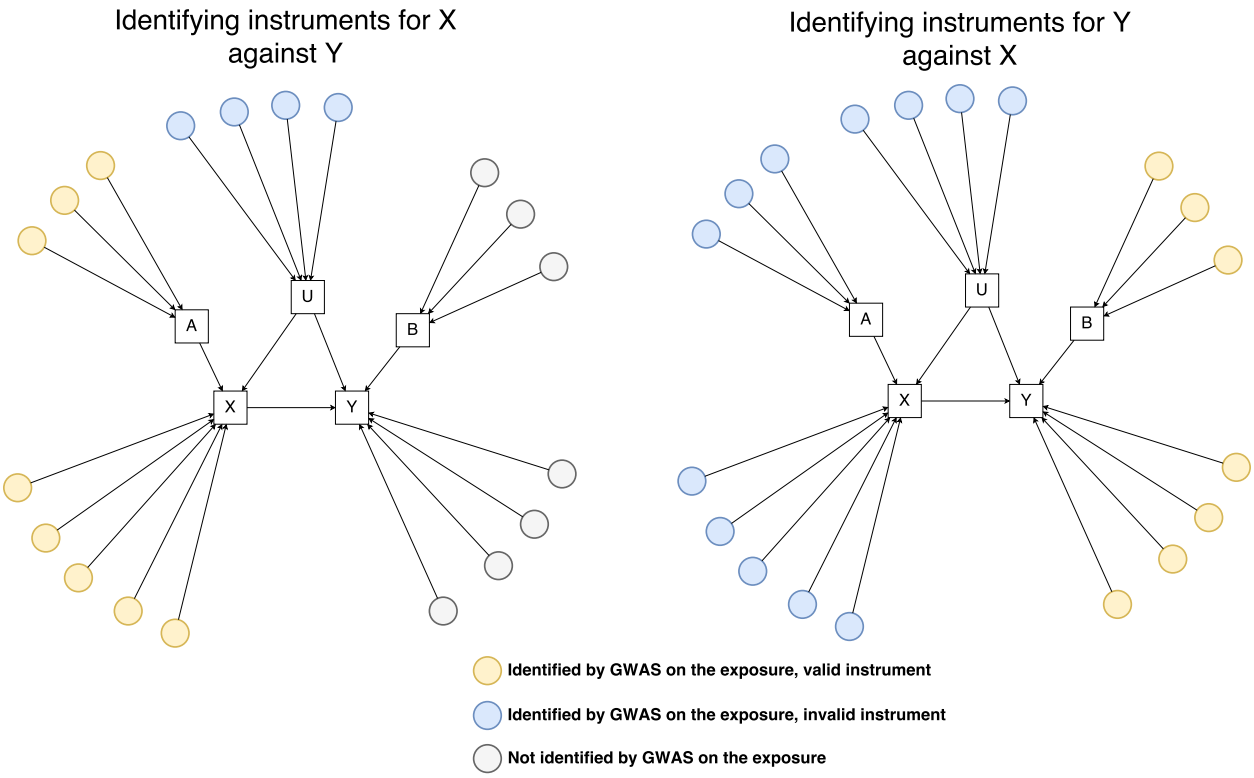



Figure 1: Figure 1

**Instrument selection**: MR analysis is performed for $X$ on $Y$ (the true causal direction) and for $Y$ on $X$, where the null hypothesis of no effect is true. There are three approaches used to select instruments for MR. First, the 'oracle' against which to benchmark methods that will be performed in practice. The oracle is obtained by selecting SNPs that have $p < 5 \times 10^{-8}$ and also are valid instruments (e.g. valid instruments for $X$ must influence $X$ directly, or influence $A$ directly). Second, the standard approach of identifying SNPs that associate with the putative exposure at $p < 5 \times 10^{-8}$. This could include invalid instruments. Third, the instruments identified by the standard approach are filtered to only retain those that pass the Steiger test.

**MR analysis**: A range of MR analyses are performed using instruments obtained from each of the three strategies. Additionally we record the number of invalid instruments that are being retained in the for the analysis.

### Identifying invalid instruments

Here we extend previous work that introduced the MR Steiger test. Assume an MR analysis is being performed to test the effect of $Y$ on $X$. The correct answer here is that there is no effect. When identifying SNPs that robustly associate with $Y$, using GWAS with threshold $p < 5 \times 10^{-8}$ as shown in Figure 1, GWAS on $Y$ is liable to returning SNPs that actually influence $X$, $U$ or $A$ directly. Our objective is to identify only valid SNPs - those that influence $Y$ or $B$ directly.

For each SNP $g$ identified for $Y$, we perform the Steiger test. This evaluates if $\rho^2_{gy} > \rho^2_{gx}$, where $\rho_{gy}$ is the correlation between the SNP and $Y$, and $\rho_{gx}$ is the correlation between the SNP and $X$. Only SNPs for which this inequality is satisfied to some nominal significance level under the Steiger test are retained as instruments for $Y$.

### Rucker framework

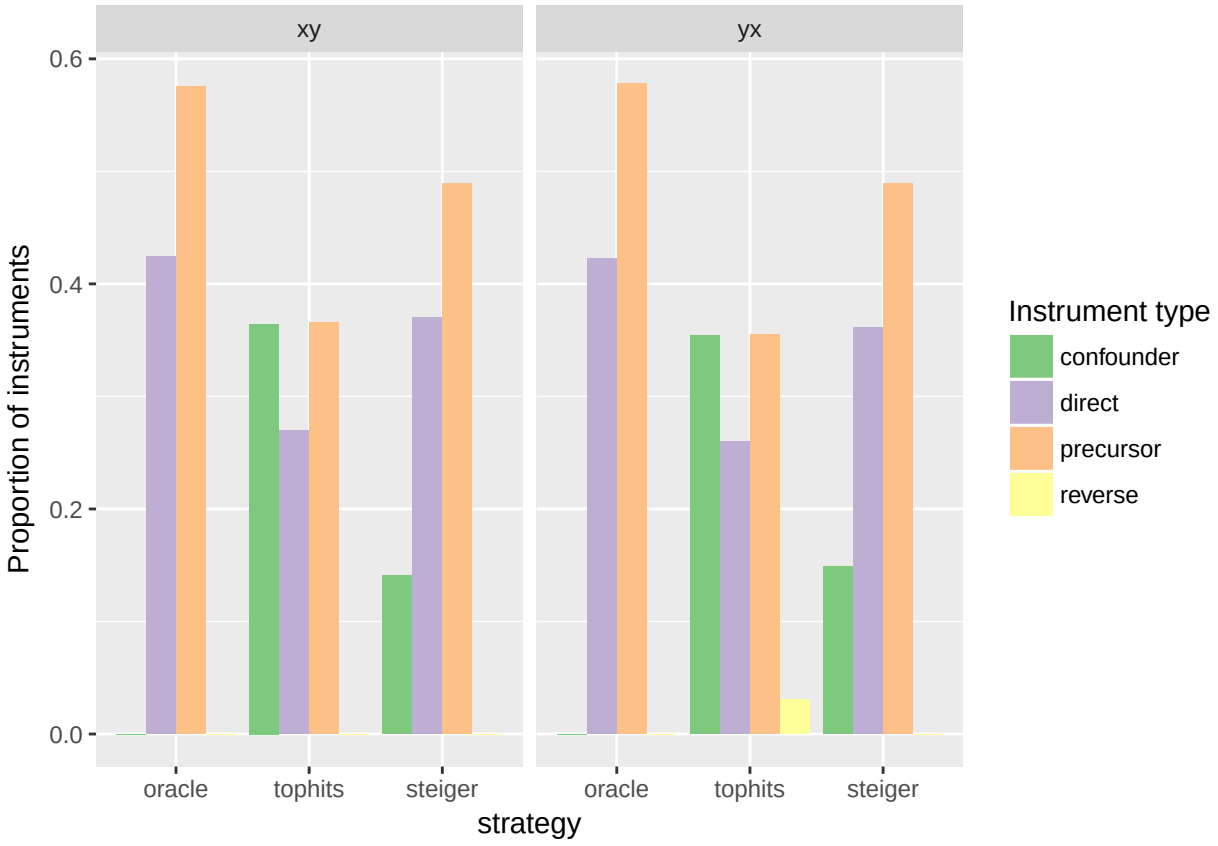
### Cook's distance

### Median approaches

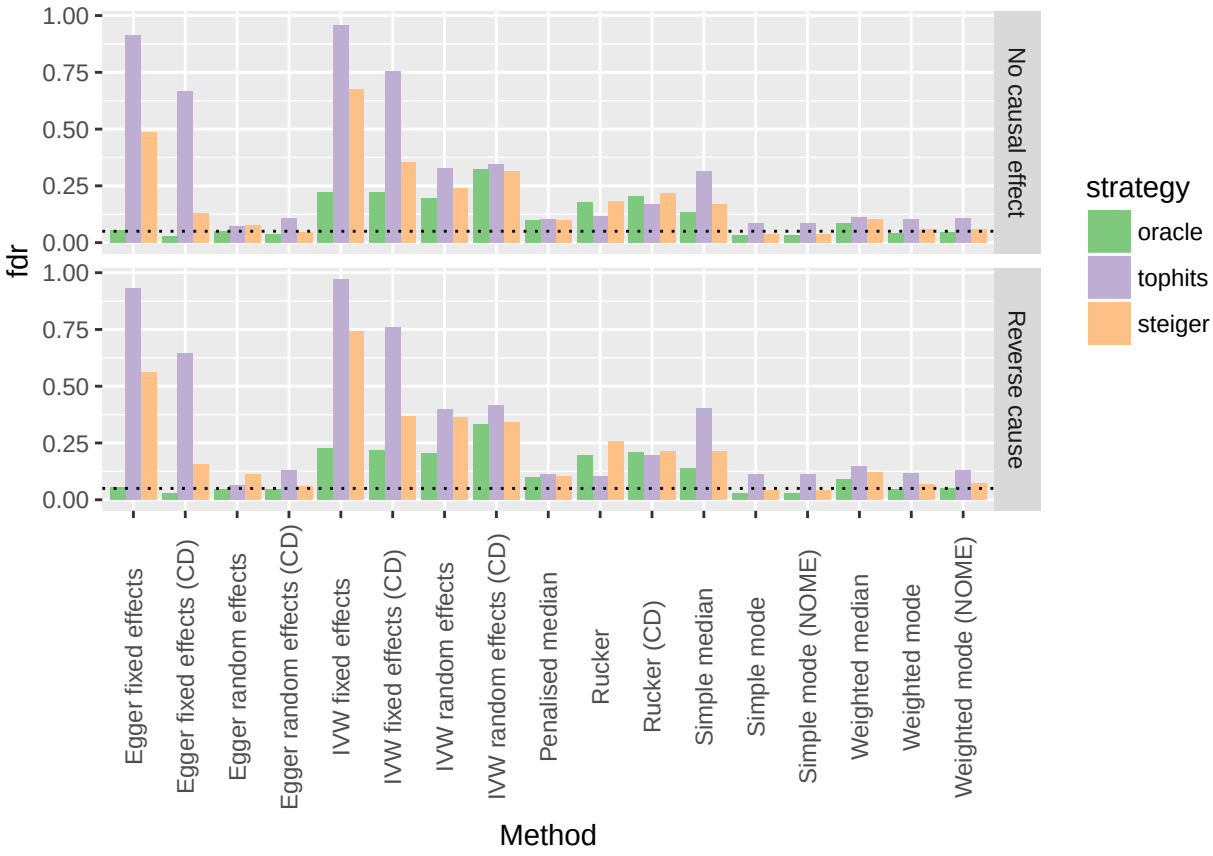
### Modal approaches

## Results

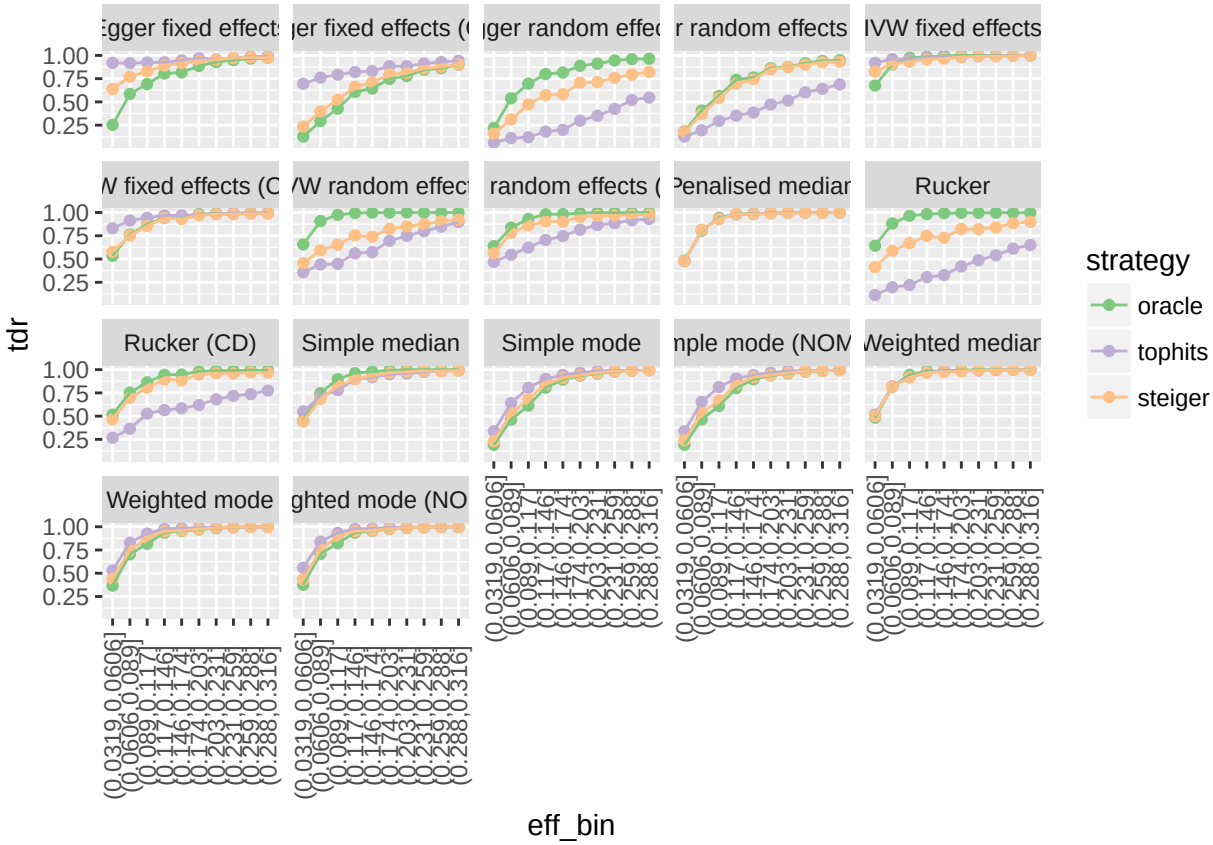
### Instrument selection



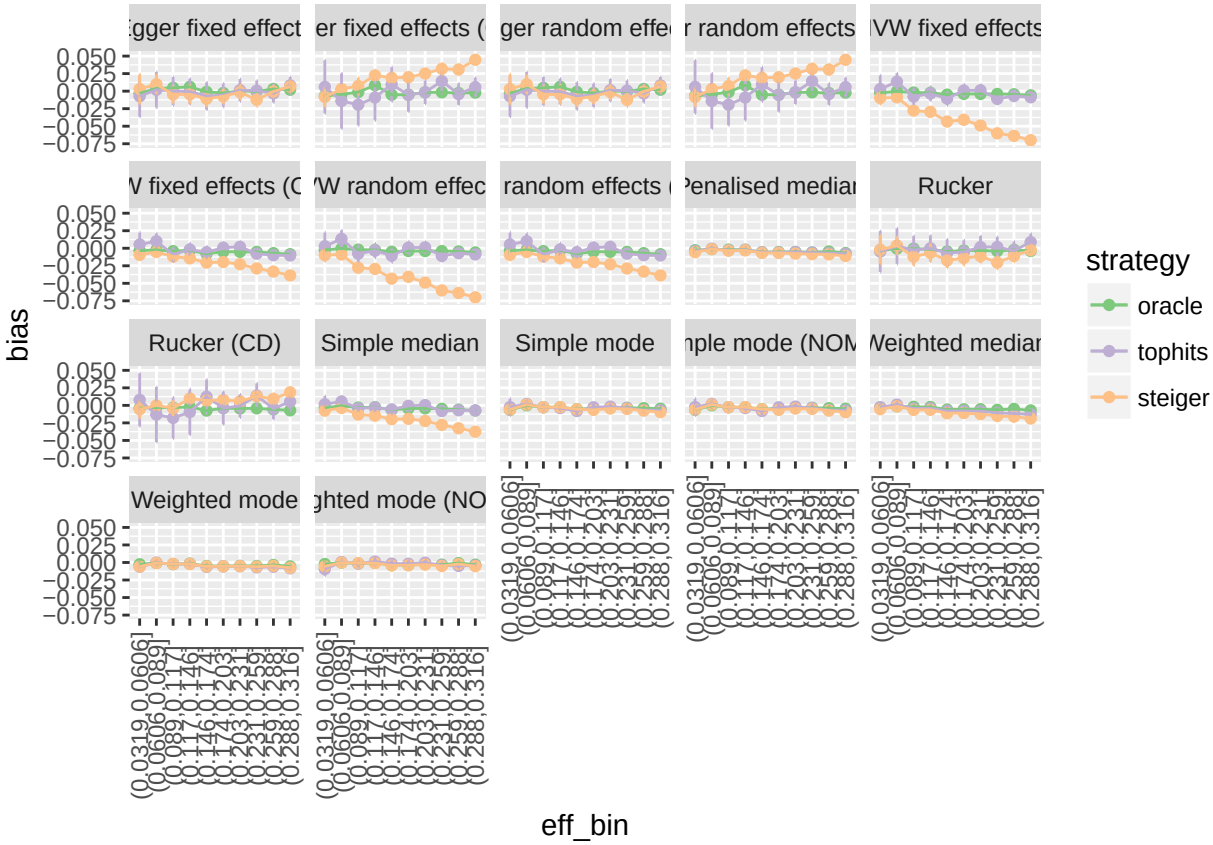
**False discovery rate**



**Power**


gger fixed effects
ger fixed effects (
ger random effec
r random effects
IVW fixed effects
W fixed effects (C
VW random effec
random effects (
Penalised median
Rucker
Rucker (CD)
Simple median
Simple mode
nple mode (NOM
Weighted median
Weighted mode
ghted mode (NO
tdr
1.00
0.75
0.50
0.25
(0.0319,0.0606]
(0.0606,0.089]
(0.089,0.117]
(0.117,0.146]
(0.146,0.174]
(0.174,0.203]
(0.203,0.231]
(0.231,0.259]
(0.259,0.288]
(0.288,0.316]
eff_bin
strategy
oracle
tophits
steiger

**Bias**


bias
eff_bin
strategy
oracle
tophits
steiger
Rucker
Rucker (CD)
Simple median
Simple mode
Weighted mode
0.050
0.025
0.000
−0.025
−0.050
−0.075