Correction for sample overlap, winner's curse and weak instrument bias in two-sample Mendelian Randomization

Ninon Mounier[1,2] **Zoltán Kutalik**[1,2,3,*]

Abstract

Inverse-variance weighted two-sample Mendelian Randomization (IVW-MR) is the most widely used approach that uses genome-wide association studies summary statistics to infer the existence and strength of the causal effect between an exposure and an outcome. Estimates from this approach can be subject to different biases due to: (i) the overlap between the exposure and outcome samples; (ii) the use of weak instruments and winner's curse. We developed a method that aims at tackling all these biases together. Assuming spike-and-slab genomic architecture and leveraging LD-score regression and other techniques, we could analytically derive and reliably estimate the bias of IVW-MR using association summary statistics only. This allowed us to apply a bias correction to IVW-MR estimates, which we tested using simulated data for a wide range of realistic scenarios. In all the explored scenarios, our correction reduced the bias, in some situations by as much as 30 folds. When applied to real data on obesity-related exposures, we observed significant differences between IVW-based and corrected effects, both for non-overlapping and fully overlapping samples. While most studies are extremely careful to avoid any sample overlap when performing two-sample MR analysis, we have demonstrated that the incurred bias is much less substantial than the one due to weak instruments or winner's curse, which are often ignored.

[1]University Center for Primary Care and Public Health, University of Lausanne, 1010, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland

[3]Department of Computational Biology, University of Lausanne, Lausanne, 1015, Switzerland

[*]Correspondence should be addressed to zoltan.kutalik@unil.ch

1 Introduction

Mendelian Randomization (MR) is a method that uses genetic variants (typically single nucleotide polymorphisms, SNPs) as instrumental variables (IVs) to infer the existence and strength of the causal effect between an exposure and an outcome[1]. In particular, two-sample summary data MR[2], that requires solely genome-wide association study (GWAS) summary statistics has become increasingly popular. The reason for this is that in the last decade, GWASs have drastically increased in sample size[3]. The resulting summary statistics are often publicly available allowing not only the identification of genetic variants independently associated with a particular exposure, i.e. IVs, but also the look up of the variants' effects on the outcome trait in different samples. Each IV provide an independent estimate for the causal effect and these estimates can then be combined using an inverse variance-weighting (IVW) approach[2]. MR relies on three main assumptions (Figure 1): (1) Relevance – IVs must be robustly associated with the exposure. (2) Exchangeability – IVs must not be associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – IVs must be independent of the outcome conditional on the exposure and all confounders of the exposure-outcome relationship (i.e. the only path between the IVs and the outcome is via the exposure).

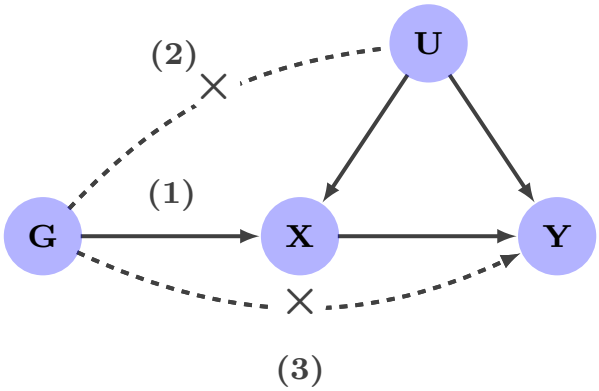

Figure 1: Main assumptions of Mendelian randomization. (1) Relevance – IVs, denoted by G, are strongly associated with the exposure. (2) Exchangeability – G is not associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – G is independent of the outcome conditional on the exposure and all confounders of the exposure-outcome relationship (i.e. the only path between the IVs and the outcome is via the exposure).

When the exposure and the outcome are coming from overlapping samples, environmental confounding factor(s) can bias the causal effect estimate toward the observational correlation, which includes the correlation induced by the environmental confounder(s)[4]. Hence, sample overlap will introduce an association between the IVs and the confounder, resulting in a violation of the second assumption of MR. Nowadays, large consortium summary statistics are often used for both the exposure and the outcome, and it is not always possible to ensure that the two samples used are not overlapping. The need for non-overlapping samples does not only substantially reduce the sample size used to generate the summary statistics, but in the era of large biobanks (such as UK-Biobank[5], UKBB) it is becoming increasingly difficult to find recent meta-analysis

datasets for an exposure and outcome without overlapping samples.

Sample overlap is not the only source of bias that needs to be considered. Two-sample MR estimates obtained from non-overlapping samples are known to be subject to weak instrument bias (toward the null). Most MR methods assume that SNP-exposure effects are measured without noise (NO Measurement Error, NOME assumption). This simplification leads to regression-dilution bias, which is particularly strong for weak instruments. For this reason, the bias introduced by the NOME assumption is referred to as weak instrument bias and it becomes more and more severe as the average variance of the exposure explained by the IVs decreases[6]. When combined with sample overlap, the effect of weak instrument bias will move toward the observational correlation[7,8]. Using IVs strongly associated with the exposure and/or increasing the sample size can mitigate weak instrument bias[7]. Although the exact multiplicative bias due to the NOME assumption can be expressed analytically (proportional to the inverse of the F-statistic), the estimator for the multiplicative constant has typically high variance and works poorly in practice[9]. The simulation-extrapolation based SIMEX method proved to yield more robust corrections both for IVW[10] and MR-Egger estimates[9].

In addition, MR estimates are subject to winner's curse bias, which occurs when the same sample is used to select IVs and estimate their effect on the exposure. In such case, the observed IV effect on the exposure is not an unbiased estimator for its true effect and is likely to be overestimated (in absolute value). This would affect the causal effect estimate (underestimation in non-overlapping sample and overestimation in fully overlapping samples)[8]. Using a third independent sample to select instruments, and therefore avoid winner's curse, is not always possible. Based on the expectation of truncated normal distribution[11], a correction can be applied for the SNP-exposure effect sizes. However, the additional estimator variance such correction entails can outweigh the benefit of the reduced bias, which can be mitigated by directly maximising the conditional likelihood[12]. Still, all these methods account for winner's curse for a single SNP, but do not model the winner's curse bias induced by the IV selection process from millions of potential markers, which is far more complex.

While previous approaches aimed at tackling one bias at a time, the intricate way these different sources of biases interplay with each other remains poorly understood, and there is currently no method that simultaneously tackles them.

In this paper, we will first introduce a two-sample MR framework that takes into account these three sources of bias simultaneously in order to obtain a corrected causal effect estimate. We will then test our approach and compare the proposed correction of the IVW-MR causal effect estimate against its uncorrected counterpart using a wide range of simulation settings. Finally, to demonstrate its utility, we will apply our approach to obesity-related traits using UKBB data.

2 Methods

2.1 Expectation of the causal effect estimate

Let X and Y denote two random variables representing two complex traits. We intend to use MR to estimate the causal effect of X on Y. We will use m Linkage Disequilibrium (LD)-independent genetic variants as IVs. Their corresponding genotype data is denoted by G and its jth column by g_j (rows representing individuals). To simplify notation we assume that $E(X) = E(Y) = E(G) = 0$ and $Var(X) = Var(Y) = Var(G) = 1$. Let us assume that X is observed in sample A of sample size n_A, Y is measured in sample B of sample size n_B with an overlap of $n_{A\cap B}$ individuals between the two samples. The vector of realisations of Z^C is denoted by $\boldsymbol{z}^C$ for all variables ($Z = X, Y, G, g$) and samples ($C = A, B, A \cap B$). Let us assume the following models:

$$\begin{aligned} \boldsymbol{x}^A &= G^A \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^A \\ \boldsymbol{x}^B &= G^B \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^B \\ \boldsymbol{y}^B &= \alpha \cdot \boldsymbol{x}^B + G^B \cdot \boldsymbol{\gamma_y} + \boldsymbol{\epsilon}_{\boldsymbol{y}}^B \end{aligned} \tag{1}$$

where $\boldsymbol{\gamma_x}$ are the effect sizes of the instruments on X, $\boldsymbol{\gamma_y}$ are their pleiotropic effects on Y. Assuming that there is a single environmental confounder U acting linearly on both traits (as used for simulations) the error term can split into two parts : $\boldsymbol{\epsilon}_{\boldsymbol{x}}^C = \kappa_x \cdot \boldsymbol{u}^C + \boldsymbol{\varepsilon}^C$ and $\boldsymbol{\epsilon}_{\boldsymbol{y}}^C = \kappa_y \cdot \boldsymbol{u}^C + \boldsymbol{\varepsilon}^C$, where κ_x and κ_y refer to the effect of U on X and Y respectively, ε^C is independent of the confounder and C can take the values A, B or $A \cap B$ as above.
Under the INSIDE assumption[13] (INstrument Strength Independent of Direct Effect, i.e. horizontal pleiotropic effects are independent of the direct effect), $cov(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}) = 0$ and $E[\boldsymbol{\gamma_y}] = 0$. We denote $\rho := cov(\boldsymbol{\epsilon_x}, \boldsymbol{\epsilon_y}) = \kappa_x \cdot \kappa_y$. It corresponds to the part of the phenotypic correlation (r) due to a (non-genetic) confounder ($r = \rho + \alpha$). The univariate effect size estimates from GWASs for genetic variant j are as follows:

$$\begin{aligned} (\widehat{\beta}_x^A)_j &= \frac{1}{n_A} \cdot (\boldsymbol{g}_j^A)' \cdot \boldsymbol{x}^A = \frac{1}{n_A} \cdot (\boldsymbol{g}_j^A)' \cdot \left(G^A \cdot \boldsymbol{\gamma_x} + \boldsymbol{\epsilon}_{\boldsymbol{x}}^A\right) \\ (\widehat{\beta}_y^B)_j &= \frac{1}{n_B} \cdot (\boldsymbol{g}_j^B)' \cdot \boldsymbol{y}^B = \frac{1}{n_B} \cdot (\boldsymbol{g}_j^B)' \cdot \left(\alpha \cdot \boldsymbol{x}^B + G^B \cdot \boldsymbol{\gamma_y} + \boldsymbol{\epsilon}_{\boldsymbol{y}}^B\right) \end{aligned} \tag{2}$$

where the genotype data for genetic variant j for individuals in sample A is denoted by $\boldsymbol{g}_j^A$.
Let us now consider the fixed-effect inverse-variance weighting meta-analysis for the ratio estimates for the causal effect α. Each IV j provides a ratio estimate:

$$\widehat{\alpha}_j = \frac{(\widehat{\beta}_y^B)_j}{(\widehat{\beta}_x^A)_j} \tag{3}$$

$$Var(\widehat{\alpha}_j) = \frac{Var((\widehat{\beta}_y^B)_j)}{(\widehat{\beta}_x^A)_j^2} = \frac{(1 - \alpha^2 - \gamma_{y_j}^2)}{n_B \cdot (\widehat{\beta}_x^A)_j^2} \tag{4}$$

Hence the weights (w_j) of IV j for estimating the IVW causal effect are of the form:

$$w_j = \frac{1}{Var(\widehat{\alpha}_j)} = \frac{n_B \cdot (\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)} \tag{5}$$

Finally, the estimate can be written in the following form:

$$\begin{aligned}\widehat{\alpha}_{IVW} &= \frac{\sum_{j=1}^m \widehat{\alpha}_j \cdot w_j}{\sum_{j=1}^m w_j} = \frac{\sum_{j=1}^m \frac{(\widehat{\beta}_y^B)_j}{(\widehat{\beta}_x^A)_j} \times \left(\frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)}\right)}{\sum_{j=1}^m \frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)}} = \frac{\sum_{j=1}^m \frac{(\widehat{\beta}_y^B)_j \cdot (\widehat{\beta}_x^A)_j}{(1-\alpha^2-\gamma_{y_j}^2)}}{\sum_{j=1}^m \frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2-\gamma_{y_j}^2)}} \\ &\approx \frac{\sum_{j=1}^m \frac{(\widehat{\beta}_y^B)_j \cdot (\widehat{\beta}_x^A)_j}{(1-\alpha^2)}}{\sum_{j=1}^m \frac{(\widehat{\beta}_x^A)_j^2}{(1-\alpha^2)}} = \frac{\sum_{j=1}^m (\widehat{\beta}_y^B)_j \cdot (\widehat{\beta}_x^A)_j}{\sum_{j=1}^m (\widehat{\beta}_x^A)_j^2}\end{aligned} \tag{6}$$

Here, the last approximation is based on the realistic assumption that the individual pleiotropic effect of each SNP is very small. To account for winner's Curse, we need to consider the probability of being selected for each genetic variant. Let us consider a threshold T ($T = -\Phi^{-1}(p/2) \approx 5.45$ for $p = 5 \times 10^{-8}$, genome-wide significance threshold, for example) and use only IVs with $|(\widehat{\beta}_x^A)_j| \cdot \sqrt{n_A} > T$. By denoting $S_j := \left\{|(\widehat{\beta}_x^A)_j| \cdot \sqrt{n_A} > T\right\}$, the causal effect estimate (Equation 6) changes to:

$$\widehat{\alpha}_{IVW} \approx \frac{\sum_{j=1}^M ((\widehat{\beta}_y^B)_j \,|S_j\,) \cdot ((\widehat{\beta}_x^A)_j \,|S_j\,) \cdot Pr(S_j)}{\sum_{j=1}^M ((\widehat{\beta}_x^A)_j \,|S_j\,)^2 \cdot Pr(S_j)} \tag{7}$$

Note that while m denoted the number of IVs, M represents the number of genome-wide variants from which IVs are selected. By approximating the expectation of a ratio by the ratio of expectations, the expectation of the causal effect estimate (Equation 7) can be written as:

$$\begin{aligned}E[\widehat{\alpha}_{IVW}] &\approx \frac{\sum_{j=1}^M E\left[((\widehat{\beta}_y^B)_j \,|S_j\,) \cdot ((\widehat{\beta}_x^A)_j \,|S_j\,)\right] \cdot Pr(S_j)}{\sum_{j=1}^M E\left[((\widehat{\beta}_x^A)_j \,|S_j\,)^2\right] \cdot Pr(S_j)} \\ &= \frac{\sum_{j=1}^M s_j \cdot Pr(S_j)}{\sum_{j=1}^M t_j \cdot Pr(S_j)}\end{aligned} \tag{8}$$

The values of s_j, t_j and $Pr(S_j)$ can be analytically derived (presented in Supplementary Section A), and we show that Equation 8 translates to:

$$\begin{aligned}E[\hat{\alpha}_{IVW}] \approx \quad & \alpha \cdot \frac{\left(\pi_x \cdot \sigma_x^2\right) \cdot \left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\right)}{d(n_A, T, \pi_x, \sigma_x^2)} \\ + \quad & \lambda' \cdot \frac{\pi_x \cdot \left(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\right) + \left(1 - \pi_x\right) \cdot 2 \cdot c}{d(n_A, T, \pi_x, \sigma_x^2)}\end{aligned} \tag{9}$$

where π_x and σ_x^2 are characteristics of the genetic architecture of trait X (respectively, a measure of the polygenicity and the per-variant heritability, see Equation S35) and λ' is a quantity closely related to the cross-trait LD score regression (LDSC) intercept[14] (λ) : $\lambda' = \frac{\lambda}{\sqrt{n_A \cdot n_B}}$.

The constants a, b and c do not depend on the causal effect α nor the sample overlap since $n_{A\cap B}$ is only affecting λ' (see Equations S33 and S51). The same is true for the denominator $(d(n_A, T, \pi_x, \sigma_x^2))$ (see Equation S52).

We can see that in the absence of sample overlap, the second term is equal to 0, so the first term corresponds to the causal effect estimated when only weak instruments and winner's curse are biasing the estimate. In this case, $E[\hat{\alpha}_{IVW}]$ is lower than α and the observed effect will be biased toward the null (Figure S1). The only parameters affecting the bias are π_x and h_x^2, n_A and T. When π_x is smaller, or h_x^2 is larger, then IVs have stronger effects, leading to a smaller bias when all other parameters are kept constant. As expected, since these are commonly used approaches to limit weak instrument bias, using a more stringent threshold and increasing the exposure sample size both reduce the bias.
When there is sample overlap (with % overlap defined as $\frac{n_{A\cap B}}{\sqrt{n_A \cdot n_B}}$), the expression of $E[\hat{\alpha}_{IVW}]$ is more complex. The magnitude of the bias will not only depend on the parameters described for non-overlapping samples (Figures S2, S3 and S4), but also on the confounder's effect (ρ) that can affect both the magnitude and the direction of the bias. For example, when the percentage of overlap is relatively low (20%) and ρ has the same sign as α, then the estimate will be further biased toward the null, whereas a confounder acting in the opposite direction (as the causal effect) will reduce the bias (Figure S2). When the percentage of overlap increases (Figures S3 and S4), the effect of sample overlap gets stronger, and the bias direction directly depends on ρ. We can see that there exists a value of ρ (that depends on all the other parameters) for which the two opposing forces (weak instruments, sample overlap) contributing to the biases cancel out.

All parameters except α are known or can be estimated from the data. We assume that sample sizes for X and Y (respectively n_A and n_B) are known, as well as the threshold used to select IVs (T). Parameter λ' can be estimated from cross-trait LDSC[14]. π_x and σ_x^2 are estimated by matching the denominator of this formula $d(n_A, T, \pi_x, \sigma_x^2)$ to the denominator of equation 8 (see Supplementary Section B for details).
From Equation 10, we can derive a corrected effect for the causal effect :

$$\hat{\alpha}_c = \frac{\hat{\alpha}_{IVW} \cdot d(n_A, T, \pi_x, \sigma_x^2) - \lambda' \cdot (\pi_x \cdot \big(2 \cdot a + b \cdot (1 + \sigma_x^2 \cdot n_A)\big) + \big(1 - \pi_x\big) \cdot 2 \cdot c)}{\big(\pi_x \cdot \sigma_x^2\big) \cdot \big(2 \cdot a + b \cdot (1 + n_A \cdot \sigma_x^2)\big)} \tag{10}$$

We also derived the standard error of the corrected effect as well as the covariance between observed and corrected effects in Supplementary Section C. Note that the formula proposed by Burgess *et al.*[4] (Supplementary Section D) under the null is a special case of ours when all instruments are selected based on external data (i.e. there is no winner's curse).
This approach has been implemented in an R-package (`MRlap`).

2.2 Simulations

We used UKBB[5] genotypic data and restricted our analyses to unrelated individuals of British ancestry (identified using genomic principal components) and HapMap3 genetic variants[15]

($M \approx 1,150,000$) to simulate phenotypic data. From this set of 379,530 individuals, we first sampled the exposure dataset (n_A individuals) and 5 different outcome datasets of sample size n_B, with an overlap with the exposure dataset varying (from no overlap to full overlap, increasing in increments of 25%). Next, causal SNPs for the exposure were randomly drawn from the set of 1,150,000 genetic variants, based on the polygenicity of X (π_x) and their effects were simulated using the heritability of X (h_x^2) as follows:

$$\boldsymbol{\gamma_x} \sim \begin{cases} 0 & \text{with probability } 1-\pi_x\text{, for non-causal variants} \\ \mathcal{N}\left(0, \frac{h_x^2}{M \cdot \pi_x}\right) & \text{with probability } \pi_x\text{, for causal variants} \end{cases} \tag{11}$$

For simplicity, we assumed that there were no direct genetic effects on the outcome. Then, phenotypic data for X and Y were simulated for all individuals included in the exposure or in any of the outcome samples, taking into account the effect of the confounder U on X and Y (respectively κ_x and κ_y) and the causal effect of X on Y (α), using the following design:

$$\boldsymbol{u} \sim \mathcal{N}(0,1) \tag{12}$$

$$\boldsymbol{x} = G \cdot \boldsymbol{\gamma_x} + \boldsymbol{u} \cdot \kappa_x + \epsilon_x \tag{13}$$

$$\text{with } \epsilon_x \sim \mathcal{N}(0, 1-(h_x^2+\kappa_x^2))$$

$$\boldsymbol{y} = \alpha \cdot \boldsymbol{x} + \boldsymbol{u} \cdot \kappa_y + \epsilon_y \tag{14}$$

$$\text{with } \epsilon_y \sim \mathcal{N}(0, 1-(\alpha^2+\kappa_y^2+2\cdot\alpha\cdot\kappa_y\cdot\kappa_x))$$

Note that these settings ensure that both X and Y have a zero mean and a variance of 1. A GWAS was performed for each sample (the exposure one and the five outcome ones) using BGENIE[16]. The GWAS summary statistics obtained were used to perform downstream analyses. MR analyses were performed using the Two-Sample MR R-package[17]: IVs were selected based on their observed effect (with a varying T threshold) and distance-pruned (500kb) to obtain a set of independent IVs. LDSC analyses were perfomed using the LDSC implementation from the GenomicSEM R-package[18] and the 1000G LD-scores[19]. Corrected effects were estimated using Equation 10.

For each parameter setting we tested, 100 datasets were simulated. Our standard parameter settings consisted of simulating data for $n_A = 20,000$ and $n_B = 20,000$ individuals. X was simulated with moderate polygenicity and large heritability ($\pi_x = 0.001$ and $h_x^2 = 0.4$). U had a moderate effect on both X and Y ($\kappa_x = 0.3$, $\kappa_y = 0.5$), leading to a correlation (ρ) of 0.15 induced by the confounder. X had a moderate direct causal effect on Y ($\alpha = 0.2$).
In addition to these standard settings, we explored various other parameter values. We investigated the effect of a confounder acting in an opposite direction ($\kappa_x = -0.3$ and $\kappa_y = 0.5$) and tested different strengths for the confounding factor (weaker : $\kappa_x = 0.15$ and $\kappa_y = 0.3$, and stronger : $\kappa_x = 0.5$ and $\kappa_y = 0.8$). We also simulated data for a scenario without any causal relationship between X and Y ($\alpha = 0$). Finally, we explored a scenario with more realistic parameters: larger sample sizes ($n_A = 100,000$, $n_B = 100,000$), increased polygenicity

($\pi_x = 0.005$), lower heritability ($h_x^2 = 0.2$) and a smaller causal effect ($\alpha = 0.1$). For each scenario, observed and corrected causal effects were compared for different degrees of sample overlap and different instrument selection thresholds. Results quality was assessed using root-mean square error (RMSE).

For a given instrument selection threshold, we obtained 500 causal effect estimates: one for each of the 100 simulated data sets for each of the five sample overlap proportion scenarios. Causal effect estimates should ideally not depend on the extent of overlap between the exposure and outcome samples. To quantify the extend to which this holds, we grouped estimates according to which sample overlap proportion they came from and compared the between group variance relative to the within group variance of the estimates. A method that is robust to overlap between the exposure and outcome samples will have small between group variance relative to the variance of the estimator (characterised by the within group variance).

Finally, we tested for differences between observed and corrected effects using the following test statistics:

$$t_{\text{diff}} = \frac{\hat{\alpha}_{IVW} - \hat{\alpha}_c}{\sqrt{Var(\hat{\alpha}_{IVW}) + Var(\hat{\alpha}_c) - 2 \cdot Cov(\hat{\alpha}_{IVW}, \hat{\alpha}_c)}} \tag{15}$$

2.3 Application to UKBB

To assess the effect of sample overlap on real data, we used a design very similar to the one used for simulations. We used both genotypic and phenotypic data from UKBB[5] and restricted our analyses to the same subsets of individuals and genetic variants. From this set of individuals, we first sampled the exposure dataset (100,000 individuals) and 5 different outcome datasets (100,000 individuals), where the overlap with the exposure dataset varied (from no overlap to full overlap, increasing in increments of 25% - in the case of unequal sample sizes, the percentage of overlap for the samples will proportionally increase from 0 to the maximum value attainable given the difference in sample sizes). Note that we always used the full set of individuals to create the 100,000-individuals samples, so the percentage of missing data in each sample will be the same as in the UKBB. Therefore, the total number of individuals with phenotypic data (effective sample size) will vary depending on the traits. We used four obesity-related traits: body mass index (BMI), systolic blood pressure (SBP), number of cigarettes previously smoked daily (smoking) and alcohol intake frequency (alcohol). For smoking, answers "prefer not answer" and "less than one a day" were considered as missing. For alcohol, answers were recoded to correspond to an increased intake frequency, and answers "do not know" were considered missing. Details about the pairs of traits analysed are available in Table 1.
Phenotypic data was normalised (inverse-normal quantile transformed) and subsequently adjusted for the following covariates: sex, age, age×age and the first 40 principal components. Similarly to what we did for simulations, a GWAS was performed for each sample (the exposure dataset and the five outcome datasets) using BGENIE[16]. The GWAS summary statistics obtained were then used to perform the downstream MR analyses using the Two-Sample

Exposure	Outcome	Field ID Exposure	Field ID Outcome	N Exposure	N Outcome
BMI	SBP	21001	4080	93,219	99,490
SBP	BMI	4080	21001	99,491	93,231
BMI	smoking	21001	2887	99,490	23,507
BMI	alcohol	21001	1558	99,493	99,808

Table 1: Description of the pairs of traits analysed
For each pair of trait analysed, we reported the field IDs corresponding to the exposure and the outcome, as well as the mean sample size (across the 100 repetitions, and the 5 overlaps for the outcome) of the phenotypic data used.

MR R-package[17]: IVs were selected based on their observed effect (exploring various selection thresholds T) and distance-pruned (500kb) to obtain a set of independent IVs. In addition, since in this case a reverse causal effect (from the outcome on the exposure) could exist, we filtered out variants that were significantly more strongly associated with the outcome than with the exposure in order to remove potentially invalid IVs. LDSC analyses were performed using the LDSC implementation from the GenomicSEM R-package[18] and the 1000G LD-scores[14]. Corrected effects were estimated using Equation (10).

We repeated this sampling approach a 100 times. For each repetition, observed and corrected causal effects were compared using the within groups and between groups variances and we tested for differences between observed and corrected effects using Equation 15.

3 Results

3.1 Overview of the method

We propose a two-sample MR framework that takes into account three sources of bias simultaneously : weak instrument bias, sample overlap and winner's curse. We analytically derived the expectation of the observed effect for IVW-based MR estimate:

$$E[\widehat{\alpha}_{IVW}] = f(\alpha, n_A, n_B, T, \pi_X, \sigma_X^2, \lambda)$$

which depends on the true causal effect size (α), the sample sizes of the exposure and outcome GWASs (n_A, n_B), the threshold used to select IVs (T), the cross-trait LDSC intercept (λ, which depends on the degree of the sample overlap, the true causal effect and the strength of the confounder) and the genetic architecture of the exposure (polygenicity π_X and per-variant heritability $\sigma_X^2 = \frac{h_X^2}{M \cdot \pi_X}$, M being the number of variants). All parameters (except α) are either known or can be estimated from the data. This allows us to adjust the IVW causal effect estimate with the aim of making it unbiased.

3.2 Simulations

Simulation results under our standard settings show a large discrepancy of the observed causal effects (IVW-based estimate) estimated using different degrees of sample overlap, while the corrected effects are more closely aligned with the true causal effects (Figure 2 - A, Table S1). We observe a 10% overestimation of the causal effect for fully overlapping samples, and a 15% underestimation of the causal effect for non-overlapping samples, due to a combination of winner's curse and weak instrument bias. As expected from Equation 10, the bias is larger when using less stringent thresholds T. For all the thresholds tested, the ratio of the between groups and the within groups variances is larger (up to 26 times) for observed effects than for corrected effects (Table 2), highlighting the differences in observed effects when estimated using different degrees of overlaps. The fact that the within group variance is 1.3 times higher for the corrected effects is due to the slightly increased variance of the bias-corrected estimator. The RMSE of the observed effects is very dependent on the degree of overlap (being larger for non-overlapping and fully overlapping samples) while the RMSE of the corrected effects is consistent across varying degrees of overlap and up to 1.75 times lower for non-overlapping samples (Figure 2 - B). The corrected effects are significantly different from the observed effects for all overlaps values except 50% and all thresholds (Table S1). The absolute bias of the observed effects goes up to 0.03 while for the corrected effect it is smaller than 0.012 for all overlaps and thresholds.

The bias of observed effects depends on the strength of the confounder. If the confounder is weak, observed effects are mostly biased toward the null for low overlaps because of winner's curse and weak instrument bias (Figure S5, Table S2, Table S3). When we simulate a stronger confounder, the bias of the observed effects for fully overlapping samples increases (Figure S6, Table S4, Table S5). In both cases, the corrected and the observed effects are significantly

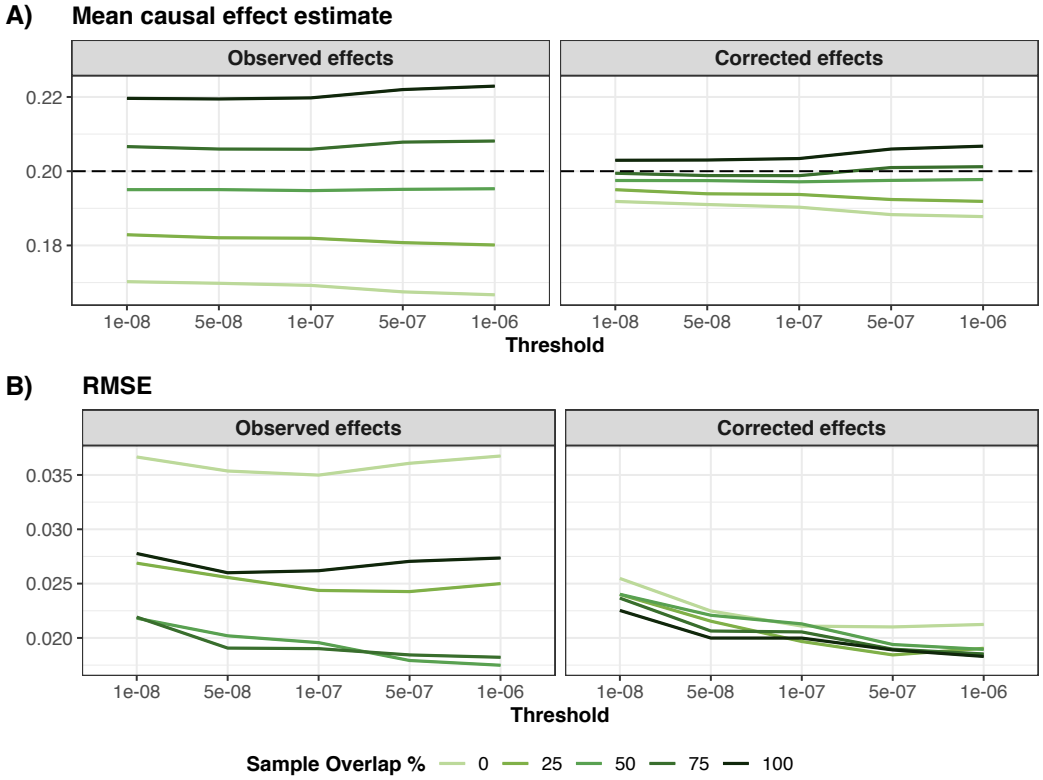

Figure 2: Simulation results for standard settings
$n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0.2$
Panel A) shows the mean observed and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represent the true causal effect). Panel B) shows the mean RMSE obtained for observed and corrected effect for each overlap and threshold.

		Observed effects			**Corrected effects**		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	51.5	0.000437	0.0375	86	0.000558	0.00178	3.20
5e-08	65.3	0.000339	0.0380	112	0.000434	0.00212	4.88
1e-07	72.3	0.000308	0.0391	127	0.000394	0.00248	6.30
5e-07	92.5	0.000260	0.0463	178	0.000331	0.00484	14.64
1e-06	102.9	0.000251	0.0493	196	0.000320	0.00562	17.56

Table 2: Analysis of variance for standard settings
$n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0.2$
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.

different for almost all overlaps and thresholds, and the corrected effects are substantially less biased than the observed effects (Table S2, Table S4). When the confounder effect (ρ) and the causal effect (α) are acting in opposite directions, the results are particularly interesting because winner's curse, weak instrument bias and sample overlap are all biasing the results toward the null (Figure 3 - A, Table S6). In this case, observed effects are more similar across the different

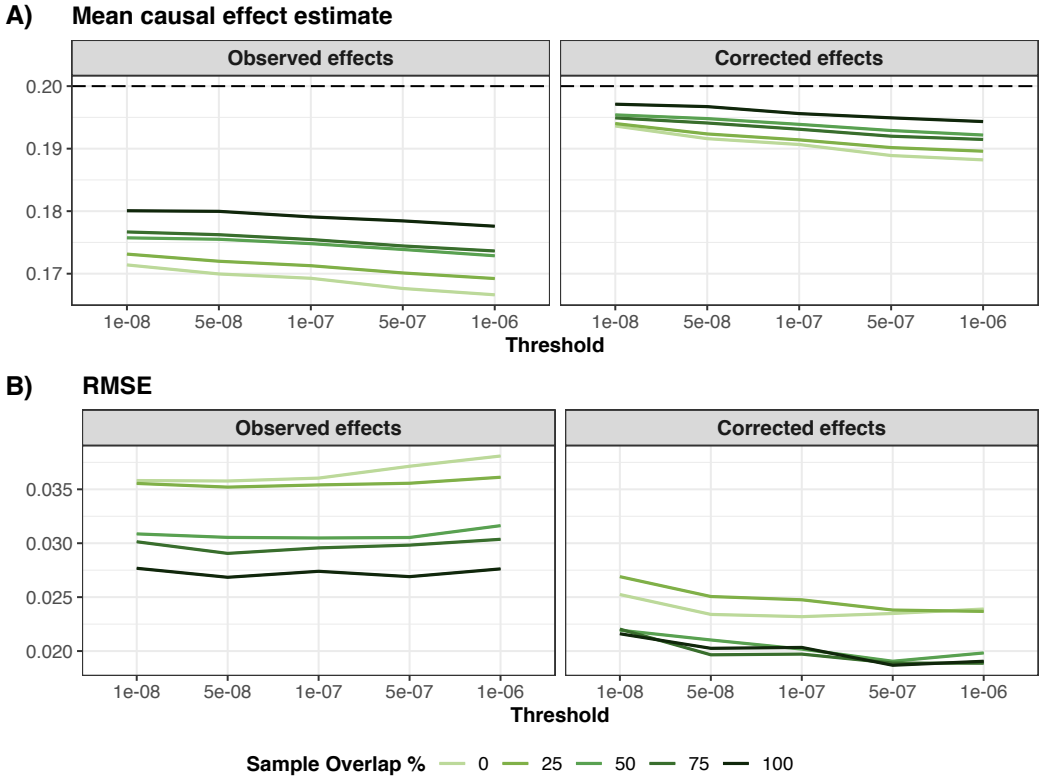

Figure 3: Simulation results for a scenario with a negative confounder $n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = -0.3, \kappa_y = 0.5, \alpha = 0.2$
Panel A) shows the mean observed and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represent the true causal effect). Panel B) shows the mean RMSE obtained for observed and corrected effect for each overlap and threshold.

		Observed effects			**Corrected effects**		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	51.2	0.000425	0.00112	2.63	0.000538	0.000187	0.348
5e-08	65.3	0.000356	0.00152	4.26	0.000447	0.000413	0.925
1e-07	72.7	0.000336	0.00146	4.35	0.000423	0.000389	0.920
5e-07	92.7	0.000293	0.00174	5.94	0.000369	0.000552	1.496
1e-06	102.7	0.000292	0.00178	6.10	0.000370	0.000559	1.512

Table 3: Analysis of variance for a scenario with a negative confounder $n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = -0.3, \kappa_y = 0.5, \alpha = 0.2$ For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.

degrees of sample overlap tested, but all are underestimating the true causal effect. For this reason, we do not observe such a large decrease in the heterogeneity of the estimates across different sample overlaps compared to the corrected effects (ratio between the between groups and the within groups variance is 5 times larger for observed effects), but the latter ones show lower RMSE and bias for all overlaps and thresholds (Figure 3 - B, Table 3). In this scenario, observed and corrected effects significantly differ for all overlaps and thresholds, with an average

underestimation of 13% for observed effects (Table S6).

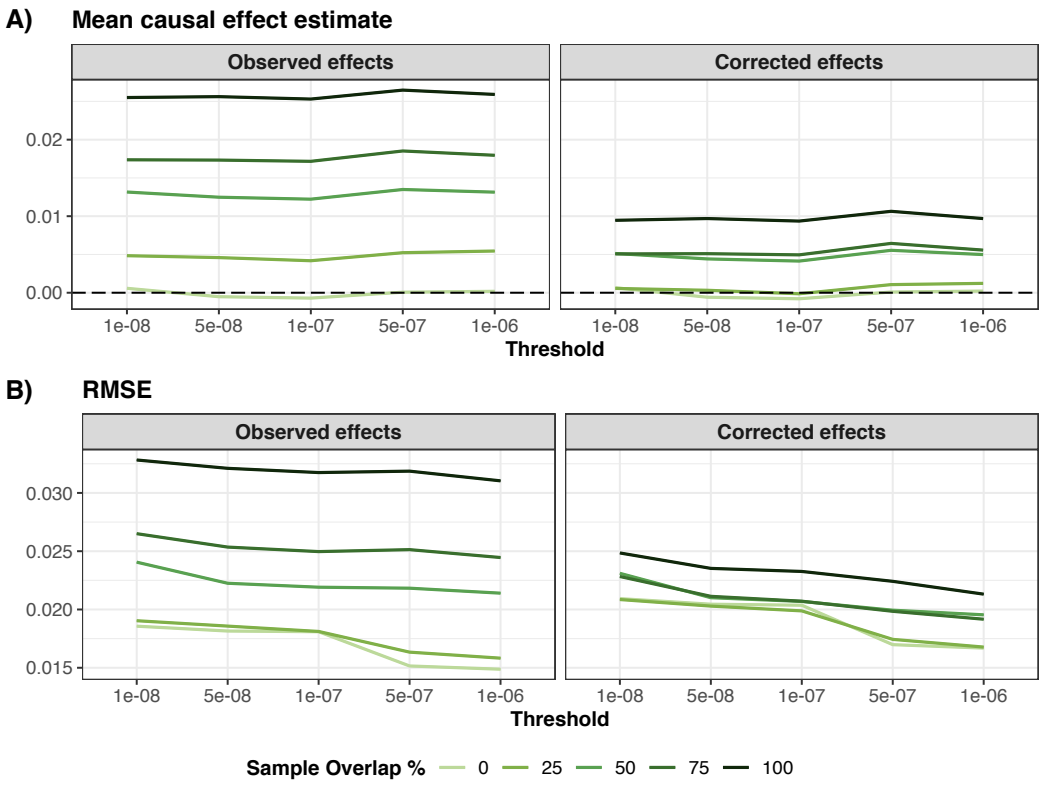

Figure 4: Simulation results in the absence of a causal effect
$n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0$
Panel A) shows the mean observed and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represent the true causal effect). Panel B) shows the mean RMSE obtained for observed and corrected effect for each overlap and threshold.

		Observed effects			**Corrected effects**		
Threshold	IVs	Within groups	Between groups	Ratio	Within groups	Between groups	Ratio
1e-08	52.2	0.000387	0.00985	25.4	0.000485	0.00138	2.84
5e-08	65.3	0.000345	0.01064	30.8	0.000431	0.00170	3.96
1e-07	72.0	0.000336	0.01064	31.6	0.000420	0.00171	4.08
5e-07	92.1	0.000276	0.01099	39.8	0.000343	0.00184	5.35
1e-06	102.2	0.000262	0.01029	39.3	0.000326	0.00143	4.40

Table 4: Analysis of variance in the absence of a causal effect
$n_A = n_B = 20,000, \pi_x = 0.001, h_x^2 = 0.4, \kappa_x = 0.3, \kappa_y = 0.5, \alpha = 0$
For each threshold, the mean number of instruments used (IVs), the within groups and between group variances, their ratio (between/within) for both observed and corrected effects are reported.

In the absence of a true causal relationship between the exposure and the outcome, observed effects from non-overlapping samples are unbiased. However, for fully overlapping samples, the

observed effects are biased toward the confounder's effect (Figure 4, Table 4). We showed that for large overlap percentages ($\geq$ 50%), the corrected effects are significantly different from the observed effects (60% smaller), and they are less biased for all overlaps and thresholds (Table S7).

Results obtained using more realistic parameters in terms of sample sizes, genetic architecture and causal effect strength show a similar pattern. We observed an important bias of observed causal effects, mostly when estimated from non-overlapping (22% underestimation) or fully overlapping samples (30% overestimation), that is strongly reduced when using our correction (Figure S7, Table S8, Table S9). Corrected effects significantly differ from observed effects for the most extreme overlaps values (0%, 75%, 100%) for which corrected effects are on average 5 times less biased than observed effects.

3.3 Application to UKBB

We tested our method on UKBB obesity-related traits using a similar approach and splitting the full dataset into samples of varying degrees of overlap. When looking at the observed effect of BMI on SBP (Figure 5), we observed that the estimates obtained using different p-value thresholds vary considerably, independently of sample overlap. Even though we expect an increase in bias when reducing the threshold used, as shown in simulations, here we see that for non-overlapping sample the observed effect is larger for less stringent thresholds. This is inconsistent with winner's curse and weak instrument bias that lean the estimate toward the null and we would expect the observed effects for less stringent thresholds to be smaller (in absolute value). Hence, we believe that this phenomenon is not related to any of the biases discussed here and is due to other reasons such as the existence of multiple causal effects depending on exposure sub-type or the presence of a heritable confounder (see Discussion). Here we will focus on the results obtained using a p-value threshold of 5e-8.

When using IVs reaching genome-wide significance, the observed effects range between 0.095 for non-overlapping samples, and 0.13 for fully overlapping samples (Table S10). After correction, the range of the estimated effect is about 2 times smaller (0.108 - 0.125). The better agreement of corrected effects across overlaps can be seen by looking at the ratio between the between groups and the within groups variance which is 4 times larger for observed effects (Table S11). For this relationship, winner's curse and weak instrument bias seem to be the strongest sources of bias, arising when using non-overlapping samples. The difference between observed and corrected effect is significant ($p_{\text{diff}} = 0.0048$), and standard two-sample MR underestimates the causal effect by about 25% (the observed effect is 0.095 while the corrected effect is 0.126). For fully overlapping samples, there is no significant difference between observed and corrected effects. We observed a similar pattern when investigating the effect of BMI on smoking (Figure S8, Table S12 , Table S13) where the largest bias occurs for non-overlapping samples. The observed causal effect at p=5e-8 is 0.131 for non-overlapping samples while results after correction point toward a causal effect of 0.172 (underestimation of 24% - significant difference between observed and corrected effects, $p_{\text{diff}} = 0.0115$). We do not see a significant difference between observed and corrected effects for larger overlap values, but it is important to note that in this case, because

of the impact of missing data on our design, the largest possible overlap was only 48.5%.

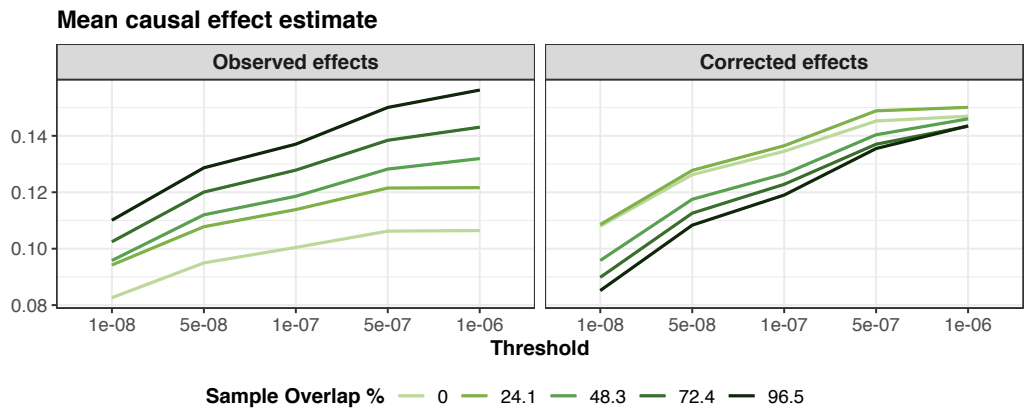

Figure 5: Effect of BMI on SBP
This figure shows the mean observed and corrected effect for each overlap and threshold obtained from 100 different sampled datasets.

When investigating the causal effect of SBP on BMI, the observed effects using a p-value threshold of 5e-8 were ranging from -0.0454 (for non overlapping samples) to 0.0031 (for fully overlapping samples) (Figure 6, Table S14). However, the corrected effects for different overlaps show a strong agreement, with a ratio between the between groups and the within groups variance that is is 360 times larger than that of observed effects (Table S15). The corrected effects across all degrees of sample overlap point toward a causal effect of -0.065 with the difference between observed and corrected effects being significant for larger overlaps ($p_{\text{diff}} = 0.028$ for 72.4% overlap, and $p_{\text{diff}} = 0.012$ for 96.5% overlap). It is known that SBP and BMI share risk factors, such as sedentary lifestyle[20], that would increase both SBP and BMI. Such a confounder would induce a positive environmental correlation between the traits, leading to a bias toward the null for all overlaps (all corrected effects are stronger than observed effects). We have shown that the extent of the bias induced by the existence of a confounder when using overlapping samples depends on several parameters, including the genetic architecture of the exposure. This could explain why we see a significant effect of the correction for fully overlapping samples when looking at the effect of SBP on BMI and not for the reverse direction (BMI on SBP). With the heritability of BMI being larger than that of SBP, we expect a smaller bias when this trait is used as an exposure. Our results also implicate the existence the existence of an environmental confounder biasing the causal effect estimate of BMI on alcohol consumption. At p=5e-8, observed effects range between -0.187 for non overlapping samples and -0.218 for fully overlapping samples, whereas the corrected effects are larger (-0.253 to -0.329) (Figure S9, Table S16). In this case however, the between groups variance is larger for corrected effects (Table S17). The difference between observed and corrected effects is significant for all overlaps, and the corrected effects being stronger than the observed ones hints at the existence of an environmental confounder having a concordant effect on BMI and alcohol frequency intake, biasing all estimates toward the null (as shown in simulations, Figure 3). While we could not identify any plausible confounder of

this relationship, its existence is supported by the fact that the phenotypic correlation between BMI and alcohol frequency (-0.13 among the 379,530 genetically British individuals in the UK Biobank) is weaker than the standardised causal effect.

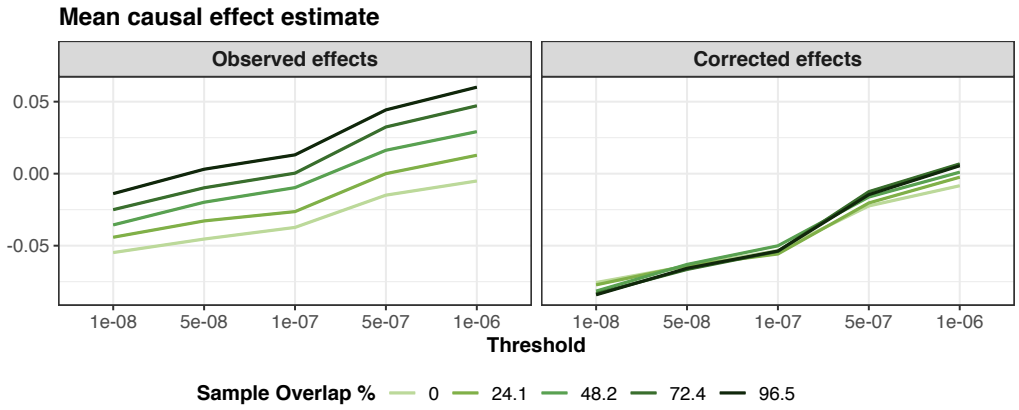

Figure 6: Effect of SBP on BMI
This figure shows the mean observed and corrected effect for each overlap and threshold obtained from 100 different sampled datasets.

4 Discussion

We developed a method that simultaneously accounts for biases arising from winner's curse, weak instruments and sample overlap, even when the degree of overlap is unknown. To the best of our knowledge, no other method can handle all these biases simultaneously while using summary statistics from exposure and outcome GWASs with arbitrary sample overlap. We tested our approach using a wide range of simulations scenarios: varying the strength of the causal effect, the strength of the confounder effect, sample sizes for the exposure and the outcome, as well as the genetic architecture of the exposure and demonstrated that both estimates for non-overlapping and fully overlapping samples can be biased. The direction and the magnitude of the bias depends on sample overlap and is strongly influenced by the effect of the confounder. When the confounder and the causal effect are acting in the same direction, observed effects are overestimated for fully-overlapping samples and underestimated for non-overlapping samples. However, when they are acting in opposite directions, these are underestimated for all overlaps because the direction of the three sources of biases is toward the null. We also showed that in the absence of a causal effect, results from overlapping samples would be biased, potentially leading to elevated type I error.

The correction we proposed worked remarkably well under all scenarios, and allows to drastically reduce the bias. For standard settings for example, we observed a 15% overestimation for fully overlapping samples and 10% underestimation for non-overlapping samples, that were respectively reduced to a 5% overestimation and a 2% underestimation after correction. We also found significant differences between observed and corrected effects for fully-overlapping samples under all scenarios. For non-overlapping samples, observed and corrected effects were significantly different under all scenarios expect in the case of the absence of causal effect. The decreased bias comes with increased variance (due to the correction) however in all of our simulation scenarios it reduced estimation error (RMSE) for at least one, if not all, sample overlap degrees. Moreover, while the RMSE of observed effects strongly depends on the degree of overlap (because the bias is overlap-dependent), the RMSE of corrected effects is very similar for all overlaps.

For real data, strong discrepancies were observed for low degrees overlap (i.e. winner's Curse and weak instrument bias) in three out of the four relationships we looked at, with significant differences between observed and corrected effects. This means that standard two-sample MR settings are often leading to an underestimation of the true causal effect, that can be corrected using our approach. We also demonstrated that while most studies are extremely keen on avoiding any sample overlap while performing two-sample MR analysis fearing potential bias, it is often much less substantial than weak instrument bias or winner's curse. In addition, we identified two trait pairs (SBP→BMI and BMI→alcohol use frequency) for which estimates for fully-overlapping samples are biased by an environmental confounder. In both cases, the confounder and the causal effect were acting in opposite direction, leading to an underestimation of the IVW-based effect for all overlaps. We have also highlighted that there is an important heterogeneity in

causal effect estimates that vary with the IV selection threshold, due to heterogeneity in the estimates between the groups of genetic variants used for different thresholds. This can happen if there is a strong phenotypic heterogeneity in the exposure, in which case different groups of IVs could be affecting the exposure through different pathways[21]. Alternatively, in the presence of a genetic confounder, IVs picked up at a less stringent thresholds may be associated to the confounder, hence violating the second assumption of MR. Such phenomenon is out of the scope of our paper. In such case, IVW two-sample MR estimates would be biased, and more sophisticated approaches either specifically accounting for this genetic confounding (CAUSE[22], LHC-MR[23]) or others allowing for multiple causal effects (MR-Clust[21]) would be needed.

Our approach has its own limitations. As IVW-MR estimates, our corrected effect estimates will also be biased in case of the existence of a genetic confounder through which some of the selected instruments are primarily acting on the exposure. In addition, our analytical derivation hinges on a genetic architecture of the exposure, namely assuming a spike and slab distribution of the multivariable effect sizes. Although this is a widely used and confirmed polygenic model, deviations from it could reduce the efficiency of our bias correction. Finally, our work focused on continuous traits, and we did not explore the effect of sample overlap when using case-control designs. To do this, we would need to consider that the overlap degree might differ between cases and controls.

The effect of these three sources of bias will decrease as the sample size increases, and for very large sample sizes, such as UKBB, this bias is expected to be less pronounced. However, we showed that the overall bias is also dependent on the genetic architecture of the exposure, and low heritability and/or highly polygenic traits will be more strongly affected. Estimating the corrected effect using our approach (implemented in an R-package to facilitate its use) can be performed as a sensitivity analysis: if the corrected effect does not significantly differ from the observed effect, then the IVW-MR estimate can be safely used (with the advantage of having lower SE). However, if there is a significant difference, corrected effects should be preferred as they are be less biased, independently of the sample overlap.

Acknowledgements

The authors thank Chiara Auwerx, Liza Darrous, Sven Erik Ojavee, Marion Patxot, Eleonora Porcu, Marie Sadler and Tabea Schoeler for the helpful discussions and relevant comments.
Z.K. was funded by the Swiss National Science Foundation (# 310030-189147).
Computations have been performed on the HPC cluster of the Lausanne University Hospital.

Code availability

The `MRlap` R-package is freely available under a GPL-2 License, and can be accessed, alongside with user guides and tutorials, from https://github.com/n-mounier/MRlap.

References

[1] Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statistics in medicine.

[2] Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. Genetic Epidemiology *37*, 658–665.

[3] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. American journal of human genetics.

[4] Burgess, S., Davies, N. M., and Thompson, S. G. (2016). Bias due to participant overlap in two-sample mendelian randomization. Genetic Epidemiology *40*, 597–608.

[5] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine.

[6] Burgess, S. and Thompson, S. G. (2011). Bias in causal estimates from mendelian randomization studies with weak instruments. Statistics in Medicine *30*, 1312–1323.

[7] Burgess, S. and and, S. G. T. (2011). Avoiding bias from weak instruments in mendelian randomization studies. International Journal of Epidemiology *40*, 755–764.

[8] Zheng, J., Baird, D., Borges, M.-C., Bowden, J., Hemani, G., Haycock, P., Evans, D. M., and Smith, G. D. (2017). Recent developments in mendelian randomization studies. Current Epidemiology Reports *4*, 330–345.

[9] Bowden, J., M., F. D. G., Minelli, C., Smith, G. D., Sheehan, N. A., and Thompson, J. R. (2016). Assessing the suitability of summary data for two-sample mendelian randomization analyses using MR-egger regression: the role of the i2 statistic. International Journal of Epidemiology pp. dyw220.

[10] Bowden, J., M, F. D. G., Minelli, C., Smith, G. D., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data mendelian randomization. Statistics in Medicine *36*, 1783–1802.

[11] Palmer, C. and Pe'er, I. (2017). Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. PLOS Genetics *13*, e1006916.

[12] Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics *9*, 621–634.

[13] Bowden, J., Smith, G. D., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. International Journal of Epidemiology *44*, 512–525.

[14] Bulik-Sullivan, B., , Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nature Genetics *47*, 1236–1241.

[15] Consortium, I. H. . (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

[16] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature.

[17] Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. eLife.

[18] Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., Ip, H. F., Marioni, R. E., McIntosh, A. M., Deary, I. J., et al. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nature Human Behaviour *3*, 513–525.

[19] Bulik-Sullivan, B. K., , Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics *47*, 291–295.

[20] Sohn, M.-W., Manheim, L., Chang, R., Greenland, P., Hochberg, M., Nevitt, M., Semanik, P., and Dunlop, D. (2014). Sedentary behavior and blood pressure control among osteoarthritis initiative participants. Osteoarthritis and Cartilage *22*, 1234–1240.

[21] Foley, C. N., Mason, A. M., Kirk, P. D. W., and Burgess, S. (2020). MR-clust: clustering of genetic variants in mendelian randomization with similar causal estimates. Bioinformatics.

[22] Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nature Genetics *52*, 740–747.

[23] Darrous, L., Mounier, N., and Kutalik, Z. (2020). Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. medRxiv.