Received: 20 September 2022 | Revised: 2 March 2023 | Accepted: 14 March 2023

DOI: 10.1002/gepi.22522

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

Bias correction for inverse variance weighting Mendelian randomization

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

[1]Department of Epidemiology and Health Systems, University Center for Primary Care and Public Health, Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

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

Correspondence
Zoltán Kutalik, University Center for Primary Care and Public Health, 1010 Lausanne, Switzerland.
Email: zoltan.kutalik@unil.ch

Funding information
Department of Computational Biology (University of Lausanne); Swiss National Science Foundation

Abstract

Inverse-variance weighted two-sample Mendelian randomization (IVW-MR) is the most widely used approach that utilizes genome-wide association studies (GWAS) summary statistics to infer the existence and the strength of the causal effect between an exposure and an outcome. Estimates from this approach can be subject to different biases due to the use of weak instruments and winner's curse, which can change as a function of the overlap between the exposure and outcome samples. We developed a method (`MRlap`) that simultaneously considers weak instrument bias and winner's curse while accounting for potential sample overlap. Assuming spike-and-slab genomic architecture and leveraging linkage disequilibrium score regression and other techniques, we could analytically derive, reliably estimate, and hence correct for the bias of IVW-MR using association summary statistics only. We tested our approach using simulated data for a wide range of realistic settings. In all the explored scenarios, our correction reduced the bias, in some situations by as much as 30-fold. In addition, our results are consistent with the fact that the strength of the biases will decrease as the sample size increases and we also showed that the overall bias is also dependent on the genetic architecture of the exposure, and traits with low heritability and/or high polygenicity are more strongly affected. Applying `MRlap` to obesity-related exposures revealed statistically significant differences between IVW-based and corrected effects, both for nonoverlapping and fully overlapping samples. Our method not only reduces bias in causal effect estimation but also enables the use of much larger GWAS sample sizes, by allowing for potentially overlapping samples.

KEYWORDS

Mendelian randomization, sample overlap, weak instrument bias, winner's curse

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 the strength of the causal effect between an exposure and an outcome (Lawlor et al., 2008). In particular, two-sample summary data MR (Burgess et al., 2013), which 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 (Visscher et al., 2017) and the resulting summary statistics are often publicly available. This allows not only the identification of genetic variants independently associated with a particular exposure, that is, IVs but also the look-up of the effect of such variants on a wide range of outcome traits in different samples. Each IV provides an independent estimate for the causal effect and these estimates can then be combined using a fixed effect, inverse variance-weighting (IVW) meta-analysis (Burgess et al., 2013). When the summary statistics for the two-sample IVW-MR estimator come from the same sample, it is equivalent to the two-stage least squares estimator applied to individual-level data from that sample. 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.

In addition, two-sample MR methods expect the exposure and the outcome GWAS summary statistics to be obtained from independent samples. Sample overlap acts as a modifier for two well-known sources of bias in MR, weak instrument bias and winner's curse. Two-sample MR estimates obtained from nonoverlapping samples are known to be biased toward the null. When using overlapping samples, the causal effect estimate can be biased toward the observational correlation, which includes the correlation induced by confounders (Burgess et al., 2016). Most MR methods assume that SNP-exposure effects are measured without noise (NO Measurement Error [NOME] assumption) (Bowden, Del Greco, et al., 2016). This simplification

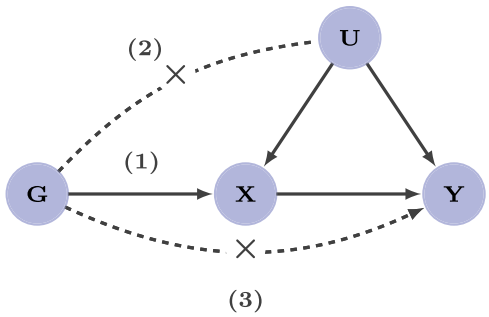

FIGURE 1 Main assumptions of Mendelian randomization. (1) Relevance—instrumental variables (IVs), denoted by G, are strongly associated with the exposure. (2) Exchangeability—G is not associated with any confounder of the exposure–outcome relationship. (3) Exclusion restriction—G is independent of the 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).

leads to regression dilution bias, which increases as the instruments get weaker. 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 (Burgess & Thompson, 2011b). When combined with sample overlap, the effect of weak instrument bias will move toward the observational correlation (Burgess & Thompson, 2011a; Zheng et al., 2017). Using IVs strongly associated with the exposure and/or increasing the sample size can mitigate weak instrument bias (Burgess & Thompson, 2011a). 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 (Bowden, Del Greco, et al., 2016). The simulation-extrapolation-based SIMEX method proved to yield more robust corrections both for IVW (Bowden et al., 2017) and MR-Egger estimates (Bowden, Del Greco, et al., 2016). More recently, methods that incorporate measurement errors have been proposed to tackle weak instrument bias. This is the case, for example, of radial MR that uses modified second-order weights (Bowden et al., 2018), MR-RAPS (Zhao et al., 2020) that maximizes the profile likelihood of the ratio estimate, or dIVW (Ye et al., 2021) that uses an explicit bias correction factor, based on asymptotic properties, to de-bias the IVW estimator. In addition, MR estimates are subject to winner's curse, which occurs when the same sample is used to select IVs and estimate their effect on the exposure. In such a 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 nonoverlapping samples and bias towards the observational correlation in fully overlapping samples) (Zheng et al., 2017). Using a third independent sample to select instruments (Zhao et al., 2019), and therefore avoiding winner's curse, is a valid solution, but summary statistics from such additional samples are rarely available. Based on the expectation of truncated normal distribution (Palmer and Pe'er, 2017), 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 maximizing the conditional likelihood (Zhong & Prentice, 2008). Still, all these methods account for winner's curse for a single SNP but do not model the bias induced by the IV selection process from millions of potential markers, which is far more complex and depends on the underlying genetic architecture of the exposure. 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 handles them. To fill this gap, we propose a new method called `MRlap`, which is a summary statistics-based MR framework that simultaneously takes into account weak instrument bias and winner's curse, while accounting for potential sample overlap, which modifies these biases.

Another major source of bias in MR is pleiotropy. In the presence of uncorrelated pleiotropy (reducing the exclusion restriction to the INSIDE assumption; Bowden et al., 2015), the causal effect estimate from IVW-MR is still consistent. Correlated pleiotropy (which can be induced by the existence of a genetic confounder acting on both the exposure and the outcome), however, can lead to the violation of the second assumption and more severe biases. Some approaches can be used to relax this assumption by assuming that at least 50% of the instruments are valid (Bowden, Davey Smith, et al., 2016) or that non-pleiotropic instruments are the most frequent (Hartwig et al., 2017). Since our `MRlap` approach does not explicitly tackle pleiotropy, we compared the impact of different sources of biases (pleiotropy, weak instrument bias, winner's curse in the presence of potential sample overlap) on IVW-MR, MR-RAPS, dIVW, weighted median, weighted mode, and `MRlap`.

In this paper, we will first introduce a two-sample MR framework that simultaneously takes into account weak instrument bias and winner's curse, while accounting for potential sample overlap and its effect as a modifier of these biases, 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 (and some pleiotropy robust methods) using a wide range of simulation settings, including scenarios with pleiotropy. Finally, to demonstrate its utility, we will apply our approach to obesity-related traits using UK Biobank (UKBB; Sudlow et al., 2015) data.

2 | METHODS

2.1 | Expectation of the causal effect estimate

Let X and Y denote two random variables representing two complex traits. Genotype data is denoted by G and its jth column by g_j (columns representing genetic variants, rows representing individuals). To simplify notation we assume that $\mathrm{E}[X] = \mathrm{E}[Y] = \mathrm{E}[G] = 0$ and $\mathrm{Var}(X) = \mathrm{Var}(Y) = \mathrm{Var}(G) = 1$. Let us assume that X is observed in sample A of sample size n_A, Y is observed in sample B of sample size n_B with an overlap of $n_{A\cap B}$ individuals between the two samples. The vector of realizations of Z^C is denoted by z^C for all variables $(Z = X, Y, G, g)$ and samples $(C = A, B, A \cap B)$. Let us assume the following models

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

where $\boldsymbol{\gamma_x}$ are the effect sizes of the genetic variants on $\boldsymbol{X}$, $\boldsymbol{\gamma_y}$ are their pleiotropic effects on $\boldsymbol{Y}$. Assuming that there is a single environmental confounder $\boldsymbol{U}$ (with $\mathrm{E}[\boldsymbol{U}] = 0$ and $\mathrm{Var}(\boldsymbol{U}) = 1$) acting linearly on both traits (as used for simulations) the error term can split into two parts: $\boldsymbol{\epsilon}_{\boldsymbol{x}}^C = \kappa_x \times \boldsymbol{u}^C + \boldsymbol{\varepsilon}^C$ and $\boldsymbol{\epsilon}_{\boldsymbol{y}}^C = \kappa_y \times \boldsymbol{u}^C + \boldsymbol{\varepsilon}^C$, where κ_x and κ_y refer to the effect of $\boldsymbol{U}$ on $\boldsymbol{X}$ and $\boldsymbol{Y}$, 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 (Bowden et al., 2015) (INstrument Strength Independent of Direct Effect, i.e., horizontal pleiotropic effects are independent of the direct effect), $\mathrm{Cov}(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}) = 0$ and $\mathrm{E}[\boldsymbol{\gamma_y}] = 0$. We denote $\rho := \mathrm{Cov}(\boldsymbol{\epsilon_x}, \boldsymbol{\epsilon_y}) = \kappa_x \times \kappa_y$. It corresponds to the part of the observational correlation (r) due to a (nongenetic) confounder ($r = \rho + \alpha$). Note that genetic confounding, as well as reverse causal effect, are also affecting observational correlation, but as long as the instruments used for MR are not associated with the confounder nor the outcome, their effect would be captured by ρ.

The univariable effect size estimates from GWASs for genetic variant j are as follows:

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

where the genotype data for genetic variant j for individuals in sample A is denoted by $\boldsymbol{g}_j^A$.

We intend to use MR to estimate the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$. We will use m linkage disequilibrium (LD) independent genetic variants as IVs. Let us now consider the fixed-effect IVW meta-analysis for the ratio estimates for the causal effect α. Each IV j provides a ratio estimate

$$\hat{\alpha}_j = \frac{\left(\hat{\beta}_y^B\right)_j}{\left(\hat{\beta}_x^A\right)_j}, \tag{3}$$

$$\mathrm{Var}(\hat{\alpha}_j) = \frac{\mathrm{Var}\left(\left(\hat{\beta}_y^B\right)_j\right)}{\left(\hat{\beta}_x^A\right)_j^2} = \frac{\left(1-\alpha^2-\gamma_{y_j}^2\right)}{n_B \times \left(\hat{\beta}_x^A\right)_j^2}. \tag{4}$$

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

$$w_j = \frac{1}{\mathrm{Var}(\hat{\alpha}_j)} = \frac{n_B \times \left(\hat{\beta}_x^A\right)_j^2}{\left(1-\alpha^2-\gamma_{y_j}^2\right)}. \tag{5}$$

Finally, the estimate can be written in the following form

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

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 (for example, $T = -\Phi^{-1}(p/2) \approx 5.45$ for $p = 5 \times 10^{-8}$, genome-wide significance threshold) and use only IVs with $|(\hat{\beta}_x^A)_j| \times \sqrt{n_A} > T$. By denoting $S_j := \{|(\hat{\beta}_x^A)_j| \times \sqrt{n_A} > T\}$, the causal effect estimate (6) changes to

$$\hat{\alpha}_{\mathrm{IVW}} \approx \frac{\sum_{j=1}^{M} \left(\left(\hat{\beta}_y^B\right)_j | S_j\right) \times \left(\left(\hat{\beta}_x^A\right)_j | S_j\right) \times Pr(S_j)}{\sum_{j=1}^{M} \left(\left(\hat{\beta}_x^A\right)_j | S_j\right)^2 \times Pr(S_j)}. \tag{7}$$

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 (see Supporting Information Section SA for details about this assumption), the expectation of the causal effect estimate (7) can be written as

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

The values of s_j, t_j, and $Pr(S_j)$ can be analytically derived (presented in Supporting Information Section SA), and we show that (8) translates to

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

with

$$\begin{aligned}
a &= \Phi\left(-\frac{T}{\sqrt{1 + n_A \times \sigma_x^2}}\right) \\
b &= 2T \times \frac{\exp\left(-\frac{T^2}{2\left(n_A\sigma_x^2 + 1\right)}\right)}{\sqrt{2\pi}\left(1 + n_A\sigma_x^2\right)^{3/2}} \\
c &= \Phi(-T) + T \times \phi(T),
\end{aligned} \tag{10}$$

and

$$d\left(n_A, T, \pi_x, \sigma_x^2\right) = \pi_x \times \left(2 \times \left(\sigma_x^2 + \frac{1}{n_A}\right) \times a + b \times \left(n_A \times \sigma_x^4 + 2\sigma_x^2 + \frac{1}{n_A}\right)\right) + (1 - \pi_x) \times \frac{2}{n_A} \times c. \quad (11)$$

π_x and σ_x^2 are characteristics of the genetic architecture of trait $\boldsymbol{X}$ (respectively, a measure of the polygenicity and the per variant heritability, see S35) (Zeng et al., 2018) and λ' is a quantity closely related to the cross-trait LD score regression (LDSC) intercept (Bulik-Sullivan, Finucane, et al., 2015) (λ): $\lambda' = \frac{\lambda}{\sqrt{n_A \times n_B}}$. 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 10 and S33). The same is true for the denominator (11).

We can see that in the absence of sample overlap, the second term is equal to 0 and the bias is multiplicative. In this case, $\mathrm{E}[\hat{\alpha}_{\mathrm{IVW}}]$ is lower than α and the IVW-based effect will be biased towards 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/or increasing the exposure sample size reduces the bias.

When there is sample overlap (with % overlap defined as $\frac{n_{A\cap B}}{\sqrt{n_A \times n_B}}$), the expression of $\mathrm{E}[\hat{\alpha}_{\mathrm{IVW}}]$ is more complex. The magnitude of the bias will not only depend on the parameters described for nonoverlapping samples (Figures S2–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 towards the null, whereas a confounder having an opposite sign (as the causal effect) will reduce the bias (Figure S2). When the percentage of overlap increases (Figures S3 and S4), the secondterm starts to dominate, and the bias direction strongly depends on ρ.

All parameters except α are known or can be estimated from the data. We assume that sample sizes for $\boldsymbol{X}$ and $\boldsymbol{Y}$ (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 (Bulik-Sullivan, Finucane, et al., 2015). π_x and σ_x^2 are estimated by matching the denominator of this formula, $d\left(n_A, T, \pi_x, \sigma_x^2\right)$, to the denominator of (8) (see Supporting Information Section SB for details).

From (9), we can derive a corrected effect for the causal effect

$$\hat{\alpha}_c = \frac{\hat{\alpha}_{\mathrm{IVW}} \times d\left(n_A, T, \pi_x, \sigma_x^2\right) - \lambda' \times \left(\pi_x \times \left(2 \times a + b \times \left(1 + \sigma_x^2 \times n_A\right)\right) + (1 - \pi_x) \times 2 \times c\right)}{\left(\pi_x \times \sigma_x^2\right) \times \left(2 \times a + b \times \left(1 + n_A \times \sigma_x^2\right)\right)}. \quad (12)$$

We also derived the standard error of the corrected effect as well as the covariance between IVW-based and corrected effects in Supporting Information Section SC. Note that the formula proposed by Burgess et al. (2016) (Supporting Information Section SD) 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`—https://github.com/n-mounier/MRlap/), using existing functions from the two-sample MR R-package (Hemani et al., 2018) and the LDSC implementation from the GenomicSEM R-package (Grotzinger et al., 2019). All analyses presented in this paper have been performed using version 0.0.2. IVs were selected for different T thresholds and independent IVs were identified using distance pruning (500 kb, which is equivalent to using LD pruning with an LD cutoff of 0). For the LDSC analyses, we used the 1000G LD scores (Bulik-Sullivan, Loh, et al., 2015).

2.2 | Simulations

We used UKBB (Sudlow et al., 2015) genotypic data and restricted our analyses to unrelated individuals of British ancestry (identified using genomic principal components) and HapMap3 genetic variants (International HapMap 3 Consortium, 2010) ($M \approx 1,150,000$) to simulate phenotypic data. From this set of 379,530 individuals, we first sampled the exposure dataset (n_A individuals) and five 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 $\boldsymbol{X}$ (π_x), and their effects were simulated using the heritability of $\boldsymbol{X}$ (h_x^2) as follows:

$$\boldsymbol{\gamma_x} \sim \begin{cases} 0 & \text{with probability } 1-\pi_x, \\ & \text{for non-causal variants} \\ \mathcal{N}\left(0, \dfrac{h_x^2}{M \times \pi_x}\right) & \text{with probability } \pi_x, \\ & \text{for causal variants} \end{cases}. \quad (13)$$

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

$$\boldsymbol{u} \sim \mathcal{N}(0,1), \quad (14)$$

$$\boldsymbol{x} = G \times \boldsymbol{\gamma_x} + \boldsymbol{u} \times \kappa_x + \epsilon_x \quad \text{with } \epsilon_x \sim \mathcal{N}\left(0, 1 - \left(h_x^2 + \kappa_x^2\right)\right), \quad (15)$$

$$\boldsymbol{y} = \alpha \times \boldsymbol{x} + \boldsymbol{u} \times \kappa_y + \epsilon_y \quad \text{with } \epsilon_y \sim \mathcal{N}\left(0, 1 - \left(\alpha^2 + \kappa_y^2 + 2 \times \alpha \times \kappa_y \times \kappa_x\right)\right). \quad (16)$$

Note that this design ensures that both $\boldsymbol{X}$ and $\boldsymbol{Y}$ have a zero mean and a variance of 1. Slightly different designs, with the same property, were used to simulate phenotypic data when the exposure was binary (case–control) and in the presence of uncorrelated or correlated pleiotropy (Supporting Information Section SE). A GWAS was performed for each sample (one for the exposure and five GWASs for the outcome, one for each sample overlap) using BGENIE (Bycroft et al., 2018). We then applied MRlap to these GWAS summary statistics to obtain the IVW-based and the corrected effect estimates.

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. $\boldsymbol{X}$ was simulated with moderate polygenicity and large heritability ($\pi_x = 0.001$ and $h_x^2 = 0.4$). $\boldsymbol{U}$ had a moderate effect on both $\boldsymbol{X}$ and $\boldsymbol{Y}$ ($\kappa_x = 0.3, \kappa_y = 0.5$), leading to a correlation (ρ) of 0.15 induced by the confounder. We varied the size of the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ from null ($\alpha = 0$) to moderate ($\alpha = 0.2$).

In addition to these standard settings, we explored various other parameter values. We investigated the effect of a confounder of the opposite sign ($\kappa_x = -0.3$ 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 considered a case–control design, where the exposure was first simulated on the liability scale before being converted to the observed scale for our analyses. For this setting, we used larger sample sizes ($n_A = 100,000$, $n_B = 100,000$) and a prevalence of 0.1 to define cases and controls. 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 ($\alpha = 0.1$). Finally, we simulated data in presence of both uncorrelated and correlated pleiotropy. For the uncorrelated pleiotropy, we used the standard settings parameters and added direct genetic effects on $\boldsymbol{Y}$ ($\pi_y = 0.002$ and $h_y^2 = 0.3$). Sixty percent of the SNPs having a direct effect on $\boldsymbol{X}$ were also directly affecting $\boldsymbol{Y}$ and their direct effects on each trait were uncorrelated. To simulate data in presence of correlated pleiotropy, we added to the model a genetic confounder ($\boldsymbol{U_g}$) acting both on $\boldsymbol{X}$ and $\boldsymbol{Y}$, with respective effects q_x and q_y. First, we modeled a genetic confounder that was highly polygenic and fairly heritable ($\pi_u = 0.0001$ and $h_u^2 = 0.2$), with moderate effects on the two traits ($q_x = 0.4$ and $q_y = 0.3$). Then we slightly increased the confounder's polygenicity and, to make the genetic confounding effect stronger, we increased its heritability and its effects on $\boldsymbol{X}$ and $\boldsymbol{Y}$ ($\pi_u = 0.0005$, $h_u^2 = 0.3$, $q_x = 0.5$, and $q_y = 0.7$).

For each scenario, IVW-based and corrected causal effects were compared for different degrees of sample overlap and different instrument selection thresholds. The results quality was assessed using root-mean-square error (RMSE), coverage, 95% confidence interval width, and power. Note that RMSE is a measure of the bias-variance trade-off, and that coverage might be influenced by the severity of the bias but also by the standard error of the estimator.

For a given instrument selection threshold, we obtained 500 causal effect estimates: one for each of the 100 simulated datasets for each of the five degrees of sample overlap (ranging from 0% to 100%). Causal effect estimates should ideally not depend on the extent of overlap between the exposure and outcome samples. To quantify the extent to which this holds, we grouped estimates according to which sample overlap degree 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 a small between-group variance relative to the variance of the estimator (characterized by the within-group variance).

Finally, we tested for differences between IVW-based and corrected effects using the following test statistic

$$t_{\text{diff}} = \frac{\hat{\alpha}_{IVW} - \hat{\alpha}_c}{\sqrt{\text{Var}(\hat{\alpha}_{\text{IVW}}) + \text{Var}(\hat{\alpha}_c) - 2 \times \text{Cov}(\hat{\alpha}_{\text{IVW}}, \hat{\alpha}_c)}}. \quad (17)$$

In addition to those scenarios, we wanted to assess the potential gain, in terms of bias and variance, arising from the possibility of performing analyses using the full UKBB sample instead of having to split it into two halves to avoid sample overlap. To do so, we simulated data using the following parameters: large polygenicity and moderate heritability ($\pi_x = 0.01$ and $h_x^2 = 0.15$), moderate confounder effect ($\kappa_x = 0.3, \kappa_y = 0.5$), and a fairly small causal effect ($\alpha = 0.1$). In this case, we only compared the IVW-based causal effect estimates from nonoverlapping samples ($n_A = 180,000, n_B = 180,000$, roughly half of UKBB sample size) and the corrected causal effect estimates from fully overlapping samples ($n_A = 360,000, n_B = 360,000$) by measuring the bias, the variance, and the RMSE.

Finally, we also compared our `MRlap` results to those of pleiotropy robust approaches, such as weighted median (Bowden, Davey Smith, et al., 2016), weighted mode (Hartwig et al., 2017), MR-RAPS (Zhao et al., 2020), and dIVW (Ye et al., 2021), both under the standard settings and in all scenarios with pleiotropy, for all degrees of sample overlap. For all methods, we use a p-value threshold of 5×10^{-8} to select IVs. Since dIVW does not directly accounts for winner's curse, the authors suggest using this estimator without instrument screening, and additional analyses using dIVW were performed without selection (a threshold of 1), using observed p values but also random p values for pruning, for nonoverlapping samples.

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 (Sudlow et al., 2015) 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 five 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 proportionally increased 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 individual 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. First, we performed "same-trait" analyses to estimate the causal effect of body mass index (BMI) on itself, and the causal effect of systolic blood pressure (SBP) on itself, using only the nonoverlapping samples. In addition, for all degrees of sample overlap, we assessed the effect of BMI on SBP, on the number of cigarettes previously smoked daily (smoking), and on alcohol intake frequency (alcohol). For smoking, answers coded as "−10" or "−1" were considered missing. For alcohol, answers were recoded to correspond to an increased intake frequency, and answers coded as "−3" were considered missing. Details about the pairs of traits analyzed are available in Table S1.

Phenotypic data were normalized (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, one for each sample overlap) using BGENIE (Bycroft et al., 2018). We then applied `MRlap` to these GWAS summary statistics to obtain the IVW-based and the corrected effect estimates. In addition, since in this case, a reverse causal effect (from the outcome on the exposure) could exist, we filtered out variants that were statistically significantly more strongly associated with the outcome than with the exposure to remove potentially invalid IVs (Steiger filter).

We repeated this sampling approach 100 times. For each repetition, IVW-based and corrected causal effects were compared using the within-group and between-group variances and we tested for differences between the IVW-based and the corrected effects using (17).

3 | RESULTS

3.1 | Overview of the method

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

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

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 (characterized by the polygenicity π_X and per variant heritability $\sigma^2_X = \frac{h^2_X}{M \times \pi_X}$, M being the number of potential IVs). 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

To compare the standard IVW approach and our method, we used simulated data for a wide range of scenarios: in the presence and in absence of a true causal effect, varying the strength of the environmental confounder, simulating correlated and uncorrelated pleiotropy, and so forth. We assessed the extent of the bias for both approaches, compared the bias-variance trade-off as well as coverage, and explored how the bias is affected by sample overlap and instrument selection threshold.

Simulation results under our standard settings show a large discrepancy between the IVW-based causal effects estimated using different degrees of sample overlap, while the corrected effects are more closely aligned with the true causal effects (Figure 2a). We observe a 10% overestimation of the causal effect for fully overlapping samples, and a 15% underestimation of the causal effect for nonoverlapping samples, due to winner's curse and weak instrument bias both biasing the estimate towards the null. As expected from (9), the bias is larger when using less stringent thresholds T. For all the thresholds tested, the ratio of the between-group and the within-groups variances is larger (up to 26 times) for IVW-based effects than for corrected effects (Table S2), highlighting the differences in IVW-based effects when estimated using different degrees of overlaps. The fact that the within-group variance is 1.3 times higher for the corrected effects compared to the uncorrected counterpart is due to the slightly increased variance of the bias-corrected estimator. The RMSE of the IVW-based effects is very dependent on the degree of overlap (being larger for nonoverlapping 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 nonoverlapping samples (Figure 2b). The corrected effects also yield a much better coverage, close to 95% for all degrees of overlap (Figure 2c). The corrected effects are statistically significantly different from the IVW-based effects for all overlaps values except 50% and all thresholds. The absolute bias of the IVW-based effects goes up to 0.033 while for the corrected effect it is smaller than 0.012 for all overlaps and thresholds.

The bias of IVW-based effects depends on the strength of the confounder. If the confounder is weak, IVW-based effects are mostly biased towards the null for low overlaps (Figure S5; Table S3). When we simulated a stronger confounder, the bias of the IVW-based effects for fully overlapping samples increased (Figure S6; Table S4). In both cases, the corrected and the IVW-based effects are statistically significantly different for almost all overlaps and thresholds, and the corrected effects are substantially less biased than the IVW-based effects.

When the confounder effect ($\rho := \kappa_x \times \kappa_y$) and the causal effect (α) have different signs (i.e., one >0 and one <0), the results are particularly interesting because winner's curse and weak instrument bias are biasing the results towards the null regardless of the sample overlap degree (Figure 3a). In this case, IVW-based effects are more similar across the different degrees of sample overlap tested, but all are underestimating the true causal effect. For this reason, we do not observe a less striking decrease in the heterogeneity of the estimates across different sample overlaps upon correction (the ratio of the between- and the within-group variance is five times larger for IVW-based effects), but still, the correction reduces RMSE and bias, as well as leads to better coverage, for all overlaps and thresholds (Figure 3; Table S5). In this scenario, IVW-based and corrected effects statistically significantly differ for all overlaps and thresholds, with an average underestimation of 13% for IVW-based effects.

In the absence of a true causal relationship between the exposure and the outcome, IVW-based effects from nonoverlapping samples are unbiased. However, for fully overlapping samples, the IVW-based effects are biased towards the confounder-induced correlation (Figure S7; Table S6). We showed that for large overlap percentages ($\geq$50%), the corrected effects are statistically significantly different from the IVW-based effects (60% smaller), and they are less biased for all overlaps and thresholds. Moreover, while the false positive rate (at a 5% level) for IVW-based effects using nonoverlapping samples is below 5%, it is much larger (between 20% and 40% depending on the threshold) when using fully overlapping samples. The correction proposed here provides much better control of the false positive rate for all degrees of sample overlap (Figure S8).

Simulating binary, instead of continuous, exposure led to very similar results. The absolute bias was on

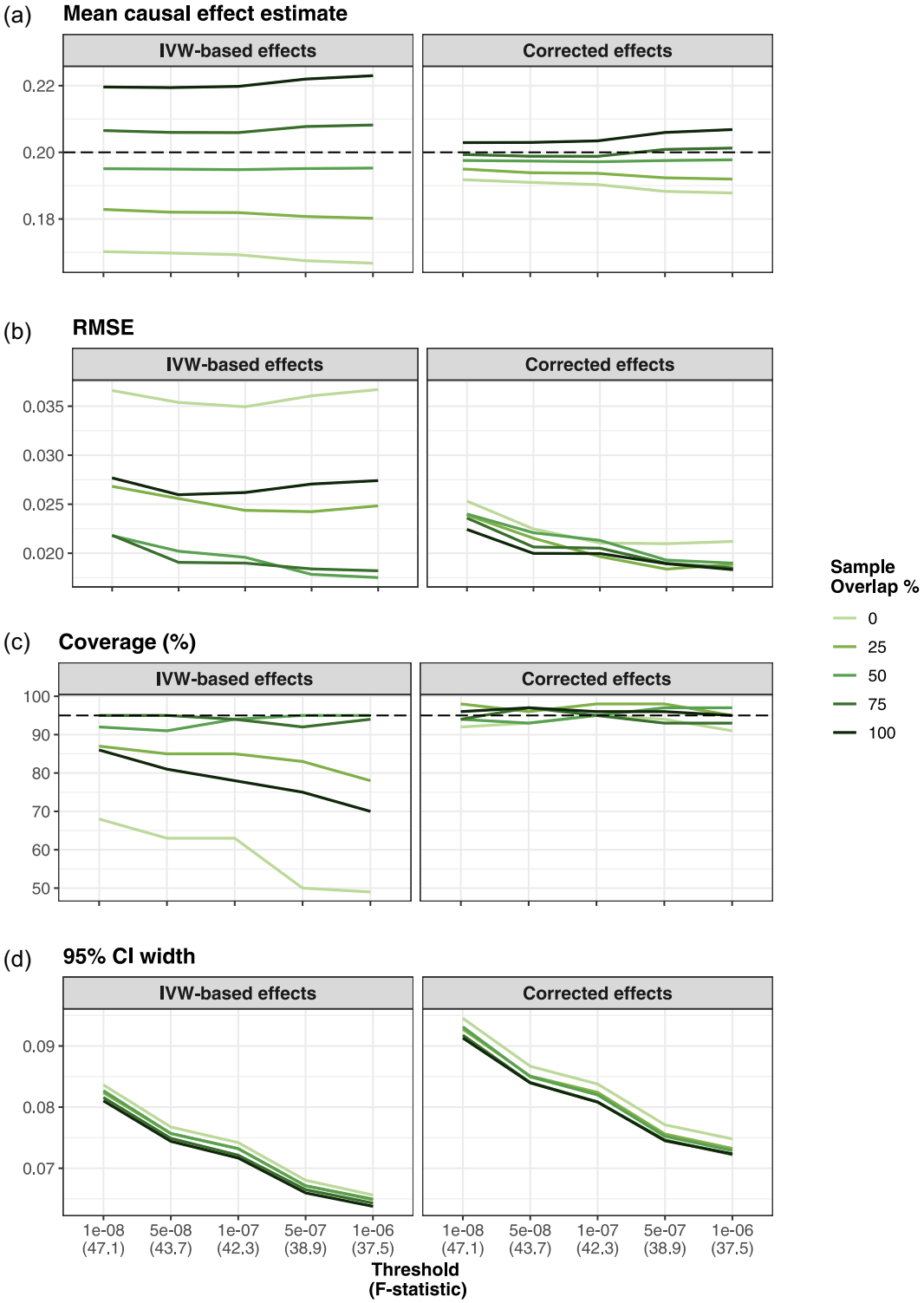

FIGURE 2 Simulation results for standard settings. $n_A = n_B = 20,000$, $\pi_x = 0.001$, $h_x^2 = 0.4$, $\kappa_x = 0.3$, $\kappa_y = 0.5$, $\alpha = 0.2$ Panel (a) shows the mean IVW-based and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represents the true causal effect). Panel (b) shows the mean RMSE obtained for IVW-based and corrected effect for each overlap and threshold. Panel (c) shows the coverage of the 95% confidence interval for IVW-based and corrected effect for each overlap and threshold. Panel (d) shows the width of the 95% confidence interval for IVW-based and corrected effect for each overlap and threshold.

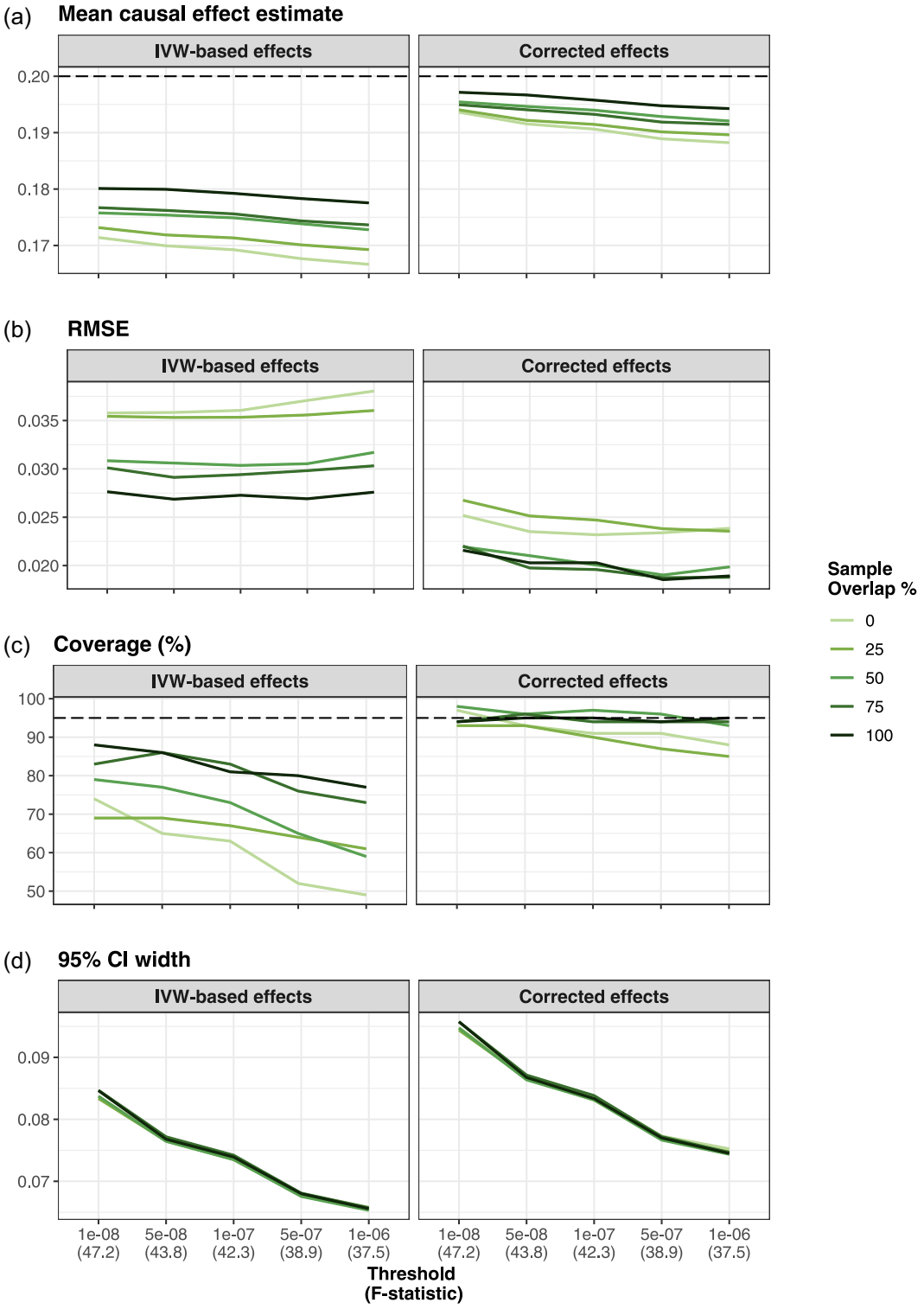

FIGURE 3 Simulation results for a scenario with a negative confounder. $n_A = n_B = 20,000$, $\pi_x = 0.001$, $h_x^2 = 0.4$, $\kappa_x = -0.3$, $\kappa_y = 0.5$, $\alpha = 0.2$ Panel (a) shows the mean IVW-based and corrected effect for each overlap and threshold obtained from 100 simulations (the dashed line represents the true causal effect). Panel (b) shows the mean RMSE obtained for IVW-based and corrected effect for each overlap and threshold. Panel (c) shows the coverage of the 95% confidence interval for IVW-based and corrected effect for each overlap and threshold. Panel (d) shows the width of the 95% confidence interval for IVW-based and corrected effect for each overlap and threshold.

average 2.3 times larger for IVW-based effects than for corrected effects and the causal effect estimates were much more consistent across varying degrees of overlap after correction (Figure S9; Table S7). Results obtained using more realistic parameters in terms of sample sizes, genetic architecture, and causal effect strength show a similar pattern. We observe an important bias of IVW-based causal effects, mostly when estimated from nonoverlapping (22% underestimation) or fully overlapping samples (30% overestimation), that is strongly reduced when using our correction (Figure S10; Table S8). Corrected effects statistically significantly differ from IVW-based effects for the most extreme overlaps values (0%, 75%, 100%) for which corrected effects are on average five times less biased than IVW-based effects.

We compared corrected effects obtained using the full (overlapping) sample to IVW-based effects obtained by splitting it into two halves to avoid sample overlap. We showed that the IVW estimators were 3.27 times more biased than those from `MRlap` and the latter reduced the estimator variance by more than threefold (3.43) thanks to the elevated sample size. In addition, the corrected effects estimates had better coverage (88% vs. 72%) and higher power (100% vs. 94%). Thus, we have demonstrated that applying `MRlap` to the full (overlapping) sample is a better strategy than ensuring no sample overlap and applying the IVW estimator because it is less biased and has considerably lower variance (Table S9).

We also investigated the extent of the bias, for both IVW-based and corrected effects, in scenarios with either uncorrelated pleiotropy (Figure S11; Table S10), moderately correlated pleiotropy (Figure S12; Table S11) and strongly correlated pleiotropy (Figure S13; Table S12). Results obtained in presence of uncorrelated pleiotropy are very similar to the ones obtained for the standard settings, with slightly larger within-group variances. MR-RAPS and dIVW, however, did not perform particularly well, even in the standard settings scenario, potentially because the instruments are relatively strong and their estimates are therefore most affected by winner's curse. Using a third sample to select instruments and avoid winner's curse was out of the scope of the paper, but we did try to use alternative selection approaches for dIVW, focusing on nonoverlapping samples (Figure S14). We ignored the selection step (only pruning the SNPs to obtain independent instruments, but not using any selection threshold) to reduce the impact of winner's curse, as suggested in Ye et al. (2021). Surprisingly, the causal effect estimates for both methods were more strongly biased than when using a p value threshold of 5e−08. We believe that is due to the pruning step (keeping the most strongly associated instrument in each region), as using random p values instead of observed p values for pruning yielded unbiased estimates (at the cost of much larger variance). In presence of correlated pleiotropy, we observe a bias towards the ratio of the genetic confounder effects on $\boldsymbol{Y}$ and $\boldsymbol{X}$ (0.75 for moderately correlated pleiotropy and 1.4 for strongly correlated pleiotropy) for both IVW-based and corrected effects for all sample overlap degrees. This can be explained by the fact that there is now a third source of bias, correlated pleiotropy, affecting the causal effect estimates. This source of bias is independent of sample overlap, and corrected effect estimates are able to recover consistent causal effect estimates across varying degrees of overlap, corresponding to the sum of the true causal effect and of the bias induced by correlated pleiotropy. When comparing our results to results obtained using pleiotropy robust methods (Figure 4), we observe that for nonoverlapping samples, weighted median, weighted mode, and MR-RAPS estimates are able to recover the value of the IVW-based estimate from the standard settings scenarios (i.e., a downward biased causal effect estimate). However, for large degrees of sample overlap (≥50%), these approaches are biased toward the observational correlation, and their causal effect estimates are strongly overlap-dependent.

3.3 | Application to UKBB

We tested our method on UKBB obesity-related exposures using a similar approach and splitting the full dataset into samples of varying degrees of overlap. We started by estimating the causal effect of BMI on BMI and the causal effect of SBP on SBP as these are expected to be equal to 1. In this case, we only looked at nonoverlapping samples and compared the IVW-based and the corrected effects to the true expected effect (Figure 5). IVW-based effects are biased towards the null for all thresholds (95% confidence intervals do not include 1, coverage between 0% and 18%), with the bias being stronger for less stringent thresholds. Corrected effects however were less biased (at the cost of a slightly higher variance) and statistically nonsignificantly different from 1 for all thresholds (coverage of between 65% and 96%), illustrating the importance of correcting for weak-instrument bias and winner's curse even in the absence of sample overlap.

When looking at the IVW-based effect of BMI on SBP (Figure 6a), 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

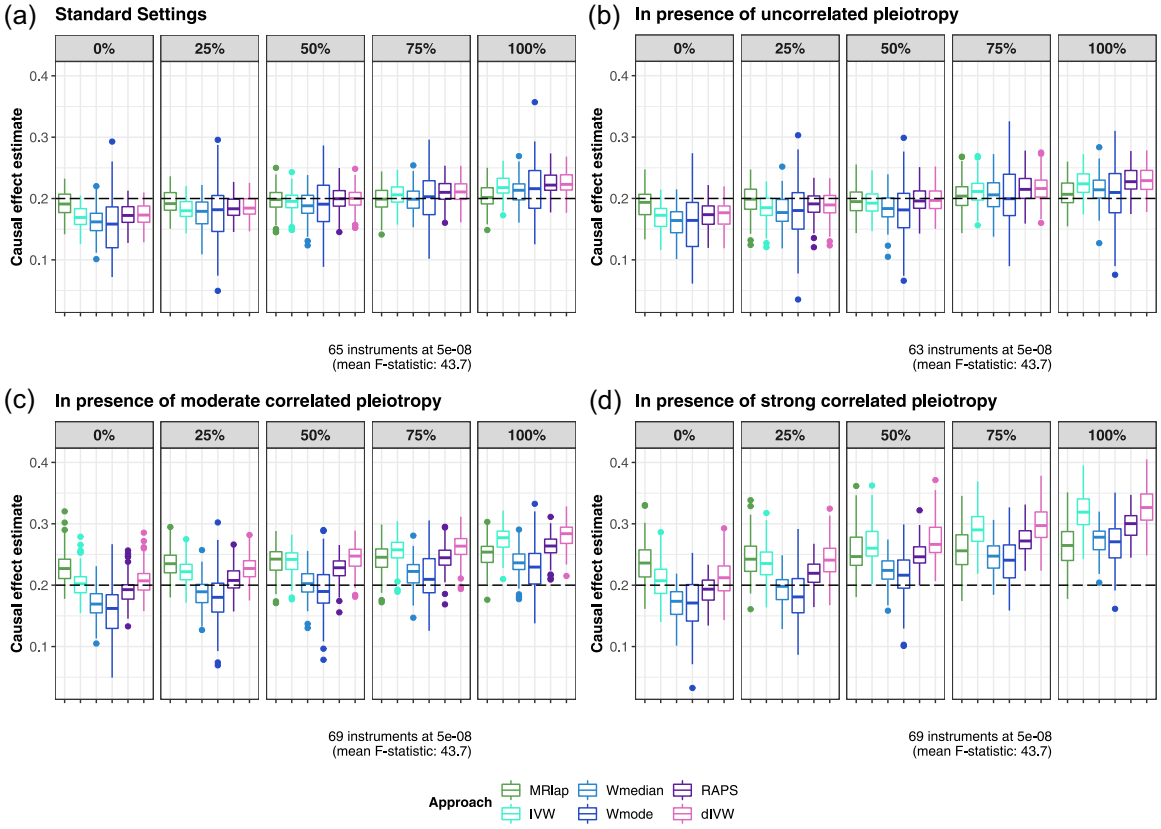

FIGURE 4 Comparison of different MR approaches. Causal effects estimates were obtained from 100 simulations using six different methods (MRlap in green, IVW in turquoise blue, weighted median in light blue, weighted mode in dark blue, MR-RAPS in purple, and dIVW in pink). The dashed line represents the true causal effect. Panel (a) shows results for the standard settings scenario (no pleiotropy). Panel (b) shows results in presence of uncorrelated pleiotropy. Panel (c) shows results in presence of moderately correlated pleiotropy. Panel (d) shows results in presence of strongly correlated pleiotropy. The average number of instruments and mean F-statistic (at 5e−08) are indicated for each scenario.

nonoverlapping samples the IVW-based effect is larger for less stringent thresholds. This is inconsistent with winner's curse and weak instrument bias that leans the estimate towards the null and we would expect the IVW-based 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 subtype or the presence of a heritable confounder (see Section 4). Here, we will focus on the results obtained using a p value threshold of 5e−8.

When using IVs reaching genome-wide significance, the IVW-based effects range between 0.095, for non-overlapping samples, and 0.129, for fully overlapping samples. After correction, the range of the estimated effect is about two times smaller (0.108–0.128). 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 reduced fourfold upon correction (Table S13). For nonoverlapping samples, the difference between the IVW-based and corrected effect is statistically significant ($p_{\text{diff}} = .0049$), and standard two-sample MR underestimates the causal effect by about 25% compared to the corrected effect (the IVW-based effect is 0.095 while the corrected effect is 0.126). For fully overlapping samples, there is no statistically significant difference between IVW-based and corrected effects. We observed a similar pattern when investigating the effect of BMI on smoking (Figure S15; Table S14) where the largest bias occurs for nonoverlapping samples. The IVW-based causal effect at $p = 5e{-}8$ is 0.130 for nonoverlapping samples while results after correction point towards a causal effect of 0.170 (underestimation of 24%—statistically significant difference between IVW-based and corrected effects, $p_{\text{diff}} = .0137$). We do not see a statistically significant difference between IVW-based 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%.

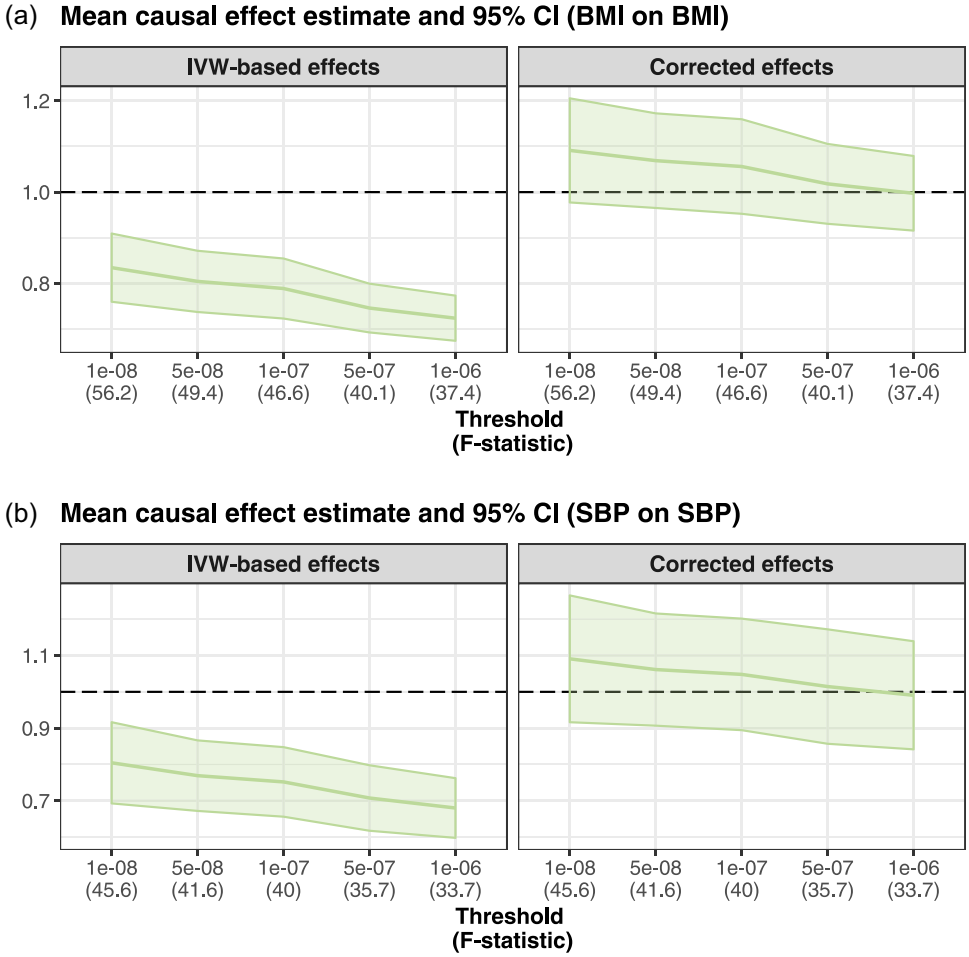

FIGURE 5 Effect of BMI on BMI and effect of SBP on SBP. This figure shows the mean IVW-based and corrected effect (and 95% confidence interval) for each threshold obtained from 100 different sampled datasets, using nonoverlapping samples. The true causal effect is expected to be 1, as represented by the dashed line. Panel (a) corresponds to the effect estimates of BMI on BMI. Panel (b) corresponds to the effect estimates of SBP on SBP.

Our results implicate the existence of an environmental confounder biasing the causal effect estimate of BMI on alcohol intake frequency. At $p = 5e{-}8$, IVW-based effects range between −0.187 for nonoverlapping samples and −0.218 for fully overlapping samples, whereas the corrected effects are larger (−0.244 to −0.2515) (Figure 6b; Table S15). The between-group variance is much smaller for the corrected effects. The difference between IVW-based and corrected effects is statistically significant for all overlaps and the corrected effects being stronger than the IVW-based ones hints at the existence of an environmental confounder having a concordant effect on BMI and alcohol intake frequency, biasing all estimates towards 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 observational correlation between BMI and alcohol intake frequency (−0.13 among the 379,530 genetically British individuals in the UKBB) is weaker than the standardized causal effect.

4 | DISCUSSION

We developed a method that models the entire instrument selection process, sample overlap, and exposure effect estimation error. As a result, our approach reduces winner's curse and weak instrument bias, even when the degree of overlap is unknown. Descriptive work has recently focused on a single relationship, investigating the effect of winner's curse in the estimation of the causal effect of BMI on coronary artery disease, including comparisons of the effect estimates for nonoverlapping and fully overlapping samples (Jiang et al., 2022). To the best of our knowledge, while some empirical work describing the interplay between these biases has been done (Sadreev et al., 2021), no other method can tackle

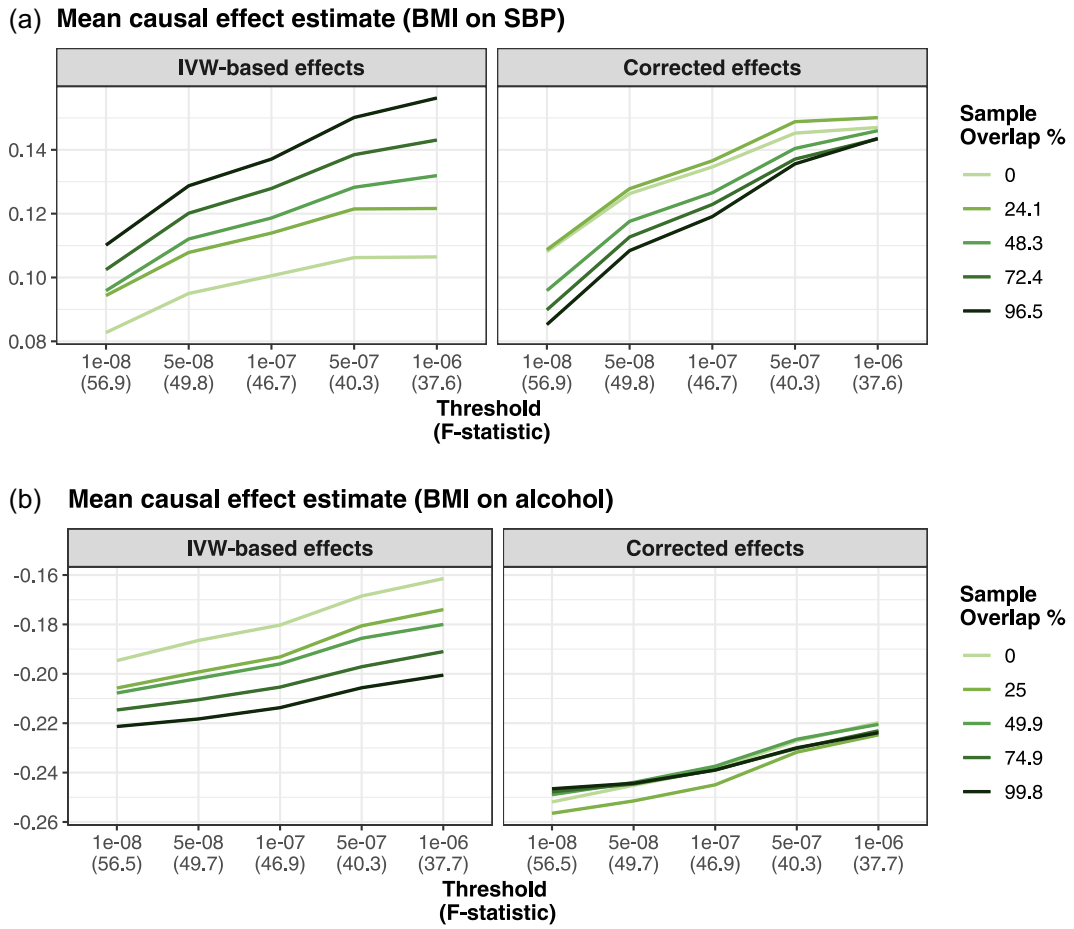

FIGURE 6 Effect of BMI on SBP and effect of BMI on alcohol intake frequency. This figure shows the mean IVW-based and corrected effect for each overlap and threshold obtained from 100 different sampled datasets. Panel (a) corresponds to the effect estimates of BMI on SBP. Panel (b) corresponds to the effect estimates of BMI on alcohol.

all the aforementioned biases 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 nonoverlapping and fully overlapping samples can be biased. The direction and the magnitude of the bias depend on sample overlap and are strongly influenced by the effect of the confounder. When the confounder and the causal effect have the same sign, observed effects are overestimated for fully overlapping samples and underestimated for nonoverlapping samples. However, when they have opposite signs, these are underestimated for all overlaps because the direction of the biases is towards the null for any degree of sample overlap. 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 permits drastically reducing the bias. For standard settings for example, we observed a 15% overestimation for fully overlapping samples and a 10% underestimation for nonoverlapping samples, which were respectively reduced to a 5% overestimation and a 2% underestimation after correction. We also found statistically significant differences between IVW-based and corrected effects for fully overlapping samples under all scenarios. For nonoverlapping samples, IVW-based and corrected effects were statistically significantly different under all scenarios except in the case of the absence of a causal effect. The decreased bias leads to better coverage of the 95% confidence interval, but the correction also comes with increased variance. Still, in all of our simulation scenarios, the correction yielded reduced estimation error (RMSE) for at least one, if not all, sample overlap degrees. Moreover, while the RMSE of IVW-based 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.

In the simulations, we mostly compared our corrected effect to its IVW-based counterpart, and only considered methods that can deal with weak instruments, such as MR-RAPS (Zhao et al., 2020) and dIVW (Ye et al., 2021), for a small number of scenarios. Comparisons with these approaches are not optimal, as they require a third sample to be robust to winner's curse, and are expected to perform better using a larger set of weaker instruments. We only investigated `MRlap` performance for thresholds smaller than 1e−06 since using less stringent thresholds will increase the chances of using IVs that would violate the relevance assumption. We believe that the thresholds that have been used, both for simulations and real data analyses are realistic and that more lenient thresholds are unlikely to be used for IVW-MR in practice.

For real data, we used a sampling strategy to compare results obtained using varying degrees of sample overlap. We first focused on "same-trait" (BMI on BMI and SBP on SBP) analyses for which the true causal effect is expected to be 1, using nonoverlapping samples. IVW-based effects were strongly biased towards the null (between 18% and 30% depending on the threshold) and it is important to note that the 95% confidence intervals did not overlap with one for any of the traits. Corrected effects were less biased and their confidence intervals were overlapping with one (partly because of the lower precision of the estimator). We observed a slight overcorrection for the most stringent thresholds that could be due to potential violations of our assumptions regarding the genetic architecture of the exposure (spike-and-slab distribution). For the three other relationships we looked at, strong discrepancies were observed for low degrees of overlap, with statistically significant differences between IVW-based and corrected effects. This means that standard two-sample MR settings often lead to an underestimation of the true causal effect, which 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, the bias is often much less substantial for higher degrees of sample overlap. Among our many examples, we found that the IVW-MR estimate for the effect of BMI on alcohol intake frequency using the fully overlapping samples is biased by a confounder. In this case, the confounder and the causal effect had opposite signs, leading to an underestimation of the IVW-based effect for all overlaps. We have also highlighted that there is 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 strong phenotypic heterogeneity in the exposure, in which case different groups of IVs could be affecting the exposure through different pathways (Foley et al., 2020). Alternatively, in the presence of a genetic confounder, IVs picked up at a less stringent threshold may be associated with a confounder, hence violating the second assumption of MR. Such a 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 account for this genetic confounding (CAUSE [Morrison et al., 2020]; LHC-MR [Darrous et al., 2021]) or others allowing for multiple causal effects (MR-Clust [Foley et al., 2020]) 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, as shown in our simulations with correlated pleiotropy. 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. It is also important to note that our work focused on continuous traits, and our approach would only work using case–control designs if the sample overlap degree does not differ between cases and controls. A simplification of the simulated model is that we assumed only a single confounder, but the bias estimation does not depend on this assumption. Finally, we have not explicitly modeled the local LD in the IV selection process, whereby a small winner's curse bias may be introduced when selecting the SNP with the strongest effect (at a given locus) as the IV.

Nowadays, samples from large biobanks are often used to estimate SNP effect sizes for both the exposure and the outcome, and hence it will be less and less possible to ensure that the two samples used are not overlapping. Thus, the need for nonoverlapping samples forces researchers to use summary statistics from reduced sample sizes. Most published MR analyses go to great lengths to ensure nonoverlapping samples are used, for example, Davies et al. (2019), Cornish et al. (2020), Yang et al. (2022), Brumpton et al. (2020) (one of their seven estimates split the sample to avoid sample overlap). Sample overlap is a key point in the STROBE guidelines for MR (Skrivankova et al., 2021), in item 10d. It is also included in rule #7 in a popular MR guideline (Taliun & Evans, 2021). Avoiding sample overlap remains the predominant approach in the MR field, without major attempts to quantify the extent of bias it gives rise to. While our results have shown that for nonoverlapping samples the biases usually do not

strongly reduce power or change the clinical conclusions, which is in line with results from (Jiang et al., 2022), not accounting for these biases can still dramatically decrease coverage. We believe that our approach is of particular interest when it is not possible to use nonoverlapping samples (for example, exposure and outcome only measured in a specific cohort). In this case, weak instrument bias and winner's curse could increase the false positive rate quite drastically if there is a strong observational correlation. For these reasons, 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 statistically significantly differ from the IVW-based effect, then the IVW-MR estimate can be safely used (with the advantage of having lower variance). However, if there is a statistically significant difference, corrected effects should be preferred as they are less biased, independently of sample overlap.

AUTHOR CONTRIBUTIONS

Zoltán Kutalik designed and supervised the project. Zoltán Kutalik and Ninon Mounier developed the methods. Ninon Mounier performed the analyses. Ninon Mounier and Zoltán Kutalik wrote the manuscript.

ACKNOWLEDGMENTS

The authors thank Chiara Auwerx, Liza Darrous, Sven Erik Ojavee, Marion Patxot, Eleonora Porcu, Marie Sadler, and Tabea Schoeler for the helpful discussions and constructive comments. Computations have been performed on the HPC cluster of the Lausanne University Hospital. This research has been conducted using the UK Biobank Resource under Application Number 16389. Zoltán Kutalik was funded by the Swiss National Science Foundation (# 310030-189147) and the Department of Computational Biology (UNIL). Open access funding provided by Universite de Lausanne.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Ninon Mounier http://orcid.org/0000-0001-6663-5402

REFERENCES

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, *44*, 512–525. https://doi.org/10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., & Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, *40*(4), 304–314. https://doi.org/10.1002/gepi.21965

Bowden, J., Del Greco, F. M., Minelli, C., Davey Smith, G., Sheehan, N., & Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine*, *36*, 1783–1802. https://doi.org/10.1002/sim.7221

Bowden, J., Del Greco, F. M., Minelli, C., Davey Smith, G., Sheehan, N. A., & Thompson, J. R. (2016). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-egger regression: The role of the I^2 statistic. *International Journal of Epidemiology*, *45*, 1961–1974. https://doi.org/10.1093/ije/dyw220

Bowden, J., Del Greco, F. M., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., Thompson, J., & Davey Smith, G. (2018). Improving the accuracy of two-sample summary-data Mendelian randomization: Moving beyond the NOME assumption. *International Journal of Epidemiology*, *48*(3), 728–742. https://doi.org/10.1093/ije/dyy258

Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. Å., Cho, Y., Howe, L. D., Hughes, A., Boomsma, D. I., Havdahl, A., Hopper, J., Neale, M., Nivard, M. G., Pedersen, N. L., Reynolds, C. A., Tucker-Drob, E. M., Grotzinger, A., ... Howe, L. (2020). Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications*, *11*(1):3519. https://doi.org/10.1038/s41467-020-17117-4

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium, Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., & Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, *47*, 1236–1241. https://doi.org/10.1038/ng.3406

Bulik-Sullivan, B., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*, 291–295. https://doi.org/10.1038/ng.3211

Burgess, S., Butterworth, A., & Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, *37*, 658–665. https://doi.org/10.1002/gepi.21758

Burgess, S., Davies, N. M., & Thompson, S. G. (2016). Bias due to participant overlap in two-sample Mendelian randomization. *Genetic Epidemiology*, *40*, 597–608. https://doi.org/10.1002/gepi.21998

Burgess, S., & Thompson, S. G. (2011a). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, *40*, 755–764. https://doi.org/10.1093/ije/dyr036

Burgess, S., & Thompson, S. G. (2011b). Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine*, *30*, 1312–1323. https://doi.org/10.1002/sim.4197

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature*, *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z

Cornish, A. J., Law, P. J., Timofeeva, M., Palin, K., Farrington, S. M., Palles, C., Jenkins, M. A., Casey, G., Brenner, H., Chang-Claude, J., Hoffmeister, M., Kirac, I., Maughan, T., Brezina, S., Gsur, A., Cheadle, J. P., Aaltonen, L. A., Tomlinson, I., Dunlop, M. G., & Houlston, R. S. (2020). Modifiable pathways for colorectal cancer: A Mendelian randomisation analysis. *The Lancet Gastroenterology & Hepatology*, *5*(1), 55–62. https://doi.org/10.1016/s2468-1253(19)30294-8

Darrous, L., Mounier, N., & Kutalik, Z. (2021). Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. *Nature Communications*, *12*(1):7274. https://doi.org/10.1038/s41467-021-26970-w

Davies, N. M., Hill, W. D., Anderson, E. L., Sanderson, E., Deary, I. J., & Smith, G. D. (2019). Multivariable two-sample Mendelian randomization estimates of the effects of intelligence and education on health. *eLife*, *8*, e43990. https://doi.org/10.7554/elife.43990

Foley, C. N., Mason, A. M., Kirk, P. D. W., & Burgess, S. (2020). MR-Clust: Clustering of genetic variants in Mendelian randomization with similar causal estimates. *Bioinformatics*, *37*, 531–541. https://doi.org/10.1093/bioinformatics/btaa778

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., Koellinger, P. D., Harden, K. P., Nivard, M. G., & Tucker-Drob, E. M. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour*, *3*, 513–525. https://doi.org/10.1038/s41562-019-0566-x

Hartwig, F. P., Davey Smith, G., & Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, *46*, 1985–1998. https://doi.org/10.1093/ije/dyx102

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Davey Smith, G., Gaunt, T. R., & Haycock, P. C. (2018). The MR-base platform supports systematic causal inference across the human phenome. *eLife*, *7*, e34408. https://doi.org/10.7554/elife.34408

International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*, 52–58. https://doi.org/10.1038/nature09298

Jiang, T., Gill, D., Butterworth, A. S., & Burgess, S. (2022). An empirical investigation into the impact of winner's curse on estimates from Mendelian randomization. *International Journal of Epidemiology*. https://doi.org/10.1093/ije/dyac233

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, *27*, 1133–1163. https://doi.org/10.1002/sim.3034

Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., & He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, *52*, 740–747. https://doi.org/10.1038/s41588-020-0631-4

Palmer, C., & 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. https://doi.org/10.1371/journal.pgen.1006916

Sadreev, I. I., Elsworth, B. L., Mitchell, R. E., Paternoster, L., Sanderson, E., Davies, N. M., Millard, L. A., Davey Smith, G., Haycock, P. C., Bowden, J., Gaunt, T. R., & Hemani, G. (2021). Navigating sample overlap, winner's curse and weak instrument bias in Mendelian randomization studies using the UK biobank. *medRxiv*. https://doi.org/10.1101/2021.06.28.21259622

Skrivankova, V. W., Richmond, R. C., Woolf, B. A. R., Davies, N. M., Swanson, S. A., VanderWeele, T. J., Timpson, N. J., Higgins, J. P. T., Dimou, N., Langenberg, C., Loder, E. W., Golub, R. M., Egger, M., Smith, G. D., & Richards, J. B. (2021). Strengthening the reporting of observational studies in epidemiology using Mendelian randomisation (STROBE-MR): Explanation and elaboration. *BMJ*, *375*, n2233. https://doi.org/10.1136/bmj.n2233

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (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*, *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779

Taliun, S. A. G., & Evans, D. M. (2021). Ten simple rules for conducting a Mendelian randomization study. *PLOS Computational Biology*, *17*(8), e1009238. https://doi.org/10.1371/journal.pcbi.1009238

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS discovery: Biology, function, and translation. *American Journal of Human Genetics*, *101*, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Yang, Q., Sanderson, E., Tilling, K., Borges, M. C., & Lawlor, D. A. (2022). Exploring and mitigating potential bias when genetic instrumental variables are associated with multiple non-exposure traits in Mendelian randomization. *European Journal of Epidemiology*, *37*(7), 683–700. https://doi.org/10.1007/s10654-022-00874-5

Ye, T., Shao, J., & Kang, H. (2021). Debiased inverse-variance weighted estimator in two-sample summary-data Mendelian randomization. *The Annals of Statistics*, *49*(4), 2079–2100. https://doi.org/10.1214/20-aos2027

Zeng, J., de Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F.,

Powell, J. E., Montgomery, G. W., Metspalu, A., Esko, T., Gibson, G., Wray, N. R., Visscher, P. M., & Yang, J. (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics*, *50*, 746–753. https://doi.org/10.1038/s41588-018-0101-4

Zhao, Q., Chen, Y., Wang, J., & Small, D. S. (2019). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *International Journal of Epidemiology*, *48*(5), 1478–1492. https://doi.org/10.1093/ije/dyz142

Zhao, Q., Wang, J., Hemani, G., Bowden, J., & Small, D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, *48*, 1742–1769. https://doi.org/10.1214/19-aos1866

Zheng, J., Baird, D., Borges, M.-C., Bowden, J., Hemani, G., Haycock, P., Evans, D. M., & Davey Smith, G. (2017). Recent developments in Mendelian randomization studies. *Current Epidemiology Reports*, *4*, 330–345. https://doi.org/10.1007/s40471-017-0128-6

Zhong, H., & Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, *9*, 621–634. https://doi.org/10.1093/biostatistics/kxn001

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mounier, N., & Kutalik, Z. (2023). Bias correction for inverse variance weighting Mendelian randomization. *Genetic Epidemiology*, 47, 314–331. https://doi.org/10.1002/gepi.22522