Barry, C. S., Liu, J., Richmond, R., Lawlor, D. A., Rutter, M. K., Dudbridge, F., & Bowden, J. (2021). Exploiting collider bias to apply two-sample summary data Mendelian randomization methods to one-sample individual level data. *PLoS Genetics*, *17*(8), [e1009703]. https://doi.org/10.1371/journal.pgen.1009703

PLOS GENETICS

METHODS

Exploiting collider bias to apply two-sample summary data Mendelian randomization methods to one-sample individual level data

Ciarrah Barry[1,2], Junxi Liu[1,2], Rebecca Richmond[1,2], Martin K. Rutter[3,4], Deborah A. Lawlor[1,2], Frank Dudbridge[5], Jack Bowden[6,1]*

1 MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom, **2** Population Health Sciences, University of Bristol, Bristol, United Kingdom, **3** Division of Diabetes, Endocrinology and Gastroenterology, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom, **4** Diabetes, Endocrinology and Metabolism Centre, Manchester University NHS Foundation Trust, Manchester, United Kingdom, **5** Department of Health Sciences, University of Leicester, Leicester, United Kingdom, **6** Exeter Diabetes Group (ExCEED), College of Medicine and Health, University of Exeter, Exeter, United Kingdom

* j.bowden2@exeter.ac.uk

Abstract

Over the last decade the availability of SNP-trait associations from genome-wide association studies has led to an array of methods for performing Mendelian randomization studies using only summary statistics. A common feature of these methods, besides their intuitive simplicity, is the ability to combine data from several sources, incorporate multiple variants and account for biases due to weak instruments and pleiotropy. With the advent of large and accessible fully-genotyped cohorts such as UK Biobank, there is now increasing interest in understanding how best to apply these well developed summary data methods to individual level data, and to explore the use of more sophisticated causal methods allowing for non-linearity and effect modification.

In this paper we describe a general procedure for optimally applying any two sample summary data method using one sample data. Our procedure first performs a meta-analysis of summary data estimates that are intentionally contaminated by collider bias between the genetic instruments and unmeasured confounders, due to conditioning on the observed exposure. These estimates are then used to correct the standard observational association between an exposure and outcome. Simulations are conducted to demonstrate the method's performance against naive applications of two sample summary data MR. We apply the approach to the UK Biobank cohort to investigate the causal role of sleep disturbance on HbA1c levels, an important determinant of diabetes.

Our approach can be viewed as a generalization of Dudbridge et al. (*Nat. Comm*. **10**: 1561), who developed a technique to adjust for index event bias when uncovering genetic predictors of disease progression based on case-only data. Our work serves to clarify that in any one sample MR analysis, it can be advantageous to estimate causal relationships by artificially inducing and then correcting for collider bias.

OPEN ACCESS

Citation: Barry C, Liu J, Richmond R, Rutter MK, Lawlor DA, Dudbridge F, et al. (2021) Exploiting collider bias to apply two-sample summary data Mendelian randomization methods to one-sample individual level data. PLoS Genet 17(8): e1009703. https://doi.org/10.1371/journal.pgen.1009703

Editor: Heather J Cordell, Newcastle University, UNITED KINGDOM

Received: December 1, 2020

Accepted: July 8, 2021

Published: August 9, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1009703

Data Availability Statement: This research has been conducted using the UK Biobank Resource under application 6818. The individual level data underlying the results presented in the study are available to qualified researchers from UK Biobank (https://www.ukbiobank.ac.uk/). Summary statistics gleaned from the individual level data and used by our Collider-Correction algorithm in applied analyses (a) and (b) are provided in S1 Data, as well as R code to aid their extraction.

Funding: D.A.L., J.L. and R.R. all work in a Unit that receives support from the University of Bristol and UK Medical Research Council (MCUU00011/6). C.B. is supported by a Wellcome Trust studentship (218495/Z/19/Z). J.L. is funded by a Diabetes UK project grant (17/0005700). D.A.L. is a National Institute of Research Senior Investigator (NF-0616-10102). F.D. is supported by the MRC (MR/S037055/1). J.B. is funded by an Expanding Excellence in England (E3) research award. D.A.L. is funded by an award from the National Institute for Health Research (NF-0616-10102). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Uncovering causal mechanisms between risk factors and disease is challenging with observational data because of unobserved confounding. Mendelian randomization offers a potential solution by replacing an individual's observed risk factor data with an unconfounded genetic proxy measure. Over the last decade an array of methods for performing Mendelian randomization studies (MR) using publicly available summary statistics gleaned from two separate genome-wide association studies. With the advent of large and accessible fully-genotyped cohorts such as UK Biobank, there is now increasing interest in understanding how best to apply these well-developed summary data methods to individual level data. In this paper we describe a general procedure for optimally applying any summary data MR method using individual level data from one cohort study. Our approach may at first seem nonsensical: we create summary statistics that are intentionally biased by confounding. This bias can, however, be very accurately estimated, and the estimate then used to correct the results of a standard observational analysis. We apply our new way of performing an MR analysis to data from UK Biobank to investigate the causal role of sleep disturbance on HbA1c levels, an important determinant of diabetes.

Introduction

Mendelian randomisation (MR) is a technique used to test for, and quantify, the causal relationship between a modifiable exposure and health outcome with observational data, by using genetic variants as instrumental variables [1, 2]. MR circumvents the need to measure and adjust for all variables which confound the exposure-outcome association, and is therefore seen as an attractive additional analysis to perform alongside more traditional epidemiological methods [3]. The following Instrumental Variable assumptions are usually invoked in order justify testing for a causal effect of an exposure X on a health outcome Y using a set of genes, G:

- IV1: G must be associated with X;
- IV2: G must be independent of unmeasured confounding between X and Y;
- IV3: G must be independent of Y conditional on X and all confounders of the X-Y relationship.

These assumptions are encoded in the causal diagram in Fig 1. Further linearity and homogeneity assumptions are needed in order to consistently estimate the magnitude of the causal effect. When performing an MR-analysis it is best practice to pre-select SNPs for use as instruments using external data, in order to avoid bias due to the winner's curse [4]. Subsequently, if the genetic variants are not as strongly associated with the exposure as in the discovery GWAS, assumption IV1 will only be weakly satisfied, which leads to so-called weak instrument bias [5, 6]. This issue is mitigated as the sample size increases as long as the true association is non-zero. When a genetic variant is in fact associated with the outcome through pathways other than the exposure, a phenomenon known as horizontal pleiotropy [7], this is a violation of assumptions IV2 and/or IV3. Horizontal pleiotropy is not necessarily mitigated by an increasing sample size and is also harder to detect. Its presence can therefore render very precise MR estimates hopelessly biased. Pleiotropy-robust MR methods have been a major focus of research in recent years for this reason [8–11].

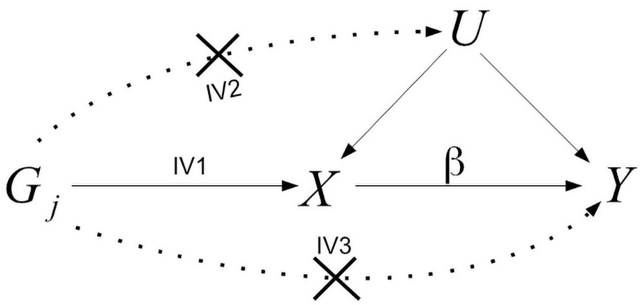

Fig 1. The IV assumptions for a genetic variant *G* are represented by solid lines in the directed acyclic graph (DAG). Dotted lines represent violations of IV assumptions as described in IV2 and IV3. The causal effect of a unit increase of the exposure, *X*, on the outcome, *Y*, is denoted by *β*. *U* represents unobserved confounders of *X* and *Y*.

https://doi.org/10.1371/journal.pgen.1009703.g001

One-sample versus Two-sample MR: Pros and cons

Obtaining access to a single cohort with measured genotype, exposure and outcome data that is large enough to furnish an MR analysis has been difficult, historically. It has instead been far easier to obtain summary data estimates of gene-exposure and gene-outcome associations from two independent studies, and to perform an analysis within the 'two-sample summary data MR' framework (see Fig 2). [12, 13]. This has made it an attractive option for the large scale pursuit of MR, through software platforms such as MR-Base [14]. The relative simplicity of these methods (which resemble a standard meta-analysis of study results) and their ability to furnish graphical summaries for the detection and adjustment of pleiotropy [15] has also acted to increase their popularity. Indeed, the array of pleiotropy robust two sample summary data methods far outstrips those available for one sample individual level data MR analysis [16]. A further advantage of two-sample over one-sample MR is that weak instruments bias causal estimates towards the null (it is often referred to as a 'dilution' bias for this reason) which is conservative [17]. Dilution bias arises precisely because uncertainty in the SNP-exposure association estimates obtained from one cohort is independent of the uncertainty in

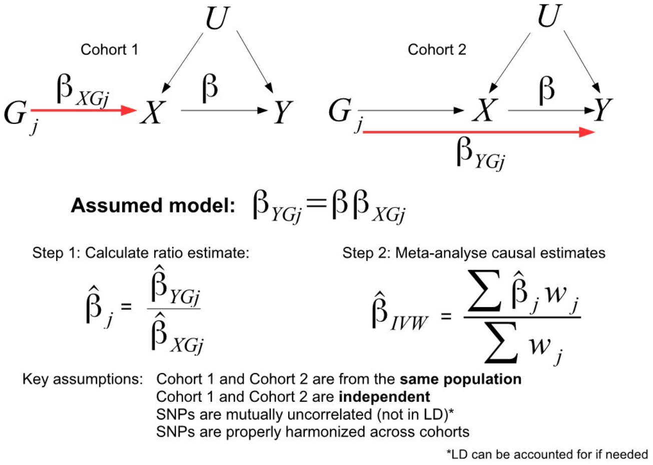
Cohort 1
U
β_{XGj}
G_j
X
β
Y
Cohort 2
U
G_j
X
β
Y
β_{YGj}
Assumed model: $\beta_{YGj} = \beta\beta_{XGj}$
Step 1: Calculate ratio estimate:
$\hat{\beta}_j = \frac{\hat{\beta}_{YGj}}{\hat{\beta}_{XGj}}$
Step 2: Meta-analyse causal estimates
$\hat{\beta}_{IVW} = \frac{\sum \hat{\beta}_j w_j}{\sum w_j}$
Key assumptions: Cohort 1 and Cohort 2 are from the **same population**
Cohort 1 and Cohort 2 are **independent**
SNPs are mutually uncorrelated (not in LD)*
SNPs are properly harmonized across cohorts
*LD can be accounted for if needed

Fig 2. In two sample summary data MR, (*G* – *X*) association estimates, $\hat{\beta}_{XGj}$, from one cohort are combined with (*G* – *Y*) association estimates, $\hat{\beta}_{YGj}$ from a separate, non-overlapping cohort, to produce a set of SNP-specific causal estimates, $\hat{\beta}_j$. These are combined using inverse variance weighted meta-analysis (w_j being the weight) to obtain an overall estimate $\hat{\beta}_{IVW}$ for the true causal effect *β*.

https://doi.org/10.1371/journal.pgen.1009703.g002

SNP-outcome association estimates from a non-overlapping cohort (Fig 2). This makes the the SNP-exposure association uncertainty akin to 'classical' measurement error [18] and enables standard approaches such as Simulation Extrapolation [19, 20] or modified weighting [6, 21] to be used to adjust for its presence. In contrast, weak instruments bias MR estimates obtained from a one sample analysis towards the observational association because uncertainty in the SNP-exposure and SNP-outcome association estimates are correlated. This bias is harder to correct for and is potentially anti-conservative.

There are, however, many disadvantages of using two sample summary data compared to individual level data from a single sample MR, some examples of which are now given: The two-sample approach assumes the two cohorts are perfectly homogeneous [13]. If the distribution of confounders is different between the samples, this can result in severe bias [22]. Alternatively, it may be that the independence assumption is violated due to an unknown number of shared subjects across the two studies [23], which cannot be easily removed [24]. Even when the homogeneity assumption is satisfied, two sample methods can give misleading results if the two sets of associations are not properly harmonized [25]. Often, summary statistics from a GWAS have been adjusted for factors that might bias MR results, and the unadjusted data are not available [26]. It may not be possible to source summary data on the exact population needed for a particular analysis, for example on either men or women only when looking at sex-specific outcomes) [27]. Finally, a richer array of analyses are possible with individual level data. For example, the estimation of non-linear causal effects across the full range of the exposure and the exploration of effect modification via covariates.

It is of course possible to naively apply summary data MR methods to the one-sample context, estimating both the gene-exposure and gene-outcome associations in the same sample, an analysis made increasingly easy by the advent of large open-access cohort studies such as the UK Biobank (UKB) [28]. This avoids problems with synthesising and harmonizing data from separate cohorts, but can result potentially anti-conservative weak instrument bias due to correlated error. A preliminary investigation has found that this naive approach is particularly bad for pleiotropy robust approaches such as MR-Egger regression [29, 30]. So far, there is no consensus on how best to implement summary data approaches in the one sample setting.

In this paper we propose a general method which we term 'Collider-Correction' that can reliably apply two-sample summary data MR methods to one-sample data, whilst maintaining the simplicity and appeal of the two-sample approach. Our method builds on the work of Dudbridge et. al. [31], who proposed a method to correct for 'index event' (or collider) bias in genetic studies of disease progression, when all subjects included in the analysis have been diagnosed with the disease. In this setting, the analysis is open to contamination from collider bias. Our work serves to clarify that the procedure can be extended to any MR analysis where the aim is to estimate the causal effect, by artificially inducing collider bias in the observational association between X and Y and then correcting for it. This allows any two sample method to be used in a one sample design, thereby benefiting from the plethora of weak instrument and pleiotropy robust approaches available. We show that this approach is (a) statistically efficient compared to artificially splitting the data in two, and (b) will deliver consistent estimates of the causal effect whenever the assumptions of the underlying two-sample approach are satisfied.

Although our method builds on the work of Dudbridge et al, there are several major differences. Firstly, whilst Dudbridge et al focus on the unbiased estimation of the direct SNP-outcome associations, we treat these as nuisance parameters and focus instead on estimation of the causal effect. Secondly, whilst the underlying method we use is closely related to the approach of Dudbridge et al when the chosen method is MR-Egger regression, our paper shows that the underlying method can actually be applied to any MR method. Thirdly, whereas Dudbridge et al propose a solution to adjust for weak instrument bias within the specific context of an MR-Egger model which relies on the InSIDE assumption [9], we propose the use of a SIMEX procedure that can be applied to any regression model, including robust regression models that do not rely on InSIDE for identification. Of course, some recent two-sample approaches have weak-instrument robust weighting built into them, for example MR-RAPS [21, 32]. In this case, SIMEX adjustment is unnecessary.

A major reason for the emergence of weak instrument and pleiotropy robust two-sample MR methods [6, 21, 32] is the avoidance of winner's curse [4], by using one discovery GWAS for instrument selection and two additional data sources for the two-sample MR analysis (i.e. a 'three-sample' design). Although this removes winner's curse by design, it generally yields far weaker instruments. In practice, it may be hard to obtain data from three independent, homogenous cohorts to enact the three-sample approach, but a nice property of Collider-Correction is that it can be enacted with two-independent data sources rather than three. In Results, we apply Collider-Correction to 1 sample individual level UK Biobank data to investigate the causal role of sleep disturbance on HbA1c levels, using both overlapping and non-overlapping GWAS data for instrument selection. In the former case winner's curse is seen to induce a dilution in the MR estimates that is not present in the latter case.

We see three scenarios where our Collider-Correction approach is applicable. Firstly, when interest lies in estimation of the causal effect of an exposure *X* on an outcome, *Y* and only summary data on 'YadjX' genetic associations are available (for example, waist/hip ratio adjusted for BMI from the GIANT consortium). The second is when researchers have direct access to individual level patient data. This is likely to become much more common over time as further international biobank studies follow the lead of UKB in opening up data access. Extracting the summary statistics for our approach then enables the efficient implementation of any two-sample method to the data. This is attractive because two-sample methods are currently more numerous than one sample methods, more familiar to researchers and more technically advanced (especially in their ability to adjust for weak instrument bias and pleiotropy). Furthermore, if one additional GWAS is available for instrument selection, Collider-Correction enables winner's curse, weak instrument bias and pleiotropy to be accounted for using two independent data sets rather than three. The third is when data custodians prefer not to grant direct access to individual level data, but are willing to provide the requisite summary statistics for implementing the Collider-Correction approach, safe in the knowledge that the individual-level data analysis can be performed whilst maintaining data security. Allowing large scale, rapid access to confidential data has obvious benefits to the research community and wider society, as demonstrated through initiatives such as OpenSAFELY [33].

Methods

To motivate ideas, we assume the following individual level data model for the exposure *X* and continuous outcome *Y* for subject *i*:

$$X_i|G_i, U_i = \sum_{j=1}^{k} \beta_{XG_j} G_{ij} + \beta_{UX} U_i + \epsilon_{X_i} \quad (1)$$

$$Y_i|X_i, G_i, U_i = \beta X_i + \sum_{j=1}^{k} \alpha_j G_{ij} + \beta_{UY} U_i + \epsilon_{Y_i} \quad (2)$$

$$= \sum_{j=1}^{k}(\alpha_j + \beta\beta_{XGj})G_{ij} + (\beta\beta_{UX} + \beta_{UY})U_i + \beta\boldsymbol{\varepsilon}_{X_i} + \boldsymbol{\varepsilon}_{Y_i}$$

$$= \sum_{j=1}^{k}\beta_{YGj}G_{ij} + \boldsymbol{\varepsilon}^{*}_{Y_i} \quad (3)$$

Here, $G_i = (G_{i1}, \ldots, G_{ik})'$ represents a set of k variants that predict X_i, β represents the target estimand, reflecting the causal effect of inducing a 1-unit change in the exposure on the outcome, and U represents unmeasured confounding predicting both X and Y. The variables $\boldsymbol{\varepsilon}_{X_i}$, $\boldsymbol{\varepsilon}_{Y_i}$ represent independent residual error terms. Since the unmeasured confounder U is common to both X and Y, the total residual errors around $X|G$, $Y|X$, G and $Y|G$ in Eqs (1)–(3) are correlated. This linear model tacitly assumes that the causal effect is the same for all individuals (that is, regardless of their observed exposure level). This is referred to as 'Homogeneity': it is an example of a fourth IV assumption that is needed to 'point identify' β (assumptions IV1-IV3 are sufficient to test for causality only). We re-write model (2) in 'reduced form' as model (3) to clarify that the underlying SNP-outcome association β_{YGj} is equal to $\alpha_j + \beta\beta_{XGj}$. When the exposure is binary, so that $X = 0$, and $X = 1$ refer to being unexposed and exposed respectively, we can again identify β by assuming Homogeneity. This would mean that the effect of intervening and changing X from 1 to 0 is equal and opposite to the effect of intervening and changing X from 0 to 1.

The standard approach to estimating β with individual level data is Two Stage Least Squares (TSLS). This assumes that all instruments are valid (not pleiotropic), so that $\alpha_j = 0$ for all j. TSLS firstly regresses the exposure on all k genotypes simultaneously to derive an estimate for subject i's genetically predicted exposure: $\hat{X}_i = \sum_{j=1}^{k}\hat{\beta}_{XG_j}G_{ij}$, where $\hat{\beta}_{XG_j}$ is the estimated association between SNP j and X. The outcome Y is then regressed on $\hat{X}_i$ and its regression coefficient is taken as the causal estimate $\hat{\beta}$. As explained in Fig 2, when the set of k SNPs which predict X are mutually independent (i.e. not in linkage disequilibrium), the TSLS estimate is asymptotically equivalent to the IVW estimate [34] obtained by:

- Calculating the causal estimate $\hat{\beta}_j$ by dividing the SNP-outcome association $\hat{\beta}_{YGj}$ (obtained from a regression of Y on G_j) by the SNP-exposure estimate $\hat{\beta}_{XGj}$ for each SNP and;
- Performing an inverse variance weighted meta-analysis of the k individual causal estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_k$.

The inverse variance weights traditionally used make the simplifying assumption that the SNP-exposure association $\hat{\beta}_{XGj}$ is sufficiently precise that its uncertainty can be ignored. This is referred to as the No Measurement Error (NOME) assumption [6]. This procedure is equivalent to fitting the following weighted regression model

$$\hat{\beta}_{YGj} = \beta\hat{\beta}_{XG_j} + \epsilon_{YGj} \quad (4)$$

where ϵ_{YGj} is the mean zero residual error with $Var(\epsilon_{YGj}) = \sigma^2_{YGj} = Var(\hat{\beta}_{YGj})$ and the intercept is constrained to zero. We will refer to this as the 'standard' IVW approach. It is commonly used in two sample summary data MR out of necessity because only summary statistics are available, but not typically in the one sample setting [29].

Inducing collider bias into SNP-outcome associations

Consider a regression of the outcome Y on G and X together (but not U). Under our assumed data generating model:

$$E[Y_i|X_i, G_i] = \beta^* X_i + \sum_{j=1}^{k} \alpha^*_j G_{ij}, \tag{5}$$

yielding estimated coefficients $\hat{\beta}^*$ and $\hat{\alpha}^*_1, \ldots, \hat{\alpha}^*_k$. Since X is a function of both G and U, conditioning on X induces a correlation between them [35]. This is commonly referred to as 'collider bias' [36]. Its presence contaminates the G_j-Y association estimate with a contribution through U so that $\hat{\alpha}^*_j$ is not a consistent estimate for α_j. For the same reason, $\hat{\beta}^*$ is not a consistent estimate for β. It instead reflects the causal effect, plus a contribution from X to Y via U. Such 'collider biased' analyses are usually avoided for this reason [36]. However, it is in a special sense advantageous to fit model (5) because under models (1) and (2), α^*_j, α_j, β^* and β are linked through the following linear relation:

$$\hat{\alpha}^*_j = \alpha_j + (\beta - \beta^*)\beta_{XG_j} + \theta_j, \tag{6}$$

where θ_j is mean zero residual error with $Var(\theta_j) = \sigma^2_{\hat{\alpha}^*_j} = Var(\hat{\alpha}^*_j)$ (see S1(A) Text for a detailed derivation). This suggests the following algorithm for estimating the causal effect:

1. Regress Y on X and G to obtain the collider biased parameter estimates $\hat{\beta}^*$ and $\hat{\alpha}^*_1, \ldots, \hat{\alpha}^*_k$.
2. Regress X on G to obtain estimates $\hat{\beta}_{XG1}, \ldots, \hat{\beta}_{XGk}$, where

$$\hat{\beta}_{XGj} = \beta_{XGj} + \delta_j, \tag{7}$$

   for independent residual error term δ_j with mean zero and variance $\sigma^2_{XGj} = Var(\hat{\beta}_{XGj})$;
3. Fit the linear model:

$$E[\hat{\alpha}^*_j|\hat{\beta}_{XGj}] = \alpha_0 + (\beta - \beta^*)\hat{\beta}_{XG_j} \tag{8}$$

   under a user-specified loss function and pleiotropy-identifying assumption in order to obtain an estimate for the Collider-Correction term $(\beta - \beta^*)$.
4. Adjust the observational estimate to obtain an estimate for the causal effect β via:

$$\hat{\beta} = \hat{\beta}^* + \widehat{\beta - \beta^*} \tag{9}$$

The above procedure, which we call 'Collider-Correction' is a modification and generalisation of the Dudbridge approach [31]. In step 3 and 4 we instead focus on estimation of the Collider-Correction term and the causal parameter β rather than, as Dudbridge et al do, the pleiotropic effects. Crucially, we clarify that, as long as the model for Y given X and G in step 1 is correctly specified, the correlation between the residual error in model (6) and residual error in the first-stage model (7) will have a mean of zero. To illustrate this we simulated 500 independent sets of data from models (1)–(2), each containing individual level data on 10,000 subjects. We fixed the number of SNPs to $k = 50$: each SNP was bi-allelic (taking the values 0,1 or 2), mutually uncorrelated with other SNPs, had a minor allele frequency of 0.3, and collectively explained 1.5% of the variance in the exposure. The correlation between the residual errors in

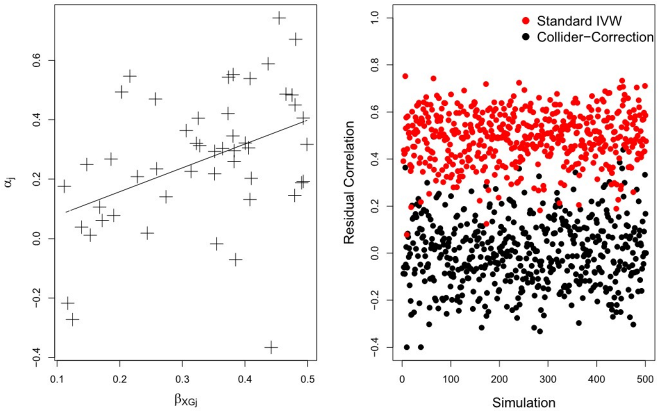

Fig 3. (A = Left): Scatter plot of β_{XGj} and α_j terms for a single simulated data set. (B = Right): Sample correlation between $\hat{\delta}_j$ and $\hat{\theta}_j$ (black) and $\hat{\delta}_j$ and $\hat{\epsilon}_{YGj}$ (red).

https://doi.org/10.1371/journal.pgen.1009703.g003

model (1) and (2) was approximately 0.5 to reflect moderate confounding. SNP-exposure and SNP outcome association parameters β_{XGj} and α_j were generated from dependent distributions, so that their average correlation was approximately 0.45. This is a clear violation of the InSIDE assumption that the sample covariance $\widehat{Cov}(\alpha_j, \beta_{XG})$ is zero [6, 9, 13]. We then applied Step 1 and 2 of the Collider-Correction algorithm to estimate the $\hat{\alpha}_j^*$ and $\hat{\beta}_{XGj}$ terms. Fig 3A shows, for a single simulated data set, the extent of correlation between the 50 β_{XGj} and α_j. Fig 3B shows across all 500 independent data sets, the sample correlation between the first stage residual $\delta_j = \hat{\beta}_{XGj} - \beta_{XGj}$ and both:

- The Collider-Correction residual: $\theta_j = \hat{\alpha}_j^* - \alpha_j - (\beta - \beta^*)\beta_{XGj}$ (shown in black);
- The 'standard' SNP-outcome residual: $\epsilon_{YGj} = \hat{\beta}_{YGj} - \alpha_j - \beta\beta_{XGj}$ (shown in red).

We see that the mean correlation of the Collider-Correction residual with 1st stage residual is zero whereas the mean correlation of the standard SNP-outcome residual with the 1st stage residual is 0.5. This residual error independence property is advantageous because it means that step 3 of the Collider-Correction algorithm can be implemented using any pleiotropy robust two-sample summary data MR method, where the estimand of interest is $\beta - \beta^*$ rather than the causal effect β directly. Crucially, the residual error independence property means that weak instrument bias will induce a dilution in the slope estimate $\widehat{\beta - \beta^*}$ towards zero, because it can be viewed as a consequence of 'classical' measurement error. This makes it easy to quantify and correct for using standard methods, as we will subsequently discuss.

In the toy example above we purposefully generated the data so that the InSIDE assumption was violated across the entire set of SNPs to demonstrate that residual error independence does not rely on InSIDE. However, the success of any subsequently applied Collider-Corrected two sample approach in consistently estimating the causal effect β (i.e. so that it is asymptotically unbiased) will of course depend on the pleiotropy identifying assumption being met, just as if it were being applied in a standard two-sample setting. Although the Collider-Correction algorithm is generalisable in theory to any MR analysis method, we now describe several canonical implementations, which require that the InSIDE assumption is satisfied across either the entire set of SNPs or a subset of SNPs.

Implementing Collider-Correction

Collider-Corrected IVW implementation. To implement the Collider-Corrected IVW approach we set the parameter α_0 to zero in Eq (8) and estimate the slope $(\beta - \beta^*)$ using weighted least squares via the model:

$$\hat{\alpha}^*_j = (\beta - \beta^*)\hat{\beta}_{XG_j} + \theta^*_j \tag{10}$$

where θ^*_j is mean zero residual error with an assumed variance $\sigma^2_{\alpha^*_j} = \mathrm{Var}(\hat{\alpha}^*_j)$. Note that under data-generating model (6) θ^*_j is actually equal to $\theta_j + \alpha_j$. Under the assumption that the mean pleiotropic effect is zero and the InSIDE assumption is satisfied, the residual error independence property of Collider-Correction will mean that θ^*_j is also independent of uncertainty in $\hat{\beta}_{XG_j}$ so that $(\beta - \beta^*)$ can be consistently estimated. The IVW approach then quantifies additional uncertainty in the estimate for $(\beta - \beta^*)$ due to the presence of pleiotropy, by increasing its variance by a factor ϕ proportional to the variance of the estimated residual $Var(\hat{\theta}_j)$ whenever this variance is greater than 1. This is equivalent to fitting a multiplicative random effects model [13].

The IVW estimate uses '1st order weights' that ignore uncertainty in the SNP-exposure association estimate by assuming that its variance $\sigma^2_{XGj} \approx 0$. This is referred to as the NO Measurement Error (NOME) assumption [6]. When this is violated the estimate $\widehat{\beta - \beta^*}$ from model (10) will be diluted towards zero by a factor of $(\bar{F} - 1)/\bar{F}$, where:

$$\bar{F} = \sum_{j=1}^{k} \frac{\hat{\beta}^2_{XGj}}{\sigma^2_{XGj}} \tag{11}$$

See Section 3.2 in [21] for a more detailed explanation. Note that, whilst the Collider-Correction slope is diluted towards zero in the presence of weak instrument bias, the causal estimate itself is still biased toward the observational association estimate $\hat{\beta}^*$, because the causal effect calculated in Step 4 of the Collider-Correction algorithm is the sum of $\hat{\beta}^*$ and $\widehat{\beta - \beta^*}$. A simple and general method for weak instrument bias adjustment that can be applied directly to the IVW estimate from model (10) is Simulation Extrapolation (SIMEX) [19]. Under SIMEX, a parametric bootstrap is used to generate 'pseudo' SNP-exposure associations, each one centred on the observed estimate, but with an increasing amount of uncertainty (i.e. with larger and larger values of σ^2_{XGj}). This subsequently induces an increasing dilution in the IVW estimate for $(\beta - \beta^*)$. A global model is then fitted to the entire set of simulated data in order to extrapolate back to the estimate for $(\beta - \beta^*)$ that would have been obtained if there were no uncertainty in the SNP-exposure associations (i.e. $\sigma^2_{XGj} = 0$, NOME satisfied). SIMEX is attractive because it can be applied to any regression model (and hence many MR methods), and reliable software is available in standard packages, such as R and Stata.

Connecting IVW to LIML and MR-RAPS. An alternative to SIMEX in the special case of the IVW approach is to find the values of $(\beta - \beta^*)$ and $\sigma^2_{\alpha^*} = \mathrm{Var}(\alpha^*_j)$ that minimises the weighted sum of squared residuals in the extended model (12):

$$\hat{\alpha}^*_j = (\beta - \beta^*)\hat{\beta}_{XG_j} + \sqrt{(\beta - \beta^*)\sigma^2_{XGj} + \sigma_{\alpha^*_j} + z\sigma^2_{\alpha^*}}\epsilon_j \tag{12}$$

When $z = 0$ in (12), the pleiotropy variance $\sigma^2_{\alpha^*} = \mathrm{Var}(\alpha^*_j)$ is fixed to zero and the above procedure is equivalent to performing Limited Information Maximum Likelihood (LIML) with summary data (see Section 3.1 in [21]). Furthermore, the weighted sum of squared residuals from (12) follows a χ^2_{L-1} distribution when the assumption that $\mathrm{Var}(\alpha^*_j) = 0$ is satisfied, thus providing a simple weak instrument bias robust test for the presence of pleiotropy. This is referred to as the 'exact' Q statistic [6] which is similar to the simulation-based MR-PRESSO test for 'global' pleiotropy [37, 38].

Unfortunately, when pleiotropy *is* present so that $\mathrm{Var}(\alpha^*_j) \neq 0$, then the LIML estimate will be biased [6]. In order to account for both weak instrument bias and non-zero pleiotropy, z can be set to 1 so that the squared residual minimisation is over both $(\beta - \beta^*)$ and $\sigma^2_{\alpha^*}$. This is equivalent to applying 'MR-RAPS' [21] when applied to the Collider-Correction summary statistics. MR-RAPS actually uses an approximation to the least-squares method because the maximum likelihood estimates are inherently unstable, this entails the use of a score function to proxy for the likelihood and a penalization term to dampen the effect of large residuals.

Collider-Corrected MR-Egger implementation

In order to account for pleiotropy with a non-zero mean but under the InSIDE assumption, we could instead allow the intercept α_0 and slope $(\beta - \beta^*)$ to be freely estimated via weighted least squares by fitting a Collider-Correction MR-Egger model [9]

$$\hat{\alpha}^*_j = \alpha_0 + (\beta - \beta^*)\hat{\beta}_{XG_j} + \theta^*_j, \tag{13}$$

where θ^*_j is mean zero residual error with an assumed variance $\sigma^2_{\alpha^*_j} = \mathrm{Var}(\hat{\alpha}^*_j)$. Note that under data-generating model (7) θ^*_j is actually equal to $\theta_j + \alpha_j - \alpha_0$. Using the same argument as for the IVW model, when InSIDE is satisfied this will consistently estimate the Collider-Correction slope (adjusted for α_0) and from there, the causal effect. Additional uncertainty due to pleiotropy can again be handled using a multiplicative random effects model [13]. To assess the vulnerability of the MR-Egger regression estimates to weak instrument bias due to violation of the NOME assumption, we use the I^2_{GX} statistic [5]:

$$I^2_{GX} = \frac{Q_{GX} - (k-1)}{Q_{GX}}, \text{where } \ Q_{GX} = \sum_{j=1}^{k} \frac{(\hat{\beta}_{XGj} - \bar{\beta}_{XGj})^2}{\sigma^2_{XGj}} \tag{14}$$

The expected dilution in the Collider-Correction $\widehat{\beta - \beta^*}$ due to weak instruments is equal to $(\beta - \beta^*)I^2_{GX}$. This can easily be adjusted for by applying SIMEX to model (13), just as for the IVW approach.

Collider-Corrected robust regression

IVW MR-Egger and MR-RAPS rely on the InSIDE assumption to consistently estimate the causal effect. This may be violated in practice, hence the rationale for the development of alternative, robust methods such as the Weighted Median [10]. In the two-sample summary data context it can consistently estimate the causal effect if the majority of the 'weight' in the MR analysis stems from genetic variants that are not pleiotropic. That is, the existence of a SNP subset S is assumed for which $\widehat{Cov}_{j \in S}(\alpha_j, \beta_{XGj}) = \widehat{Cov}_S(0, \beta_{XGj}) = 0$, but InSIDE is allowed to be violated for SNPs not in subset S. The downside of weighted median approach is that it is not directly equivalent to a regression model, which in turn means that we can not benefit from a procedure like SIMEX to perform a weak instrument bias adjustment. However, there is a close connection between the median and minimisation using a Least Absolute Deviation (LAD), or L1-norm. We therefore propose the use of LAD regression [39] instead of least squares, at Step 3 of the Collider-Correction algorithm, with α_0 set to zero. This is close in spirit to the Weighted Median, and is amenable to SIMEX-adjustment too. The exact 'break-down point' of LAD regression (or the proportion of pleiotropic SNPs above which LAD regression will not deliver a consistent estimate) depends on the data generating model, but is bounded between 1/k (k being the number of SNPs) and 1/2.

Simulation studies

In order to confirm our theoretical results and assess the performance of the Collider-Correction algorithm, data sets of between 5000 and 50,000 individuals were generated under models (1) and (2) as described previously. Across all simulations:

- The causal effect of inducing a one-unit change in the exposure on the outcome, β, was set to 0.5 for all individuals;
- The correlation between the residual errors in model (1) and (2) was set to approximately 0.9 to reflect strong confounding;
- The observational estimate for $\hat{\beta}^*$ and the true Collider-Correction term $\beta - \beta^*$ were approximately 1.12 and -0.62 respectively.

To showcase the ability of IVW-based approaches, MR-Egger regression and LAD regression, pleiotropy parameters and SNP-exposure associations were generated under three distinct models:

- For IVW simulations, pleiotropic effect parameters $\alpha_1, \ldots, \alpha_{50}$ were generated with a zero mean independently of the SNP-exposure associations $\beta_{XG1}, \ldots, \beta_{XG50}$ (InSIDE satisfied);
- For MR-Egger simulations, pleiotropic effect parameters $\alpha_1, \ldots, \alpha_{50}$ were generated with a non-zero mean independently of the SNP-exposure associations $\beta_{XG1}, \ldots, \beta_{XG50}$ (InSIDE satisfied);
- For LAD regression simulations, pleiotropic effect parameters $\alpha_1, \ldots, \alpha_{15}$ were generated with a non-zero mean dependent on the SNP-exposure associations $\beta_{XG1}, \ldots, \beta_{XG15}$ (with an average correlation of 0.5) whilst $\alpha_{16}, \ldots, \alpha_{50}$ were set to 0. InSIDE was therefore strongly violated across SNPs 1:15, satisfied across SNPs 16:50 and violated across all SNPs, respectively.

IVW simulation results. Fig 4 shows, for a range of sample sizes the average value across 1000 independent data sets of: (a) The standard IVW estimate (black line); (b) the SIMEX adjusted standard IVW estimate (blue line); (c) the Collider-Corrected IVW estimate (red line); (d) the Collider-Corrected IVW estimate with SIMEX correction (green line); (e) the TSLS estimate (orange line) and (f) the Collider-Corrected MR-RAPS estimate (implemented using the 'Tukey' penalization option). We see that methods (a), (c) and (e) give approximately the same answer, and are therefore hard to individually distinguish in the figure. The approximate equivalence of the TSLS and IVW approaches with uncorrelated SNPs is well known, but it is also reassuring that our two step approach is also equivalent. We also see that applying a direct SIMEX correction to method (a) (i.e. method (b)) dramatically increases the bias of the causal estimate beyond even that of the observational estimate for small sample sizes. This bias is slow to diminish as the sample size grows. This poor performance is because uncertainty in the SNP-exposure association estimates can **not** be viewed as classical measurement error within a standard IVW model. Conversely, we see that applying a SIMEX correction to the

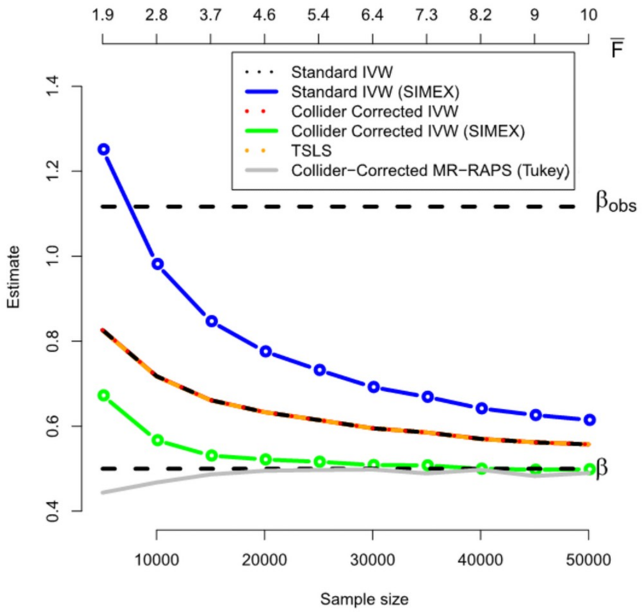

Fig 4. Performance of IVW implementations (including the Collider-Correction algorithm) using one-sample data.

https://doi.org/10.1371/journal.pgen.1009703.g004

Collider-Corrected IVW estimate (c) (i.e method (d)) yields a steadily decreasing bias which is essentially zero when the mean *F* statistic across the instruments is larger than 5. The Collider-Corrected MR-RAPS estimate performs very well too, and is essentially unbiased for mean *F* statistics greater than 3.5.

Fig 5 gives further intuition on why the correction process works. The black line shows the estimated Collider-Correction $\widehat{\beta - \beta^*}$ as a function of the given sample size. The blue line

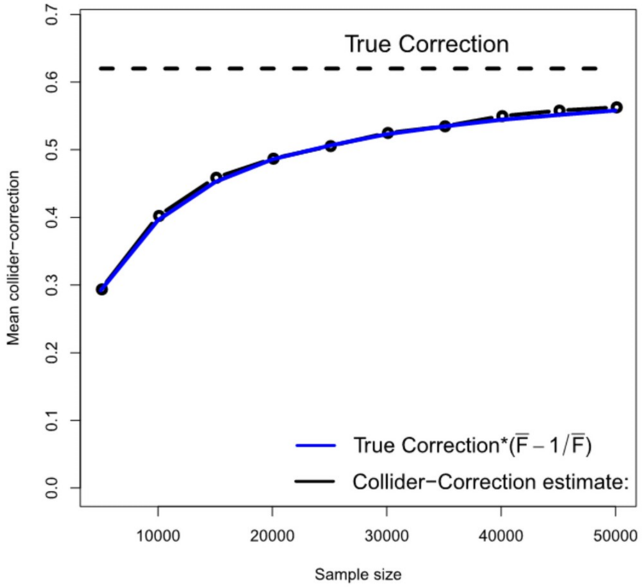

Fig 5. An illustration that the Collider-Correction slope's dilution can be accurately predicted using the F-statistic.

https://doi.org/10.1371/journal.pgen.1009703.g005

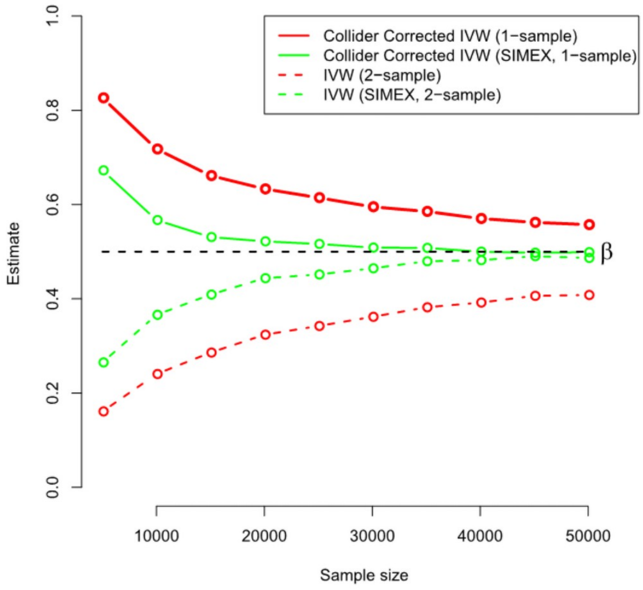

Fig 6. A comparison of the one sample Collider-Correction versus two-sample IVW approaches in terms of bias.

https://doi.org/10.1371/journal.pgen.1009703.g006

shows the true Collider-Correction multiplied by the expected dilution factor $\frac{\bar{F}-1}{\bar{F}}$, which varies as a function of the sample size. The fact that the two lines are in good agreement indicates that the dilution in $\widehat{\beta - \beta}^{*}$ can be perfectly predicted by the F-statistic formula and underlines why SIMEX can be used to correct for it. Fig 6 shows the performance of the IVW estimate implemented using the (one sample) Collider-Correction algorithm, versus that obtained from artificially splitting the data in two, and applying the 'standard' IVW approach. That is, calculating SNP-exposure associations in one half, SNP-outcome associations in the other half and combining in the usual manner. This ensures that the residual error independence property is satisfied, as it is for the one sample Collider-Correction approach. Results for each method are shown with and without SIMEX correction. We see that the absolute bias of the Collider-Correction implementations is less than that of the two-sample implementation. However, the two estimation strategies differ more substantially in terms of precision, as shown in Fig 7. Collider-correction of one sample data is shown to be far more efficient than sampling splitting.

Figs 8A and 9A show the corresponding standard deviation and mean-squared error (a measure of accuracy that equals an estimate's variance plus the squared bias) for all IVW-based methods across the same set of simulations. They show that whilst the MR-RAPS estimate is less biased for small sample sizes than the Collider-Corrected IVW method with SIMEX adjustment, it is more variable and less accurate.

MR-Egger simulation results. Fig 10 shows for a range of sample sizes the average value across 1000 independent data sets of: (a) The standard (one-sample) MR-Egger estimate (black line); (b) the SIMEX adjusted standard MR-Egger estimate (blue line); (c) the Collider-Corrected MR-Egger estimate (red line) and (d) the Collider-Corrected MR-Egger estimate with SIMEX correction (green line). As the sample size increases the I^2_{GX} statistic increases from 0.1 to 0.5. This signals that the 50 SNPs get collectively stronger as a set of instruments within MR-Egger as the sample size increases, but even at the largest sample size we expect a dilution of 50% in the MR-Egger slope. Again, we see that standard and Collider-Corrected MR-Egger methods give the same results, but the two approaches differ greatly under SIMEX

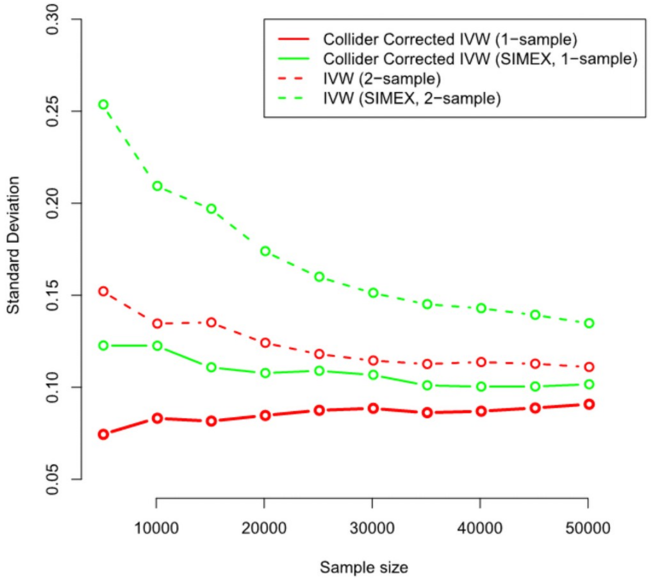

Fig 7. A comparison of the one sample Collider-Correction versus two-sample IVW approaches in terms of efficiency.

https://doi.org/10.1371/journal.pgen.1009703.g007

correction, with the SIMEX adjusted Collider-Corrected estimate being least biased. In Fig 11 we show how dilution in the Collider-Corrected slope estimate $\widehat{\beta - \beta^*}$ for MR-Egger can be accurately quantified using the I^2_{GX} statistic, just as the F-statistic predicts the dilution for IVW. This explains why SIMEX adjustment works.

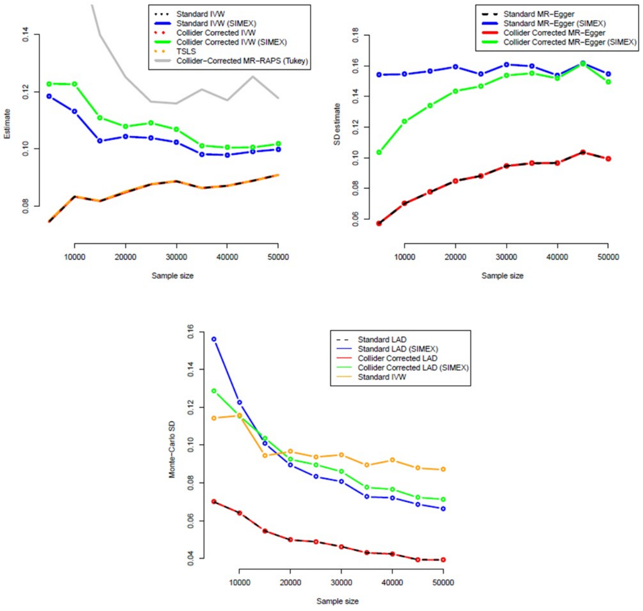

Fig 8. Monte-Carlo standard deviations for all IVW (A = top-left), MR-Egger (B = top-right) and LAD regression (C = bottom) estimators.

https://doi.org/10.1371/journal.pgen.1009703.g008

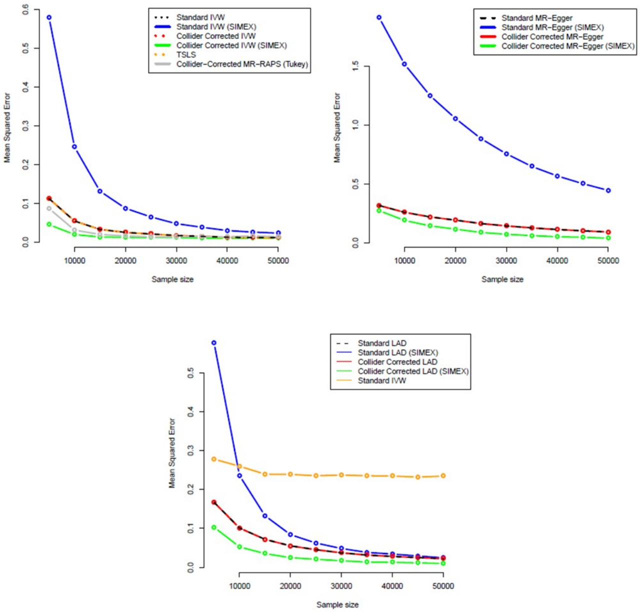

Fig 9. Mean Squared Error for all IVW (A = top-left), MR-Egger (B = top-right) and LAD regression (C = bottom) estimators.

https://doi.org/10.1371/journal.pgen.1009703.g009

Figs 8B and 9B show the corresponding standard deviation and mean squared error for all MR-Egger-based methods across the same set of simulations. Standard and Collider-Corrected MR-Egger are seen to have the joint smallest variance, but Collider-Corrected MR-Egger with SIMEX adjustment has the smallest mean-squared error because it is far less biased.

LAD-regression simulation results. Fig 12 shows for a range of sample sizes the average value across 1000 independent data sets of: (a) The standard (one-sample) LAD-regression

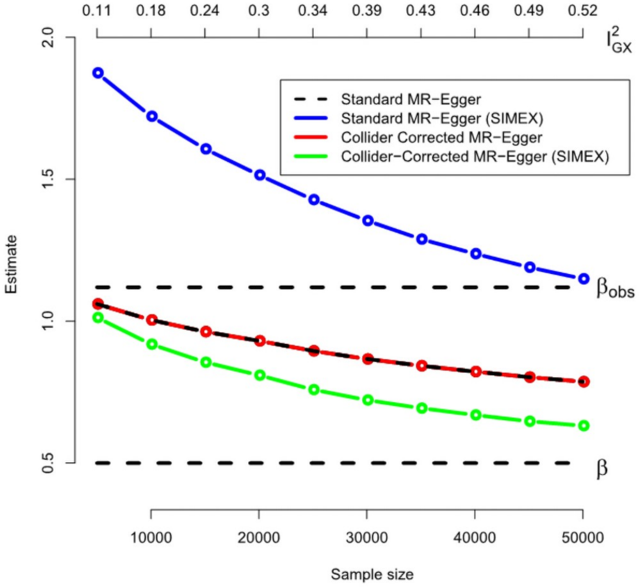

Fig 10. Performance of the MR-Egger implementation of the Collider-Correction algorithm under a directional pleiotropy scenario.

https://doi.org/10.1371/journal.pgen.1009703.g010

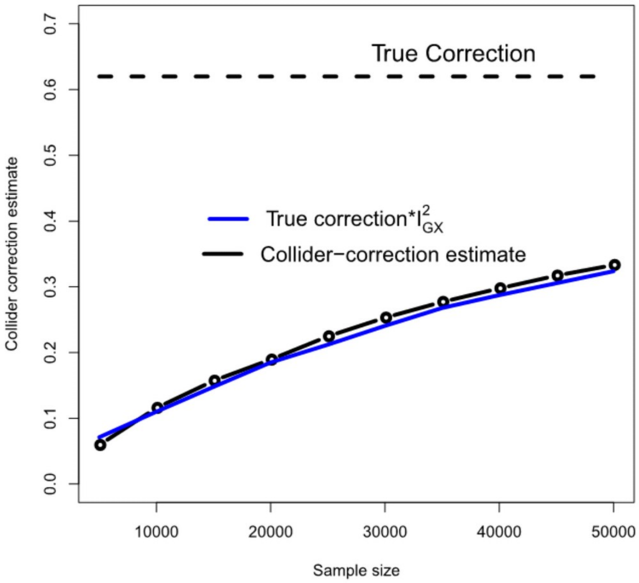

Fig 11. An illustration that the Collider-Correction slope's dilution can be accurately predicted using the I^2_{GX} statistic.

https://doi.org/10.1371/journal.pgen.1009703.g011

estimate (black line); (b) the SIMEX adjusted standard LAD regression estimate (blue line); (c) the Collider-Corrected LAD regression estimate (red line) and (d) the Collider-Corrected LAD regression estimate with SIMEX correction (green line). For comparison we also show (e) the standard IVW estimate: its bias does not approach zero as the sample size increases because of the presence of non-zero mean pleiotropy violating InSIDE, which is the very motivation for LAD regression. As in the previous simulations, standard and Collider-Corrected

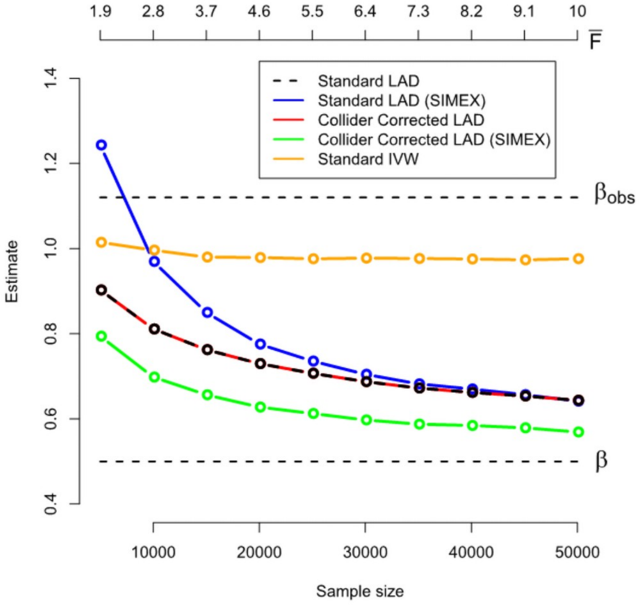

Fig 12. Performance of the LAD-regression implementation algorithm under InSIDE violating pleiotropy.

https://doi.org/10.1371/journal.pgen.1009703.g012

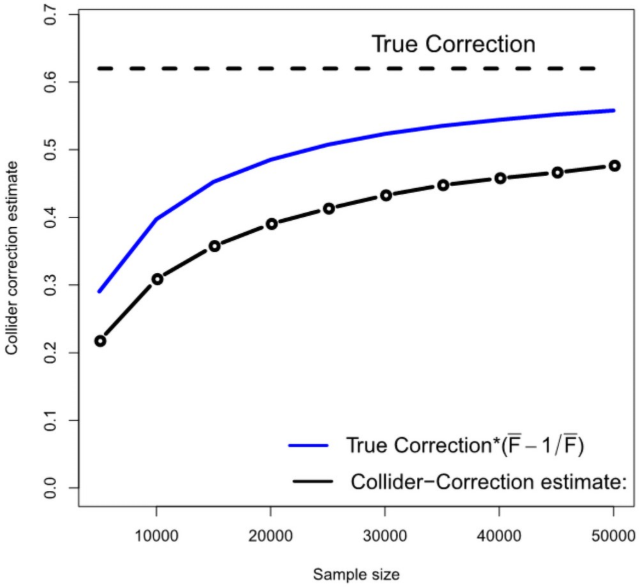

Fig 13. An illustration that the Collider-Correction slope's dilution under a LAD-regression analysis can be approximately (but not exactly) predicted using the F-statistic.

https://doi.org/10.1371/journal.pgen.1009703.g013

LAD regression give identical point estimates on average, but when SIMEX adjustment is applied the two estimates diverge substantially. Collider-Corrected LAD regression with SIMEX adjustment results in the least biased estimates of all.

Fig 13 plots the mean dilution in the Collider-Corrected LAD regression estimate, versus that predicted by the IVW dilution factor $\frac{\bar{F}-1}{\bar{F}}$. The fact that the observed dilution is below the expected IVW dilution illustrates that LAD regression is more vulnerable to weak instrument bias, because it is a less efficient but more robust technique. This emphasises the importance of being able to address its weak instrument bias.

Figs 8C and 9C show the corresponding standard deviation and mean-squared error for all LAD regression-based methods across the same set of simulations. The same pattern of higher variance but lower mean-squared error is seen for the SIMEX adjusted Collider-Corrected LAD regression approach as in the MR-Egger case.

The MR-RAPS approach can, in theory, consistently estimate the causal effect when a small proportion of SNPs are pleiotropic and violate the InSIDE assumption, as long as their contribution is strongly penalized by its robust loss function. In order to test this we also calculated the MR-RAPS estimate when applied to the simulated data for LAD regression. MR-RAPS was seen to work well for a proportion of simulated data sets, but its estimates were unstable: in many cases they were an order of magnitude larger than the true value of 0.5. To illustrate this, Fig 14 shows the distribution of its estimates at the largest sample size of 50,000 subjects, where it was most stable. Even in this case substantial instability is observed.

R code for reproducing the simulation study results is available in S1 Code.

Results: Assessing the causal role of insomnia on HbA1c

Observationally, sub-optimal sleep (i.e., low sleep quantity and quality) has been found to be associated with hyperglycaemia [40–42] and increased diabetes risk [43]. Insomnia, defined as difficulty initiating or maintaining sleep, is one of the most important indices of sleep quality [44]. It has been associated with type 2 diabetes in observational studies [44] and in a previous

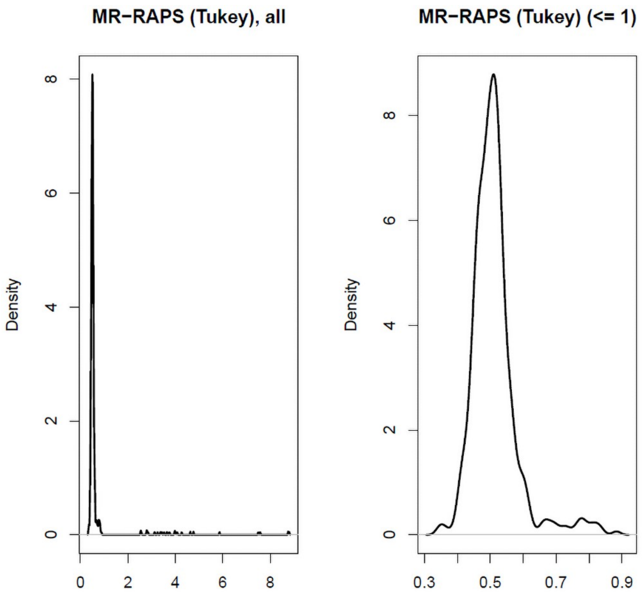

Fig 14. Distribution of MR-RAPS estimates at sample size = 50,000 when the data were generated under the LAD regression model. (A = Left: all estimates, B = Right: estimates less than 1.

https://doi.org/10.1371/journal.pgen.1009703.g014

Mendelian randomization study [45]. However, it is unclear whether associations with insomnia are mediated through HbA1c in the general population, whose glucose levels may not meet the threshold criteria for a formal diabetes diagnosis. As such, we focus on a potentially causal role of insomnia on HbA1c, a well-established clinical assessment of long-term glycaemic regulation that is central to the diagnosis of diabetes [46]. To address this question we use individual level data on approximately 320,000 individuals in UK Biobank to furnish a one sample Mendelian randomization study.

Two hundred and forty-eight independent genetic variants at 202 loci were associated with self-reported insomnia at or below the standard genome-wide significance threshold (p-value$<5\times10^{-8}$) in a recent GWAS of over 1.33 million UK Biobank and 23andMe individuals reported by Jansen [45] which collectively explained 2.6% of the total trait variance. SNP-exposure associations were measured on the log-odds scale using logistic regression. Among this set of variants, 240 SNPs were in principle available for use as instruments in UK Biobank. In this cohort, participants were asked: “Do you have trouble falling asleep at night or do you wake up in the middle of the night?” with responses “Never/rarely”, “Sometimes”, “Usually”, or “Prefer not to answer”. Those who responded “Prefer not to answer” were set to missing. To reflect the Jansen analysis, the remaining entries were treated as a binary variable for insomnia symptoms, with “Never/rarely”, “Sometimes”, and “Usually” coded as 0, 0, and 1, respectively and a logistic regression performed. HbA1c measurements were obtained from a panel of biomarkers assayed from blood samples collected at baseline from UK Biobank participants. HbA1c (mmol/mol) was measured in red blood cells by HPLC analysis using Bio-Rad VARIANT II Turbo and log-transformed.

Instrument selection and winner’s curse

The mean *F* statistic for the 240 genetic instruments in the original GWAS was 41, but in order to avoid winner’s curse we did not want to incorporate these estimates directly into our MR analysis. In UK Biobank the same SNPs had an $\bar{F}$ of approximately 8.3 and an I^2_{GX} statistic of approximately 40%, meaning that the MR analysis was susceptible to bias due to both weak instrument and pleiotropy. This motivates the use of our Collider-Correction method for causal estimation. However, the original Jansen GWAS combined data from the UK Biobank (n = 386,533) and 23andMe (n = 944,477) using METAL [47]. As such, there was an approximate 23% overlap between data used for SNP discovery and for estimation in our MR model [4]. To additionally assess the impact of winner's curse for this reason we performed our subsequent analysis using (a) all 240 SNPs and (b) a subset of 112 SNPs that were only genome-wide significant using only the 23andMe portion of the Jansen data. Analysis (b) is completely protected from winner's curse whereas (a) is not. The downside of analysis (b) is that, with an $\bar{F}$ of 6.8, it is even more susceptible to weak-instrument bias.

Methods used

We applied the TSLS, IVW, MR-Egger, LAD regression and MR-RAPS approaches to the data. The IVW, MR-Egger and LAD regression approaches were implemented in three ways (1) The 'Standard' 1-sample approach (i.e. using all the data to estimate SNP-exposure and SNP-outcome associations); (2) the Collider-Correction algorithm and (3) Collider-Correction with SIMEX adjustment. Note that MR-RAPS incorporates an internal weak instrument bias adjustment and there is no need to additionally apply a SIMEX algorithm to it. Along with MR-RAPS, we refer to approach (3) as the 'gold-standard' methods.

Causal estimates

SNP exposure associations $\hat{\beta}_{XGj}$ were obtained from a logistic regression of insomnia on the set of SNPs as well age at recruitment, sex, assessment centre, 10 genetic principal components, and genotyping chip. Estimates for collider biased SNP outcome associations $\hat{\alpha}_j^*$ were obtained from a multivariable regression of HbA1c on observed insomnia severity, all genetic variants and the same additional covariates. This second regression additionally yielded an estimate for the collider biased observational association between insomnia severity and HbA1c of $\hat{\beta}^* = 0.012$ (S.E. = 0.00057).

Fig 15 plots the collider biased SNP-outcome associations versus the SNP-exposure associations for analysis (a). Overlaid on the plot are the weak-instrument and pleiotropy adjusted Collider-Correction slopes $\widehat{\beta - \beta^*}$ estimated by the four gold standard methods. The Q statistic is 809 (df = 239) providing overwhelming evidence of heterogeneity due to pleiotropy. The 13 SNPs circled in black contribute a component to this global statistic with a bonferroni corrected p-value below (5/240)% and could therefore be classed as outliers. Adjusted causal effect estimates can be found in Table 1. Across all methods, we see a consistent picture: a unit increase in the log-odds of insomnia leads to an increase of between 0.17 and 0.24 units of log mmol/mol HbA1c. All estimates are further from the null than the collider biased observational association, $\hat{\beta}^*$. However the results highlight that, without weak-instrument adjustment, all summary data MR-methods are biased in the direction of $\hat{\beta}^*$.

Table 1 (rows 6:10) and Fig 16 show the MR results for analysis (b) using only the 112 SNPs identified in Jansen from 23andMe data, which are immune to the dilution bias caused by winner's curse. These SNPs have a weaker mean F statistic of 6.88 but a higher I^2_{GX} statistic of 52%. All causal estimates are seen to increase when compared to analysis (a). This is because the winner's curse which is present in (a) leads to an over-estimation of the SNP-exposure association (which forms the denominator of the standard ratio estimate for β) and thus an underestimation of the causal effect. Again, across all methods, we see consistent evidence that the insomnia causally increases HbA1c.

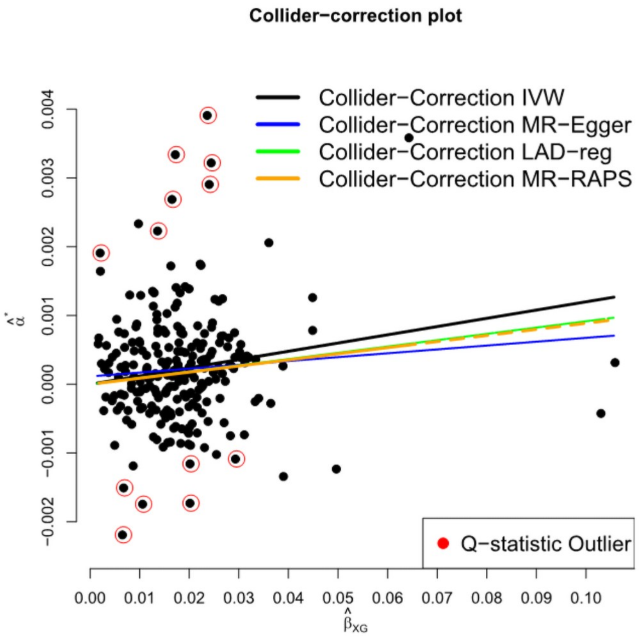

Fig 15. Collider biased SNP outcome associations, $\hat{\alpha}_j^*$, versus SNP-exposure associations, $\hat{\beta}_{XGj}$ for 240 SNPs that were genome-wide significant using 23andMe and UKB data.

https://doi.org/10.1371/journal.pgen.1009703.g015

In total there were 14 outlier SNPs (13 SNPs in analysis (a) and (6) in analysis (b), respectively), which were investigated using the GWAS Catalog (https://www.ebi.ac.uk/gwas/), a full list of which can be found in S1(B) Text and S1 Data. Most of these SNPs are only associated with insomnia except rs10758593 (type 1 and type 2 diabetes), rs12917449 (type 2 diabetes), rs1861412 (BMI) and rs429358 (70+ traits). This provides some biological evidence for the existence of pleiotropy, which further underlines the utility of using robust methods that account for its presence.

Discussion

In this paper we clarify how the principle of Collider-Correction offers a vehicle for applying any two-sample summary data MR method to one sample data, making it easy to account for both pleiotropy and weak instrument bias. Our method is closely related to the approach of Dudbridge et al [31] for genetic studies of disease progression, and primarily serves to emphasise that this procedure is in fact applicable to any MR analysis. We used our new method to provide important insights into the role of insomnia on glycated haemoglobin and, by extension, on incident diabetes.

A nice feature of our approach is that the Collider-Correction term $\beta - \beta^*$ will be large (and therefore the Collider-Corrected estimate will be clearly distinct from the observational association) precisely when there is strong confounding. Conversely, when there is weak confounding, or the confounding has been sufficiently adjusted for, $\beta - \beta^*$ will be zero and Collider-Correction estimate will equal the observational association. In this case, the observational association then becomes a consistent and likely very efficient estimate of the true causal effect. Collider-Correction therefore naturally promotes the triangulation and synthesis of observational and MR estimates, which can estimate the true causal effect with distinct but complementary assumptions.

Table 1. Point estimates, standard errors and p-values for the: TSLS, IVW, MR-Egger, LAD-regression and MR-RAPS methods. Estimates reflect the average causal effect of a unit increase in the log-odds of insomnia on HbA1c levels across the population. 'Standard' = standard 1-sample analysis. Top rows: Analysis (a)—All 240 SNPs from Jansen et al used. Bottom rows: Analysis (b)—only genome wide significant SNPs from 23andMe data (ignoring UK Biobank) used.

Method	Estimate	S.E	p-value
Analysis (a): 23andMe + UK Biobank significant SNPs			
# SNPs: 240, $\bar{F} = 8.36$, Q(p-$value$) = 809 ($<2 \times 10^{-16}$), $I^2_{GX} = 41.0\%$			
β^*	0.012	0.00057	$< 2 \times 10^{-16}$
TSLS	0.016	0.002	$< 1 \times 10^{-16}$
Standard IVW	0.013	0.008	5.04×10^{-7}
Standard MR-Egger	0.007	0.005	1.3×10^{-1}
Standard LAD	0.011	0.004	2.09×10^{-3}
Collider-Corrected IVW	0.022	0.0028	1.1×10^{-15}
Collider-Corrected IVW (SIMEX)	0.024	0.0031	3.1×10^{-14}
Collider-Corrected MR-Egger	0.015	0.0060	1.3×10^{-2}
Collider-Corrected MR-Egger (SIMEX)	0.017	0.0085	4.5×10^{-2}
Collider-Corrected LAD	0.020	0.0036	2.0×10^{-8}
Collider-Corrected LAD (SIMEX)	0.021	0.0024	$< 2 \times 10^{-16}$
Collider-Corrected MR-RAPs	0.020	0.0026	3.1×10^{-15}
Analysis (b) 23andMe significant SNPs only			
# SNPs: 112, $\bar{F} = 6.88$, Q(p-$value$) = 385 ($<2 \times 10^{-16}$), $I^2_{GX} = 52.1\%$			
β^*	0.012	0.00057	$< 2 \times 10^{-16}$
TSLS	0.017	0.003	2.39×10^{-10}
Standard IVW	0.014	0.004	5.23×10^{-4}
Standard MR-Egger	0.008	0.006	1.76×10^{-1}
Standard LAD	0.012	0.006	3.30×10^{-2}
Collider-Corrected IVW	0.024	0.0045	1.2×10^{-7}
Collider-Corrected IVW (SIMEX)	0.026	0.0051	3.3×10^{-7}
Collider-Corrected MR-Egger	0.020	0.0083	1.8×10^{-2}
Collider-Corrected MR-Egger (SIMEX)	0.023	0.0110	4.5×10^{-2}
Collider-Corrected LAD	0.021	0.0056	1.5×10^{-4}
Collider-Corrected LAD (SIMEX)	0.024	0.0042	1.4×10^{-8}
Collider-Corrected MR-RAPs	0.023	0.0043	3.6×10^{-8}

https://doi.org/10.1371/journal.pgen.1009703.t001

We showcased the Collider-Correction approach using four univariate MR approaches that estimate a single causal effect parameter. At the cutting-edge of MR methods research, new approaches are attempting to: estimate causal effects identified by different clusters of SNPs [32, 48, 49]; simultaneously estimate causal effects via multiple exposures [50, 51], or quantify non-linear effects of an exposure [52]. The Collider-Correction algorithm can in principle be adapted to fit all of these multi-parameter approaches and this is an important topic of future research.

The insomnia data was affected by a small amount of winner's curse, which we removed by design in a sensitivity analysis by restricting our SNP set to those obtained from a purely independent data source. More sophisticated approaches to adjusting for winner's curse are possible by incorporating the original Discovery data. For example, Bowden and Dudbridge [4] describe the most statistically efficient way to combine SNP discovery and validation data from two non-overlapping GWAS studies and remove winner's curse. As further work, we plan to extend this approach and combine it with Collider-Correction.

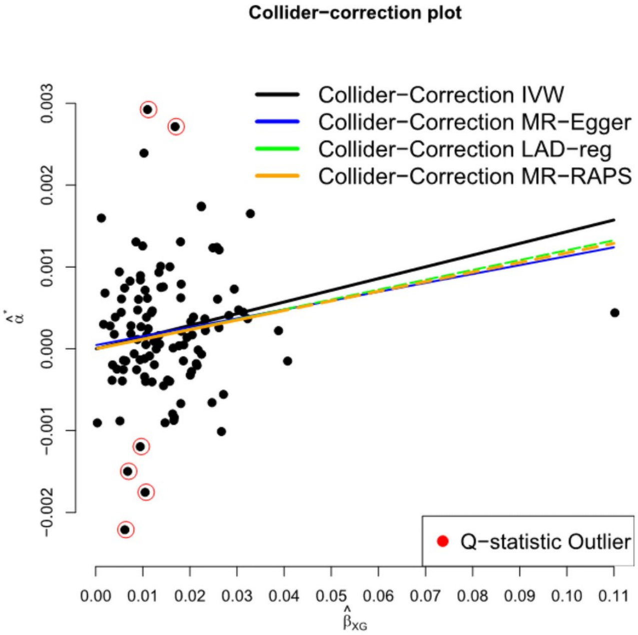

Fig 16. Collider biased SNP outcome associations, $\hat{\alpha}_j^*$, versus SNP-exposure associations, $\hat{\beta}_{XGj}$ for 112 SNPs that were genome-wide significant using 23andMe data only.

https://doi.org/10.1371/journal.pgen.1009703.g016

Often in MR analyses the outcome of interest is binary and a logistic regression model is used in place of the linear model to estimate the causal effect on the odds ratio scale. In this case, the interpretation of causal estimates from a resulting Collider-Correction analysis will be more nuanced for the following reason. Even if we replaced the assumed linear outcome model in Eq (2) with a logistic model, so that β reflected the true causal log-odds ratio for a unit increase in the exposure experienced by each individual, the causal effect estimate (which is a population average) will be diluted by a factor that is proportional to the variance of the residual error in the model not explained by the genetically predicted exposure. This is due to the fact that the odds ratio is a non-collapsible measure [53]. Although this dilution is a very general phenomenon that affects all logistic regression based analysis, three obvious options exist to the applied researcher if implementing Collider-Correction in the binary outcome case. The first would be to simply accept the interpretation of the causal estimate as a population average effect. The second would be to attempt to better approximate the individual causal effect by additionally adjusting for the first stage residual, (that is the observed exposure minus its genetically predicted value) in the second stage logistic model. This is referred to as the Control Function or adjusted IV approach [54]. The third option would be to estimate the causal effect on a risk difference scale. Since the risk difference is a collapsible measure, individual and population average effects are the same. Risk difference estimates can be estimated either by fitting a linear probability model or by extracting the risk difference contrast from the logistic model. This latter approach can be implemented using the `margins()` package in `R`. A thorough investigation of the performance of Collider-Correction in the binary outcome setting is an interesting avenue for future research.

Supporting information

S1 Text. A: A formal proof of the Collider-Correction formulae. B: A list of outlying SNPs detected in analysis (a) and analysis (b) of the data example.
(PDF)

S1 Data. Additional functional information on the outlying SNPs detected in analysis (a) and analysis (b) of the data example.
(ZIP)

S1 Code. R scripts for re-creating the simulation study results in the paper.
(R)

Author Contributions

Conceptualization: Ciarrah Barry, Frank Dudbridge, Jack Bowden.

Data curation: Junxi Liu.

Formal analysis: Ciarrah Barry, Junxi Liu, Deborah A. Lawlor, Jack Bowden.

Funding acquisition: Martin K. Rutter, Deborah A. Lawlor, Jack Bowden.

Investigation: Junxi Liu, Rebecca Richmond, Martin K. Rutter, Deborah A. Lawlor, Jack Bowden.

Methodology: Ciarrah Barry, Frank Dudbridge, Jack Bowden.

Project administration: Martin K. Rutter.

Software: Junxi Liu, Jack Bowden.

Supervision: Rebecca Richmond, Martin K. Rutter, Deborah A. Lawlor, Jack Bowden.

Writing – original draft: Junxi Liu, Jack Bowden.

Writing – review & editing: Junxi Liu, Rebecca Richmond, Martin K. Rutter, Deborah A. Lawlor, Frank Dudbridge, Jack Bowden.

References

1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 2003; 32:1–22. https://doi.org/10.1093/ije/dyg070
2. Sheehan N, Didelez V, Burton P, Tobin M. Mendelian Randomisation and Causal Inference in Observational Epidemiology *PLOS Medicine* 2008; 5:1–6. https://doi.org/10.1371/journal.pmed.0050177 PMID: 18752343
3. Davey Smith G, Lawlor D, Harbord R, Timpson N, Day I, Ebrahim S. Clustered Environments and Randomized Genes: A Fundamental Distinction between Conventional and Genetic Epidemiology *PLOS Medicine* 2007; 4:1–8.
4. Bowden J, Dudbridge F. Unbiased estimation of odds ratios: combining genomewide association scans with replication studies *Genetic Epidemiology* 2009; 33:406–418 https://doi.org/10.1002/gepi.20394 PMID: 19140132
5. Bowden J, Del Greco F, Minelli C, Davey Smith G, Sheehan N, Thompson J Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic *IJE* 2016; 45:1961–1974 https://doi.org/10.1093/ije/dyw220 PMID: 27616674
6. Bowden J, Del Greco F, Minelli C, Zhao Q, Lawlor D, Sheehan N, Thompson J, Davey Smith G. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption *IJE* 2018; 48:728–742
7. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies *HMG* 2018; 27:R195–R208 https://doi.org/10.1093/hmg/ddy163 PMID: 29771313

8. Kang H, Zhang A, Cai T, Small D. Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization *JASA* 2016; 111:132–144 https://doi.org/10.1080/01621459.2014.994705

9. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression *IJE* 2015; 44:512–525 https://doi.org/10.1093/ije/dyv080 PMID: 26050253

10. Bowden J, Davey Smith G, Haycock P, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator *Genetic Epidemiology* 2016; 40:304–314 https://doi.org/10.1002/gepi.21965 PMID: 27061298

11. Bowden J, Hartwig F, Davey Smith G. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption *IJE* 2017; 46:1985–1998 https://doi.org/10.1093/ije/dyx102 PMID: 29040600

12. Burgess S, Butterworth A, Thompson S. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data *Genetic Epidemiology* 2013; 37:685–665 https://doi.org/10.1002/gepi.21758 PMID: 24114802

13. Bowden J, Del Greco F, Minelli C, Davey Smith G, Sheehan N, Thompson J A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Statistics in Medicine 2017; 36:1783–1802. https://doi.org/10.1002/sim.7221 PMID: 28114746

14. Hemani G, Zheng J, Elsworth B, Wade K, Haberland V, Baird D et al The MR-Base platform supports systematic causal inference across the human phenome. *e-Life* 2018; 7:e34408 https://doi.org/10.7554/eLife.34408 PMID: 29846171

15. Bowden J, Spiller W, Del Greco F, Sheehan N, Thompson J, Minelli C, Davey Smith G. Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression *IJE* 2018; 47:1264–1278

16. Lawlor D, Wade K, Borges M, Palmer T, Hartwig F, Hemani G. A Mendelian Randomization Dictionary Useful Definitions and Descriptions for Undertaking, Understanding and Interpreting Mendelian Randomization Studies *OSF Preprints* 2019

17. Inoue A, Solon G. Two-sample Instrumental Variable Estimators *The Review of Economics and Statistics* 2010; 92:557–561 https://doi.org/10.1162/REST_a_00011

18. Hyslop R, Imbens G. Bias from Classical and Other Forms of Measurement Error *Journal of Business & Economic Statistics* 2001; 19:475–481 https://doi.org/10.1198/07350010152596727

19. Cook J, Stefanski L Simulation-Extrapolation Estimation in Parametric Measurement Error Models *JASA* 1994; 89: 1314–1328 https://doi.org/10.1080/01621459.1994.10476871

20. Hardin J, Schmiediche H, Carroll R. The Simulation Extrapolation Method for Fitting Generalized Linear Models with Additive Measurement Error *The Stata Journal* 2003; 3:373–385 https://doi.org/10.1177/1536867X0300300406

21. Zhao Q, Wang J, Hemani G, Bowden J, Small D. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score *The Annals of Statistics* 2020; 48: 1742–1769 https://doi.org/10.1214/19-AOS1866

22. Zhao Q, Wang J, Spiller W, Bowden J, Small D. Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples *Statistical Science* 2019; 34: 317–333 https://doi.org/10.1214/18-STS692

23. The CARDIoGRAMplusC4D ConsortiumA comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease*Nature Genetics*2015; 47: 1121–1130 https://doi.org/10.1038/ng.3396 PMID: 26343387

24. Leblanc M, Zuber V, Thompson W, Andreassen O, Frigessi A, Andreassen B. et al A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework *BMC Genomics* 2018; 19: 1471–2164 https://doi.org/10.1186/s12864-018-4859-7

25. Hartwig F, Davies N, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique *IJE* 2016; 45: 1717–1726 https://doi.org/10.1093/ije/dyx028 PMID: 28338968

26. Hartwig F, Tilling K, Davey Smith G Lawlor D, Borges M Bias in two-sample Mendelian randomization by using covariable-adjusted summary associations *IJE* 2021; In Press PMID: 33619569

27. Lawlor D. Commentary: Two-sample Mendelian randomization: opportunities and challenges *IJE* 2016; 45:908–915 https://doi.org/10.1093/ije/dyw127 PMID: 27427429

28. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age *PLOS Medicine* 2015; 12: 1–10 https://doi.org/10.1371/journal.pmed.1001779 PMID: 25826379

29. Minelli C, Del Greco F, van der Plaat D, Bowden J, Sheehan N, Thompson J. The use of two-sample methods for Mendelian randomization analyses on single large datasets *IJE* 2021; In Press PMID: 33899104

30. Hartwig F, Davies N. Why internal weights should be avoided (not only) in MR-Egger regression *IJE* 2016; 45: 1676–1678

31. Dudbridge F, Allen R, Sheehan N, Schmidt F, Lee J, Jenkins R et al Adjustment for index event bias in genome-wide association studies of subsequent events *Nature Communications* 2019; 10 1561 https://doi.org/10.1038/s41467-019-09381-w PMID: 30952951

32. Shapland C, Zhao Q, Bowden J. Profile-likelihood Bayesian model averaging for two-sample summary data Mendelian randomization in the presence of horizontal pleiotropy *Biorxiv* 2020

33. Williamson E, Walker A, Bhaskaran K, Bacon S, Bates C, Morton C et al. Factors associated with COVID-19-related death using OpenSAFELY *Nature* 2020; 584 430–436 https://doi.org/10.1038/s41586-020-2521-4 PMID: 32640463

34. Burgess S and Bowden J Integrating summarized data from multiple genetic variants in Mendelian randomization: bias and coverage properties of inverse-variance weighted methods *arXiv* 2015; 1512.04486

35. Pearl J. Causal inference in statistics: An overview *Statistics Surveys* 2009; 3: 96–146 https://doi.org/10.1214/09-SS057

36. Munafo M, Tilling K, Taylor A, Evans D, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations *IJE* 2017; 47: 226–235

37. Verbank M, Chen C, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases *Nature Genetics* 2018; 50: 693–698 https://doi.org/10.1038/s41588-018-0099-7

38. Bowden J, Hemani G, Davey Smith G, Invited Commentary: Detecting Individual and Global Horizontal Pleiotropy in Mendelian Randomization—A Job for the Humble Heterogeneity Statistic? *American Journal of Epidemiology* 2018; 187: 2681–2685 https://doi.org/10.1093/aje/kwy185 PMID: 30188969

39. Giloni A, Padberg M The Finite Sample Breakdown Point of L-1 Regression *SIAM Journal on Optimization* 2004; 14: 1028–1042 https://doi.org/10.1137/S1052623403424156

40. Spiegel K, Leproult R, Van Cauter E Impact of sleep debt on metabolic and endocrine function *Lancet* 1999; 354: 1435–1439 https://doi.org/10.1016/S0140-6736(99)01376-8 PMID: 10543671

41. Stamatakis K, Punjabi N Effects of sleep fragmentation on glucose metabolism in normal subjects *Chest* 2010; 137: 95–101 https://doi.org/10.1378/chest.09-0791 PMID: 19542260

42. Nedeltcheva A, Kessler L, Imperial J, Penev P. Exposure to Recurrent Sleep Restriction in the Setting of High Caloric Intake and Physical Inactivity Results in Increased Insulin Resistance and Reduced Glucose Tolerance *The Journal of Clinical Endocrinology & Metabolism* 2009; 94: 3242–3250 https://doi.org/10.1210/jc.2009-0483 PMID: 19567526

43. Shan Z, Ma H, Xie M, Yan P, Guo Y, Bao W et al. Sleep Duration and Risk of Type 2 Diabetes: A Meta-analysis of Prospective Studies *Diabetes Care*; 38: 529–537 https://doi.org/10.2337/dc14-2073 PMID: 25715415

44. Green M, Espie C, Popham F, Robertson T, Benzeval M Insomnia symptoms as a cause of type 2 diabetes Incidence: a 20?year cohort study *BMC Psychiatry* 2017; 17: 94 https://doi.org/10.1186/s12888-017-1268-4 PMID: 28302102

45. Jansen P, Watanabe K, Stringer S, Skene N, Bryois J, Hammerschlag A et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways *Nature Genetics* 2019; 51: 394–403 https://doi.org/10.1038/s41588-018-0333-3 PMID: 30804565

46. Guidance Diagnosis and Classification of Diabetes Mellitus*Diabetes Care*2004; 27: s5–s10 https://doi.org/10.2337/diacare.27.2007.S5 PMID: 14693921

47. Willer C, Li Y, Abecasis G. METAL: fast and efficient meta-analysis of genomewide association scans *Bioinformatics* 2010; 26: 2190–2191 https://doi.org/10.1093/bioinformatics/btq340 PMID: 20616382

48. Qi G, Chatterjee N Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects *Nature Communications* 2019; 10: 1941 https://doi.org/10.1038/s41467-019-09432-2 PMID: 31028273

49. Burgess S, Foley C, Allara E, Staley J, Howson J A robust and efficient method for Mendelian randomization with hundreds of genetic variants *Nature Communications* 2020; 11: 376 https://doi.org/10.1038/s41467-019-14156-4 PMID: 31953392

50. Sanderson E, Davey Smith G, Windmeijer F, Bowden J An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings *IJE* 2018; 48: 713–727

51. Wang J, Zhao Q, Bowden J, Hemani G, Davey Smith G, Small D et al. Causal Inference for Heritable Phenotypic Risk Factors Using Heterogeneous Genetic Instruments *PLOS Genetics* 2021, In Press https://doi.org/10.1371/journal.pgen.1009575 PMID: 34157017

52. Staley J, Burgess S. Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization Genetic Epidemiology (2017); 41: 341–352 https://doi.org/10.1002/gepi.22041 PMID: 28317167

53. Vansteelandt S, Bowden J, Babanezhad M, Goetghebeur E. On Instrumental Variables Estimation of Causal Odds Ratios *Statistical Science* 2011; 26: 403–422 https://doi.org/10.1214/11-STS360

54. Palmer T, Thompson J, Tobin M, Sheehan N, Burton P Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses *IJE* 2008; 37: 1161–1168 PMID: 18463132