Robust Mendelian randomization in the presence of residual population stratification, batch effects and horizontal pleiotropy

Carlos Cinelli[*1], **Nathan LaPierre**[2], **Brian L. Hill**[2], **Sriram Sankararaman**[2, 3, 4], **and Eleazar Eskin**[2, 3, 4]

[1]Department of Statistics, University of California, Los Angeles, CA, USA
[2]Department of Computer Science, University of California, Los Angeles, CA, USA
[3]Department of Human Genetics, University of California, Los Angeles, CA, USA
[4]Department of Computational Medicine, University of California, Los Angeles, CA, USA

ABSTRACT

Mendelian Randomization (MR) exploits genetic variants as instrumental variables to estimate the causal effect of an "exposure" trait on an "outcome" trait from observational data. However, residual population stratification, batch effects, and horizontal pleiotropy can lead to spurious findings in MR studies. We describe a suite of sensitivity analysis tools for MR that enables investigators to properly quantify the robustness of their findings against these (and other) unobserved validity threats. Specifically, we propose the routine reporting of sensitivity statistics that can be used to readily quantify the robustness of a MR result: (i) the partial R^2 of the genetic instrument with the exposure and the outcome traits; and, (ii) the *robustness value* of both genetic associations. These statistics quantify the *minimal strength* of violations of the MR assumptions that would be necessary to explain away the MR causal effect estimate. We also provide intuitive displays to visualize the sensitivity of the MR estimate to any degree of violation, and formal methods to *bound the worst-case bias* caused by violations in terms of multiples of the observed strength of principal components, batch effects, as well as putative pleiotropic pathways. We demonstrate the use of these tools by showing that the MR estimate of the causal effect of body mass index (BMI) on type 2 diabetes is relatively robust, whereas the MR estimate of the causal effect of BMI on Townsend deprivation index is fragile.

Many fundamental questions in the social and medical sciences are questions of cause and effect. For instance, what are the social and health consequences of obesity? In practice, however, it is often infeasible or unethical to perform a randomized controlled trial to answer these types of questions. Moreover, observational studies are prone to being biased due to the presence of unmeasured confounders. In such cases, the method of instrumental variables[1–4] (IVs) may be an appealing alternative, allowing one to infer cause-effect relationships even in the presence of unmeasured confounding between the exposure and the outcome.

Mendelian randomization (MR) exploits genetic variants associated with an "exposure" trait of interest as IVs to investigate whether that exposure has a causal effect on an "outcome" trait of interest[5–11]. The technique of MR has become a standard tool for inferring causal relationships, with numerous applications published in medical, genetic and epidemiological journals.[6–14] This growth has been accelerated by the availability of large genetic databases[15] and Genome-Wide Association Studies (GWAS) linking many genetic variants to complex phenotypes[8]. Nevertheless, the validity of MR studies depends on its own set of assumptions, and this rapid growth has not been accompanied with sufficient attention to those assumptions[16–19].

In particular, for a genetic variant to be a valid IV it must satisfy two fundamentally untestable conditions[6–9]: (i) the genetic variant itself must not be confounded with the outcome trait; and, (ii) the genetic variant must affect the outcome trait only through its effect on the exposure trait. These conditions may be violated in several ways due to populational and methodological artifacts, as well as biological mechanisms. Most notably, population stratification[20–25] and batch effects[25–27] are well known sources of confounding biases in high-throughput genomic data. Likewise, many genetic variants tend to exert horizontal pleiotropy, meaning they affect the outcome trait through channels other than the exposure trait[28,29].

The prevailing method for dealing with population stratification and batch effects in MR is to adjust for genomic principal components and surrogate technical covariates representing genomic batch or assessment centre[19]. In the case of horizontal pleiotropy, researchers are advised to perform alternative analyses, such as MR-Egger[30] or MR-Presso[31], that rely on modified identification assumptions. Although these methods have proved useful for partially mitigating these problems, residual biases may still remain[9,32]. Since those biases are impervious to sample size, they may lead to highly

*This version—September 5, 2020. E-mail: carloscinelli@ucla.edu

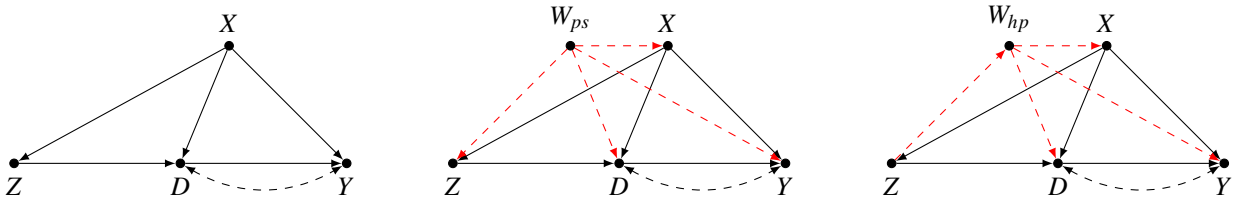


Figure 1. Directed acyclic graphs (DAGs) illustrating the traditional MR assumptions and possible violations. Graphically, conditional on X, the genetic instrument Z is a valid IV for the causal effect of trait D on trait Y, if X blocks all paths from Z to Y on the graph where the edge $D \to Y$ is removed[4]. This condition holds in Figure 1a, in which X alone accounts for all population structure. However, in Figure 1b, X does not account for all population structure (W_{ps}), and valid MR requires conditioning on *both* X and W_{ps}. Similarly, Figure 1c shows a violation of the standard MR assumptions due to horizontal pleiotropy through trait W_{hp}; again, valid MR requires conditioning on *both* X and W_{hp}.

statistically significant false findings with large genomic data, as we demonstrate later.

Here we build on recent developments of the sensitivity analysis literature in statistics[33–37] to provide a suite of sensitivity analysis tools for MR studies that quantifies the robustness of inferences to the presence of residual population stratification, batch effects, and horizontal pleiotropy. Specifically, we introduce *robustness values*[34] (RV) for MR, summarizing the *minimal strength* that residual biases must have (in terms of variance explained of the genetic instrument and of the phenotypes) in order to explain away the MR causal effect estimate. To increase transparency and facilitate the assessment of the credibility of MR studies, we propose the RV to be routinely reported alongside traditional p-values (traditional p-values assume *zero* residual biases). We also provide intuitive sensitivity plots that allow researchers to quickly inspect how their inferences would have changed under biases of any postulated strength. Finally, we show how to place formal *bounds on the worst-case bias* caused by putative unmeasured variables with strength expressed in terms of multiples of the effect of observed variables, thereby facilitating expert judgment regarding the plausibility of such strong violations of the traditional MR assumptions.

We show how these techniques can aid researchers in assessing the degree of robustness of a MR result by examining two findings of highly-cited MR studies using the UK Biobank dataset[15]—that body mass index (BMI) has a causal effect on type 2 diabetes (T2D) and Townsend deprivation index (deprivation)[38–40]. Sensitivity analysis reveals that, while the MR estimate of the causal effect of BMI on T2D is robust to relatively strong residual confounding *and* horizontal pleiotropy, the effect estimate of BMI on deprivation could be nullified by biases as weak as a fraction of current putative pleiotropic pathways, or a fraction of observed batch effects.

Results

MR-SENSEMAKR overview—a suite of sensitivity analysis tools for MR. We developed MR-SENSEMAKR, a suite of sensitivity analysis tools for MR that allows researchers to perform robust inferences of causal effect estimates in the presence of violations of the standard MR assumptions. These tools quantify both how much the inferences would have changed under a postulated degree of violation, as well as the minimal strength of violation necessary to overturn a certain conclusion. MR-SENSEMAKR builds on an extension of the "omitted variable bias" framework for regression analysis[34,35] to the Anderson-Rubin method[41] and Fieller's theorem[42] for testing null hypotheses in the IV setting. This approach has a number of benefits, such as: (i) correct test size regardless of instrument strength; (ii) handling multiple confounding or pleiotropic effects acting simultaneously, possibly non-linearly; (iii) providing simple sensitivity statistics for routine reporting; and, (iv) exploiting expert knowledge to bound the maximum strength of biases (see Methods for details).

Let D denote the "exposure" trait, Y the "outcome" trait, and Z the genetic instrument (e.g, a polygenic risk score). Additionally, let $\boldsymbol{X}$ denote a set of *observed* "control" covariates which account for potential violations of the MR assumptions, such as population stratification (e.g, genetic principal components), batch effects (e.g, batch indicators) and traits that could block putative horizontal pleiotropic pathways[19]. Traditional MR analysis assumes that $\boldsymbol{X}$ is sufficient for making Z a valid instrumental variable for identifying the effect of the exposure trait D on the outcome trait Y. An example for which this is the case is depicted in the directed acyclic graph (DAG) of Figure 1a—in this example there are no pleiotropic pathways, and although there is confounding due to population structure, adjusting for $\boldsymbol{X}$ (say, genomic principal components and batch indicators) is sufficient for eliminating all biases.

The problem arises, however, when $\boldsymbol{X}$ *does not* suffice for making Z a valid instrument; instead, an *extended* set of control covariates would be necessary to do so, but some of these variables are, unfortunately, *unobserved*. Figures 1b and 1c illustrate two of such cases. In Figure 1b, although X accounts for part of the confounding biases due to population structure (ps), it cannot account for all of it, and further adjustment for W_{ps} would be necessary for making Z a valid instrument. In Figure 1c, we have a different type of problem; there, the genetic instrument exerts horizontal pleiotropy (hp) through trait W_{hp}, which needs to be accounted for in a valid MR

analysis. In practice, of course, all these residual biases will often be acting simultaneously—we denote by $\boldsymbol{W}$ the *set* of *all additional unmeasured variables* that would be necessary for making Z a valid genetic instrument.

In this setting, MR-SENSEMAKR answers the following question: *how strong would the unmeasured variables $\boldsymbol{W}$ have to be such that, if accounted for in the analysis, they would have changed the conclusions of the MR study?* As has been extensively discussed elsewhere[7,9,16,19], MR studies are more reliable to *test* the presence or direction of a causal effect, rather than to precisely estimate its magnitude. Thus, here we focus on two problematic changes that $\boldsymbol{W}$ could cause—turning a statistically significant result into an insignificant one; or, leading to unbounded or uninformative confidence intervals due to weak instruments (when using Fieller's theorem, confidence intervals can be: (i) connected and finite; (ii) the union of two disjoint unbounded intervals; or, (iii) the whole real line; see Methods).

It can be shown that, given a significance level α, the confidence interval for the MR causal effect is unbounded if, and only if, we cannot reject the hypothesis that the *genetic association with the exposure* is zero. Likewise, the MR causal effect estimate is statistically insignificant if, and only if, we cannot reject the hypothesis that the *genetic association with the outcome* is zero (to understand this intuitively, recall that the MR estimate is the ratio of the genetic association with the outcome over the genetic association of the exposure. Note this ratio is zero if the numerator is zero; likewise, the ratio can be arbitrarily large if the denominator can be arbitrarily close to zero). Therefore, the problem of sensitivity analysis of the MR estimate can be reduced to the simpler problem of sensitivity analysis of these two genetic associations.

MR-SENSEMAKR thus performs sensitivity analysis for the MR causal effect estimate by examining how strong $\boldsymbol{W}$ needs to be to explain away *either* the observed *genetic association with the exposure or* the observed *genetic association with the outcome*. It deploys two main tools for assessing the sensitivity of these quantities. First, it computes key *sensitivity statistics* suited for *routine reporting*[34], including

- The partial R^2 of the genetic instrument with the (exposure/outcome) trait, revealing the *minimal* share of residual variation that $\boldsymbol{W}$ needs to explain of the genetic instrument in order to fully eliminate the genetic association with that trait;
- The *robustness value* (RV) of the genetic instrument with the (exposure/outcome) trait, revealing the *minimal* share of residual variation (partial R^2), both of the genetic instrument *and* of the trait that $\boldsymbol{W}$ needs to explain in order to make the genetic association with that trait statistically insignificant; and,
- *Bounds* on the *maximum* residual variation explained by unmeasured variables $\boldsymbol{W}$ if they were as strong as: (i) observed principal components; (ii) measured batch effects; and, (iii) observed pleiotropic pathways.

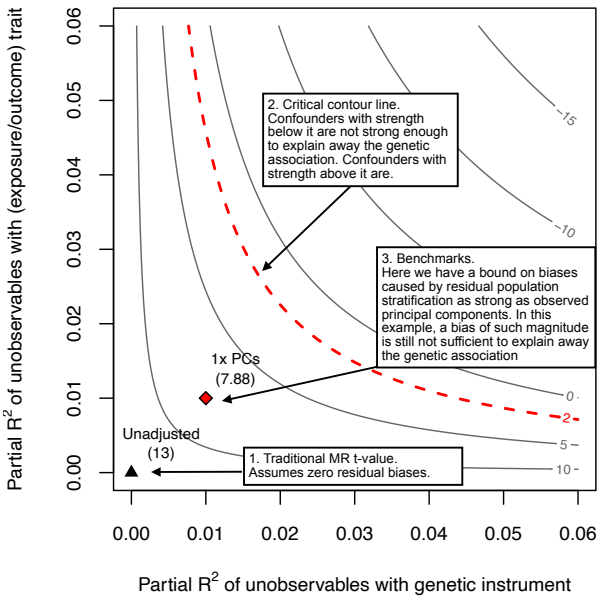


Figure 2. Sensitivity contour plot with benchmark bounds. The horizontal axis shows the partial R^2 of unobserved variables $\boldsymbol{W}$ with the genetic instrument; this corresponds to the percent of residual variation of the genetic instrument explained by $\boldsymbol{W}$. The vertical axis shows the partial R^2 of $\boldsymbol{W}$ with the trait of interest (either the treatment trait or the outcome trait); again, this stands for the percent of residual phenotypic variance explained by $\boldsymbol{W}$. Given any pair of partial R^2 values, the contour lines show the t-value that one would have obtained for testing the significance of that genetic association, had a $\boldsymbol{W}$ with such strengths been included in the analysis. The point represented by a black triangle (left lower corner) shows the t-value of a traditional MR study—note it assumes exactly *zero* biases due to unobserved variables $\boldsymbol{W}$. As we move along both axes, the biases due to $\boldsymbol{W}$ get worse, and can eventually be strong enough to reduce the t-value below a chosen critical level t^*, shown in the red dashed line (e.g, $t^* \approx 2$ for a significance level of $\alpha = 5\%$). Unobserved variables $\boldsymbol{W}$ with strength *below* the critical red line are not strong enough to change the conclusions of the original MR study; on the other hand, unobserved variables $\boldsymbol{W}$ with strength *above* the critical red line are strong enough to be problematic. The point represented by a red diamond bounds the *maximum* strength of $\boldsymbol{W}$ if it were as strong as observed genomic principal components (1x PCs). They show the maximum bias caused by residual population stratification, if it had the same explanatory power as the PCs in explaining genetic and phenotypic variation. In this example, the plot reveals that, even if there were residual population stratification as strong as the first genomic principal components, this would not be sufficient to make the genetic association statistically insignificant. Finally, we note that if the unobserved variable $\boldsymbol{W}$ is a singleton, then all the sensitivity analysis results are exact. If $\boldsymbol{W}$ consists of multiple variables, then all sensitivity analysis results are conservative, meaning that this is the worst bias that a multivariate $\boldsymbol{W}$ could cause if it had such strengths.

MR-SENSEMAKR also provides sensitivity contour plots[34] that, given *any* hypothetical strength of $\boldsymbol{W}$ (measured in terms of the partial R^2 of $\boldsymbol{W}$ with the genetic instrument and with the trait), allows researchers to investigate what would have been the result of a significance test of that particular genetic asso-

ciation, had a $\boldsymbol{W}$ with such strength been incorporated in the analysis (see Figure 2). Finally, these plots can also include several bounds on the maximum amount of residual variation that $\boldsymbol{W}$ could explain, both of the genetic instrument and of the trait, if $\boldsymbol{W}$ were multiple times stronger than observed variables. Next, we apply these tools in a real example that examines the robustness of previous MR findings regarding the causal effect of BMI on T2D and deprivation[38–40].

MR-SENSEMAKR helps distinguishing robust from fragile findings. Previous studies[38–40] used MR on the UK Biobank data[15] to assess the causal effect of body mass index (BMI) on multiple outcome traits of interest. These MR analyses found a statistically significant effect of BMI on type 2 diabetes (T2D)[39] and on Townsend deprivation index (deprivation)—a measure of socioeconomic status.[38] Following these studies, we filtered the data to only include people with self-reported white British ancestry who were not closely related, leaving a sample size of 291,274 people; the genetic instrument consisted of a polygenic risk score (PRS) derived from 97 SNPs previously found to be associated with BMI, with external weights given by the effect sizes from the GIANT study[39,43] (see Methods for details).

The first part of Table 1 reports the results of the traditional MR analysis of the effects of BMI both on T2D and on deprivation. As it is usually recommended[19] and following the original studies, these MR analyses further adjust for: age, gender, 20 leading genomic principal components, assessment center, batch indicators, as well as smoking and drinking status (both are putative pleiotropic pathways, especially for T2D[44–48]). In consonance with the previous studies, we found that the conventional MR analyses lead to a positive and statistically significant effects of BMI on both traits, at the 5% significance level. The results, however, rely on the assumptions of *zero* residual population stratification, *zero* batch effects and *zero* horizontal pleiotropy, which are unlikely to hold. We thus used MR-SENSEMAKR to investigate the robustness of these findings to potential violations of the standard MR assumptions.

We first examined the robustness of the genetic association with the exposure trait (BMI). Recall that, if confounders are strong enough to explain away the genetic association with the exposure, this can lead to unbounded or uninformative confidence intervals for the MR causal effect estimate—the exercise we are performing here is thus tantamount to assessing the "weak instrument" problem, except that now we are accounting both for sampling uncertainty and potential unmeasured confounders. The results are shown in the section entitled "Sensitivity PRS-Exposure" of Table 1. (Note the results are the same both for T2D and deprivation, since the exposure trait, BMI, is the same in both cases.) The first sensitivity measure is the partial R^2 of the PRS with BMI, which amounted to 1.67%. Although this quantity is already reported as a measure of instrument strength in many MR studies[19], it is perhaps less known that it is also a measure of its robustness to *extreme* confounding. In particular, this means that, even if the unmeasured variables $\boldsymbol{W}$ explained *all* left-out variation in BMI, they would still need to account for at least 1.67% of the variance of the genetic instrument, otherwise $\boldsymbol{W}$ cannot explain away the genetic association with the exposure. Next we obtained a robustness value of 11.88% for the PRS-exposure association. This means that any unmeasured variables $\boldsymbol{W}$ that explain less than 11.88% of the residual variation, both of the PRS and of BMI, are not strong enough to make the genetic association with the exposure statistically insignificant.

Next we examined the robustenss of the genetic association with the outcome traits; recall that any unobserved variables capable of explaining away the genetic association with the outcome trait are also capable of explaining away the MR causal effect estimate. The results are shown in the section entitled "Sensitivity PRS-Outcome" of Table 1, and here we have two separate results for each trait. Specifically, we obtained a partial R^2 of the PRS with T2D of 0.064% and a robustness value 2.13%. This means that, even if unobserved variables explained all variation of T2D, they still need to explain at least 0.064% of the residual variation of the genetic instrument to fully account for the observed PRS-T2D association; moreover, the RV reveals that unobserved variables need to account for at least 2.13% of either the variation of genetic instrument or the variation of T2D to be sufficiently strong to overturn the statistical significance found in the original MR study. Moving to the next trait, the bottom row of Table 1 shows the sensitivity statistics for the effect of BMI on deprivation. Here we found a partial R^2 of 0.002% and a robustness value of 0.08%, revealing that much weaker residual biases would be able to overturn the MR effect estimate of BMI on deprivation.

Confronted with those results, the next step is to make plausibility judgments on whether unobserved variables with the strengths revealed to be problematic can be ruled out. To aid in these plausibility judgments, MR-SENSEMAKR computes bounds on the amount of variance explained by the unmeasured variables $\boldsymbol{W}$ if it were as strong as observed vari-

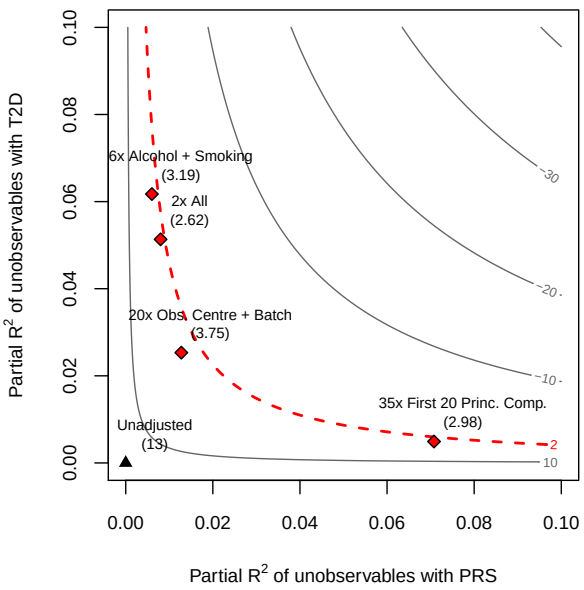
	Traditional MR		Sensitivity PRS-Outcome		Sensitivity PRS-Exposure	
Outcome	RD (95% CI)	P-value	Partial R^2	$\text{RV}_{\alpha=0.05}$	Partial R^2	$\text{RV}_{\alpha=0.05}$
T2D	0.042 (0.036 to 0.048)	$< 2 \times 10^{-16}$	0.064%	2.13%	1.67%	11.88%
Deprivation	0.033 (0.006 to 0.060)	0.017	0.002%	0.08%		

Table 1. Traditional MR results and sensitivity analyses.

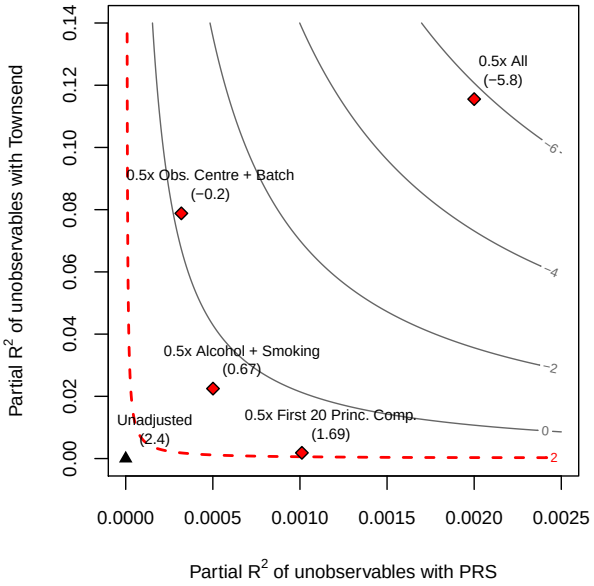
ables. For our running example, these bounds are shown in Table 2; they reveal the maximum partial R^2 of unobserved variables $\boldsymbol{W}$ if it were as strong as: (i) 20 leading genomic principal components (1 x PCs); (ii) observed batch and centre effects (1 x Batch+Centre); and, finally, (iii) smoking and drinking status (1 x Alc.+Smok.).

Starting with instrument strength, first note that all bounds on the PRS and BMI columns of Table 2 are (substantially) lower than than the RV of 11.88% for the genetic association with BMI; this means that, even if $\boldsymbol{W}$ were as strong as those variables, this would not be sufficient to result in a "weak instrument" problem. Moreover, since all values of the PRS column are less than the partial R^2 of 1.67% of the variant-exposure association , even a "worst-case" confounding that explains 100% of the variance of BMI, and as strongly associated with the genetic instrument as the observed variables, cannot account for the observed association of the genetic instrument with the exposure. Moving to statistical significance concerns, similar results hold for the PRS-T2D association. Since the bounds on both columns, for the PRS and T2D, are below the robustness value of 2.13%, Table 2 reveals that biases as strong as the observed variables are not sufficient to make the MR causal effect estimate of BMI on T2D statistically insignificant. However, in stark contrast, note that all bounds on the PRS and deprivation columns are above the RV of 0.08% for deprivation, meaning that unobserved variables $\boldsymbol{W}$ strong as those could easily overturn the original MR analysis.

Table 1 forms our proposed minimal reporting for sensitivity analysis in MR studies. Often, when supplemented with bounds such as those of Table 2, these metrics are sufficient to give a broad picture of the robustness of MR findings, as demonstrated above. Researchers, however, can refine their analyses and fully explore the whole range of robustness of their inferences with sensitivity contour plots, placing several different bounds on the strength of confounding multiple times stronger than observed variables. The plots for T2D and deprivation are shown in Figure 3 (see caption of Figure 2 for details on how to read the plot). For T2D, note that neither residual population stratification up to 35x stronger than observed principal components nor residual batch-effects up to 20x stronger than observed batch-effects are sufficient to make the MR estimate statistically insignificant. Likewise, if residual pleiotropy were up to 6x stronger than important observed pleiotropic pathways, such as alcohol and smoking,



(a) Sensitivity contour for T2D.



(b) Sensitivity contour for deprivation.

Figure 3. Sensitivity contours for the null hypothesis of zero effect.

	Bound partial R^2 with genetic IV	Bound partial R^2 with trait		
$\boldsymbol{W}$ *as strong as*	PRS (Genetic IV)	BMI (Exposure)	T2D (Outcome)	Deprivation (Outcome)
1 x PCs	0.20%	0.11%	0.01%	0.37%
1 x Batch+Centre	0.06%	0.07%	0.13%	15.76%
1 x Alc.+Smok.	0.10%	2.97%	1.03%	4.50%

Table 2. Bounds on the maximum explanatory power of $\boldsymbol{W}$ (partial R^2), if it were as strong as: (i) 20 leading genomic principal components (1 x PCs); (ii) observed batch and centre (1 x Batch+Centre); and, (iii) smoking and drinking status (1 x Alc.+Smok.).

this is also not sufficient to change the original conclusions. Finally, even if unobserved variables $\boldsymbol{W}$ had twice the explanatory power of *all* the observed variables combined, this again would not change the results for T2D. In contrast, the sensitivity plot for deprivation reveals that the MR causal effect estimate BMI on deprivation is sensitive to confounding with explanatory power as weak as a fraction (e.g, 0.5) of current observed variables.

Putting these results in context requires assessing the quality of the benchmarks involved. For example, it is not unreasonable to argue that genomic principal components correct for most, or at least a large part, of population structure[22], and that it is thus implausible to imagine residual population stratification multiple times stronger than what has been already corrected by observed principal components. Benchmarks for horizontal pleiotropy, on the other hand, require specific knowledge of the aetiology of the disease under study, or of the social process under investigation. In this application, for instance, previous studies suggest that alcohol and smoking are indeed suspected to be important channels for horizontal pleiotropy in the case of T2D[44–48]. Therefore, one could plausibly argue that it is unlikely (although, of course, not impossible) that residual horizontal pleiotropy multiple times as strong as those still remain. The case for deprivation, however, reveals a more fragile finding; not only there is no a priori reason to suspect that alcohol and smoking should be among the strongest pleiotropic pathways, but the bounding exercise shows that residual pleiotropy a fraction as strong as those could overturn the original results.

Overall, the sensitivity analyses suggest that: (i) the genetic association of the instrument (PRS) with the exposure (BMI) is relatively robust, and that instrument strength is unlikely to be an issue; (ii) that it would take substantial residual confounding as well as substantial residual pleiotropic pathways to reverse the original MR finding of the causal effect of BMI on T2D; and that, in contrast, (iii) the previous MR causal effect estimate of BMI on deprivation is fragile, meatning that there is little room for small residual biases, which could easily overturn the original analysis.

Current proposals for MR "sensitivity analyses" can lead to false positive findings in the presence of small residual biases in large samples. Prevailing proposals for sensitivity analyses of MR studies have focused on replacing traditional instrumental variable assumptions with alternative assumptions about how pleiotropy operates, such as the InSIDE assumption[30,31]. Although an improvement of traditional MR, under the presence of residual population stratification, batch effects, and certain forms of pleiotropy, such approaches may still lead to statistically significant false findings given large enough samples. Therefore, the sensitivity statistics and exercises we propose here can be a useful complement to those alternative analyses.

To demonstrate this, we performed a simulation study in which the InSIDE assumption is only *slightly* violated through small pleoitropic effects via confounders of the exposure and outcome trait. Our simulation largely follows the same specification of previous work[31,49,50], with the following data-generating model:

$$W_i = \sum_{j=1}^{J} \phi_i G_{ij} + \varepsilon_{i,W}, \qquad X_i = \sum_{j=1}^{J} \delta_i G_{ij} + \varepsilon_{i,X} \tag{1}$$

$$D_i = \sum_{j=1}^{J} \beta_i G_{ij} + X_i + W_i + U_i + \varepsilon_{i,D} \tag{2}$$

$$Y_i = \tau D_i + \eta X_i + \gamma W_i + U_i + \varepsilon_{i,Y} \tag{3}$$

where D is the exposure trait; Y is the outcome trait; W is an *unobserved* confounder, and X an *observed* confounder of D and Y, both carriers of pleiotropy in a way that violates the InSIDE assumption. The genetic variants G_{ij} are drawn independently from a Binomial distribution, $\text{Binom}(2, 1/3)$; the remaining error terms U_i, $\varepsilon_{i,W}$, $\varepsilon_{i,X}$, $\varepsilon_{i,D}$ and $\varepsilon_{i,Y}$ are drawn from standard gaussians.

We set the number of variants $J = 90$, similar to our previous BMI analysis, and consider genetic effects drawn from an uniform distribution from 0.01 to 0.05 for ϕ_i, δ_i and β_i. The parameters η and γ give further control to the level of pleiotropy, and here we set both to 0.05. To put this value in context, for the usual simulated sample size considered in previous work (10,000-30,000 individuals), this level of pleitropy is small enough that it does not meaningfully affect type I errors for MR-Egger. Here, however, we simulate larger sample sizes, similar to those found in large genetic databases, ranging from 100,000 to 500,000 individuals.

We investigated the performance of alternative MR methods in a two-sample Mendelian randomization setting, meaning that only summary level data was used in the analyses, and the genetic associations with the exposure trait and the outcome trait were obtained in separate simulated data. Table 3 shows the results of 10,000 simulations of the data generating process for each of the sample sizes, considering two cases: (i) a true null causal effect with $\tau = 0$; (ii) and a true positive causal effect of $\tau = 0.1$. Note that X_i and W_i have similar strengths—a fact that, if known, can be exploited for sensitivity analysis (see Supplement for simulations of residual population structure and alternative parameterizations).

We first focus on the case of a null causal effect. The first three columns of the table shows the proportion of cases in which the null hypothesis of zero effect was rejected, using the three different MR methods: (i) the traditional inverse variance weighted (IVW); (ii) MR-Presso; and, (iii) MR-Egger. Since the true causal effect is zero, these results indicate the proportion of *false positives*. We see that IVW and MR-Presso give similar results with a virtually 100% false positive rate for all sample sizes, and that MR-Egger starts with a false positive rate of 13% for $N = 100,000$, and this rate grows up to 52% at $N = 500,000$. Next, the last three columns show how the sensitivity exercises could help interpreting the results in such cases. Starting with the fourth column, here we have the proportion of false positives if the researcher knew that X were among the most important pleiotropic pathways (such as, in

	Proportion of rejections of the null (5% significance level)				$RV_{\alpha=0.05}$	
Sample Size	IVW	MR-Presso	MR-Egger	Bound (W as strong as X)	Min	Max
Scenario 1: true null causal effect ($\tau = 0$)						
100,000	99%	99%	13%	3.0%	0%	1.1%
200,000	100%	100%	23%	2.6%	0%	0.9%
300,000	100%	100%	33%	2.1%	0%	1.0%
400,000	100%	100%	40%	1.7%	0%	0.9%
500,000	100%	100%	52%	1.2%	0%	0.8%
Scenario 2: true positive causal effect ($\tau = 0.1$)						
100,000	100%	100%	86%	90%	1.5%	3.6%
200,000	100%	100%	99%	99%	2.0%	3.5%
300,000	100%	100%	100%	100%	2.1%	3.5%
400,000	100%	100%	100%	100%	2.3%	3.4%
500,000	100%	100%	100%	100%	2.3%	3.4%

Table 3. Simulation of weak pleiotropic pathways violating the InSIDE assumption.

our previous example, smoking) and that residual pleiotropy could be as strong as X. Using the bounding procedure delineated in the previous section, if the researcher accounted for this possibility of confounding as strong as X, she would then only falsely conclude that there is an effect roughly around 1% to 3% of the time. Moreover, the last two columns show the *minimum* and the *maximum* robustness values for the association of the genetic instrument with the exposure, over *all* simulations. Note these are the *most extreme* results one could get, and they still always remain roughly below 1%, correctly warning the researcher that residual biases of those magnitudes are capable of overturning those MR findings.

We now turn to the second scenario, in which there is a true positive causal effect of D on Y. Here all MR methods correctly reject the null hypothesis of zero effect from 86% to 100% of the time. The challenge in this setting, thus, comes not from rejecting the null hypothesis, but from the fact that potential critics of the study could *correctly* be skeptical of the results, and conjecture that the reason why the null was rejected was simply due to residual pleiotropic pathways. To mitigate those concerns, the researcher could again use the bounding procedure, and around 90% to 100% of the time she would conclude that one would still reject the null, even when allowing for residual pleiotropy as strong as that due to the observed X. Likewise, the results for the RV show that a researcher would never obtain a robustness value below 1.5%, meaning that, in all cases, the critic would need to argue that biases of *at least* these magnitudes are plausible in order to forcefully dismiss the observed MR finding.

Discussion

We have described a suite of sensitivity analysis tools for performing valid MR inferences under the presence of residual biases of *any* postulated strength. The approach we proposed here starts from the premise that all MR studies will be imperfect in some way or another, but also that a study does not have to be perfect in order to be informative—what matters is not whether certain assumptions hold exactly, but the extent to which certain conclusions are robust to violations of those assumptions, and whether such strong violations are plausible.

We showed how two simple sensitivity statistics, the partial R^2 and the *robustness value*, can be used to easily communicate the *minimum strength* of residual biases necessary to invalidate the results of a MR study. Since researchers are already well advised to report the partial R^2 of the genetic instrument with the exposure trait, routinely reporting the partial R^2 of the genetic instrument with the outcome trait and the robustness value is but a small addition to current practices, and can greatly improve the transparency regarding the robustness of MR findings.

We also showed that, whenever researchers are able to argue that, although not perfect, they have credibly accounted for most of the population structure with genomic principal components, most of possible batch effects with technical dummies, and have measured known important pleiotropic pathways, this knowledge can be leveraged to formally bound the worst possible inferences due to residual biases. Such bounding exercises can be an important piece of the scientific debate when arguing in favor or against the robustness of a certain finding.

Finally, we remind readers that these tools cannot and should not be used to replace expert judgment. On the contrary, the tools described here can aid leveraging certain types of expert knowledge that would have been otherwise neglected, such as judgments regarding the maximum plausible strength of residual biases, or knowledge regarding the relative importance of certain causal pathways. In sum, strong conclusions from Mendelian randomization studies still need to rely on the quality of the research design, and substantive understanding both of the genetic variants as well as the traits under investigation.

References

1. Wright, P. G. *Tariff on Animal and Vegetable Oils* (Macmillan Company, New York, 1928).
2. Bowden, R. J. & Turkington, D. A. *Instrumental variables*, vol. 8 (Cambridge university press, 1990).
3. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of causal effects using instrumental variables. *J. Am. statistical Assoc.* **91**, 444–455 (1996).
4. Brito, C. & Pearl, J. Generalized instrumental variables. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, 85–93 (Morgan Kaufmann Publishers Inc., 2002).
5. Katan, M. Apolipoprotein e isoforms, serum cholesterol, and cancer. *The Lancet* **327**, 507–508 (1986).
6. Davey Smith, G. & Ebrahim, S. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. journal epidemiology* **32**, 1–22 (2003).
7. Didelez, V. & Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat. methods medical research* **16**, 309–330 (2007).
8. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. medicine* **27**, 1133–1163 (2008).
9. Burgess, S. & Thompson, S. G. *Mendelian randomization: methods for using genetic variants in causal estimation* (CRC Press, 2015).
10. Baiocchi, M., Cheng, J. & Small, D. S. Instrumental variable methods for causal inference. *Stat. medicine* **33**, 2297–2340 (2014).
11. Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for mendelian randomization. *Stat. methods medical research* **26**, 2333–2355 (2017).
12. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. molecular genetics* **23**, R89–R98 (2014).
13. Timpson, N. J. *et al.* C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. *The Lancet* **366**, 1954–1959 (2005).
14. Casas, J. P., Bautista, L. E., Smeeth, L., Sharma, P. & Hingorani, A. D. Homocysteine and stroke: evidence on a causal link from mendelian randomisation. *The Lancet* **365**, 224–232 (2005).
15. Sudlow, C. *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* **12** (2015).
16. VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M. & Kraft, P. Methodological challenges in mendelian randomization. *Epidemiology* **25**, 427 (2014).
17. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity analyses for robust causal inference from mendelian randomization analyses with multiple genetic variants. *Epidemiol. (Cambridge, Mass.)* **28**, 30 (2017).
18. Burgess, S. & Thompson, S. G. Interpreting findings from mendelian randomization using the mr-egger method. *Eur. journal epidemiology* **32**, 377–389 (2017).
19. Burgess, S. *et al.* Guidelines for performing mendelian randomization investigations. *Wellcome Open Res.* **4**, 186 (2019).
20. Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat. genetics* **36**, 388–393 (2004).
21. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. genetics* **36**, 512–517 (2004).
22. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. genetics* **38**, 904 (2006).
23. Novembre, J. *et al.* Genes mirror geography within europe. *Nature* **456**, 98–101 (2008).
24. Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: Confounding factors and mixed models. *PLoS genetics* **14**, e1007309 (2018).
25. Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. genetics* **37**, 1243–1246 (2005).
26. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
27. Tom, J. A. *et al.* Identifying and mitigating batch effects in whole genome sequencing data. *BMC bioinformatics* **18**, 351 (2017).
28. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *The Am. J. Hum. Genet.* **89**, 607–618 (2011).
29. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open biology* **7**, 170125 (2017).
30. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *Int. journal epidemiology* **44**, 512–525 (2015).
31. Verbanck, M., Chen, C.-y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nat. genetics* **50**, 693–698 (2018).

32. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. genetics* **50**, 1728–1734 (2018).

33. Cinelli, C., Kumor, D., Chen, B., Pearl, J. & Bareinboim, E. Sensitivity analysis of linear structural causal models. *Int. Conf. on Mach. Learn.* (2019).

34. Cinelli, C. & Hazlett, C. Making sense of sensitivity: extending omitted variable bias. *J. Royal Stat. Soc. Ser. B* **82**, 39–67, DOI: 10.1111/rssb.12348 (2020).

35. Cinelli, C. & Hazlett, C. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Work. Pap.* (2020).

36. Cinelli, C. & Hazlett, C. sensemakr: sensitivity analysis tools for OLS. *R package version 0.2* (2020).

37. Cinelli, C., Ferwerda, J. & Hazlett, C. Sensemakr: Stata module to provide sensitivity tools for OLS. *Stat. Softw. Components (SSC), Boston Coll. Dep. Econ.* (2020).

38. Tyrrell, J. *et al.* Height, body mass index, and socioeconomic status: mendelian randomisation study in uk biobank. *bmj* **352**, i582 (2016).

39. Lyall, D. M. *et al.* Association of body mass index with cardiometabolic disease in the uk biobank: a mendelian randomization study. *JAMA cardiology* **2**, 882–889 (2017).

40. Millard, L. A., Davies, N. M., Tilling, K., Gaunt, T. R. & Smith, G. D. Searching for the causal effects of body mass index in over 300 000 participants in uk biobank, using mendelian randomization. *PLoS genetics* **15**, e1007951 (2019).

41. Anderson, T. W., Rubin, H. *et al.* Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals Math. Stat.* **20**, 46–63 (1949).

42. Fieller, E. C. Some problems in interval estimation. *J. Royal Stat. Soc. Ser. B (Methodological)* **16**, 175–185 (1954).

43. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

44. Rimm, E. B., Chan, J., Stampfer, M. J., Colditz, G. A. & Willett, W. C. Prospective study of cigarette smoking, alcohol use, and the risk of diabetes in men. *Bmj* **310**, 555–559 (1995).

45. Will, J. C., Galuska, D. A., Ford, E. S., Mokdad, A. & Calle, E. E. Cigarette smoking and diabetes mellitus: evidence of a positive association from a large prospective cohort study. *Int. journal epidemiology* **30**, 540–546 (2001).

46. Willi, C., Bodenmann, P., Ghali, W. A., Faris, P. D. & Cornuz, J. Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. *Jama* **298**, 2654–2664 (2007).

47. Song, Y. *et al.* Pancreatic beta-cell function and type 2 diabetes risk: quantify the causal effect using a mendelian randomization approach based on meta-analyses. *Hum. molecular genetics* **21**, 5010–5018 (2012).

48. US Department of Health and Human Service. The health consequences of smoking—50 years of progress: a report of the surgeon general (2014).

49. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. epidemiology* **40**, 304–314 (2016).

50. Rees, J. M., Wood, A. M., Dudbridge, F. & Burgess, S. Robust methods in mendelian randomization via penalization of heterogeneous causal estimates. *PloS one* **14**, e0222362 (2019).

51. Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

52. Hou, K. *et al.* Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* **51**, 1244–1251, DOI: 10.1038/s41588-019-0465-0 (2019). Number: 8 Publisher: Nature Publishing Group.

Acknowledgements

This research was conducted using the UK Biobank Resource under Application 33127.

Methods

Study Design and Participants

Study population. The UK Biobank[51] is a resource that links genetic data to a variety of physiological and social traits in a cohort of 503,325 British people aged 37-73 years. It has been a valuable resource for estimating causal effects of exposures on a multitude of outcomes using MR[38–40]. We filtered the data to only include people with self-reported white British ancestry who were not closely related, (e.g. no first, second, or third degree relatives), as defined by pairs of individuals who had a kinship coefficient $< (1/2)^{(9/2)}$ (following[52]), leaving 291,274 people. We also removed individuals who were not measured for BMI (non-impedence). For our analysis of the Lyall et al.[39] study, we also excluded patients who responded to a question on whether they were taking anti-hypertensive medication with "don't know".

Polygenic Risk Score. The Polygenic Risk Scores (PRS) was constructed in the same manner as in Lyall et al.[39]. This PRS score was derived from 97 SNPs that were genome-wide significantly associated with BMI in the GIANT consortium study[43]. Two of these SNPs were not directly genotyped in the UK Biobank, and two failed Hardy-Weinberg equilibrium, leaving 93 SNPs to comprise the PRS. The PRS was computed as a weighted score based on these SNPs, with the weights derived from the effect estimated reported by GIANT (β per 1-SD unit of BMI)[39,43]. We used the exact same weights computed by Lyall et al.[39].

Exposure, outcome and control traits. During the initial visit to the UK Biobank assessment center, height was measured to the nearest centimeter using a Seca 202 device and weight was measured to the nearest 0.1 kilogram using a Tanita BC418MA body composition analyser. These measurements were subsequently used to calculate body mass index (BMI), in kg/m^2 (field category ID: 21001). The two outcomes of interest were the Townsend deprivation index and diabetes. The Townsend deprivation index was calculated using the postcode of the participant at the time of recruitment (field category ID: 189). Presence of diabetes was confirmed by questionnaire, where participants were asked if they had ever been previously diagnosed with diabetes by doctor (field category ID: 2443).

In our analyses, we adjusted for age, sex, assessment centre, genetic batch effects, drinking and smoking status, given by the following variables: "Sex" (field category ID: 31); "Age when attended assessment centre" (field category ID: 21003); "UK Biobank assessment centre" (field category ID: 54); "Genotype measurement batch" (field category ID: 22000); "Smoking status" (field category ID: 20116); "Frequency of drinking alcohol" (field category ID: 20414); "Alcohol intake frequency" (field category ID: 1558).

Statistical Methods

Traditional Mendelian Randomization. Suppose we are interested in assessing the causal effect of an exposure trait D on an outcome trait Y, by performing a Mendelian Randomization study with a polygenic risk score (PRS) $Z = \sum_{j=1}^{k} \beta_j G_j$ (comprised of a linear combination of SNPs G_j with weights β_j) as the putative instrumental variable. Note the weights β_j of the PRS could have been obtained either from external data (such as a previous GWAS), or via cross-validation as well as other methods[9]. To give credibility to the study, the researcher considers a *set* of *observed* control covariates $\boldsymbol{X}$ that accounts for potential IV violations of population stratification, batch effects and horizontal pleiotropy[19]. That is, $\boldsymbol{X}$ consists of,

$$\boldsymbol{X} = \{\boldsymbol{X}_{\text{ps}}, \boldsymbol{X}_{\text{batch}}, \boldsymbol{X}_{\text{hp}}, \boldsymbol{X}_{\text{ind}}\}$$

Where $\boldsymbol{X}_{\text{ps}}$ denotes the variables to adjust for population stratification, such as, for instance, genomic principal components; $\boldsymbol{X}_{\text{batch}}$ denotes variables to adjust for batch effects, for example, dummy variables for the assessment centre and genotype batches; $\boldsymbol{X}_{\text{hp}}$ denotes *measured* variables which are suspected to be capable of blocking suspected pleoitropic pathways; and, finally, $\boldsymbol{X}_{\text{ind}}$ are participant characteristics that are usually included in MR, such as the age and sex of the individual.

The traditional MR estimate of the causal effect of D on Y, here denoted by $\hat{\tau}_{\text{res}}$, would consist of the ratio of the genetic association with the outcome trait, $\hat{\beta}_{YZ|\boldsymbol{X}}$, and the genetic association with the exposure trait, $\hat{\beta}_{DZ|\boldsymbol{X}}$, after adjusting for observed covariates $\boldsymbol{X}$, namely,

$$\hat{\tau}_{\text{res}} = \frac{\hat{\beta}_{YZ|\boldsymbol{X}}}{\hat{\beta}_{DZ|\boldsymbol{X}}}$$

Confidence intervals that have nominal coverage regardless of instrument strength can be obtained via Fieller's theorem[42] or via the Anderson-Rubin regression[41]. These confidence intervals can be of three forms: (i) a connected closed interval $[a,b]$; (ii) the union of disjoint unbounded intervals, $(-\infty,a] \cup [b,\infty)$; or, (iii) the whole real line $(-\infty,\infty)$.

Violation of traditional assumptions. The traditional MR estimate, $\hat{\tau}_{\text{res}}$, however adjusts for $\boldsymbol{X}$ only, and it is unlikely that $\boldsymbol{X}$ controls for *all* possible threats to the study validity. Instead, the researcher would have preferred to have also adjusted for additional unobserved variables $\boldsymbol{W}$ to satisfy the MR assumptions. For instance, we would like to have controlled for the *true* population indicators $\boldsymbol{W}_{\text{ps}}$ instead of its approximation as recovered by the principal components $\boldsymbol{X}_{\text{ps}}$; likewise, the researcher suspects that $\boldsymbol{X}_{\text{hp}}$ *is not enough* to block all pleiotropic pathways, and would have liked to have further adjusted for covariates $\boldsymbol{W}_{hp}$.

In sum, instead, of performing the MR analysis using $\boldsymbol{X}$ alone, resulting in $\hat{\tau}_{\text{res}}$ as our MR estimate, the researcher would have wanted to compute instead

$$\hat{\tau} = \frac{\hat{\beta}_{YZ|\boldsymbol{X}\boldsymbol{W}}}{\hat{\beta}_{DZ|\boldsymbol{X}\boldsymbol{W}}}$$

which adjusts for the extended set of covariates $\{\boldsymbol{X},\boldsymbol{W}\}$, such that Z is a valid instrument for estimating the causal effect of D

on Y, conditional on $\{\boldsymbol{X},\boldsymbol{W}\}$. Likewise, confidence intervals should have also been computed adjusting for $\{\boldsymbol{X},\boldsymbol{W}\}$. How would accounting for the omitted variables $\boldsymbol{W}$ have changed our inferences regarding the causal effect of D on Y?

The sensitivity analysis of the MR estimate can be reduced to the sensitivity of the genetic associations. We now explain how to perform sensitivity analysis within the Anderson-Rubin (AR) approach[41], which as we show is also numerically equivalent to Fieller's proposal[42] when considering a single instrumental variable Z.

Let τ denote the causal effect of interest, and define a new random variable $Y_{\tau_0} := Y - \tau_0 D$, in which we subtract from Y the causal effect of D, considering a hypothetical value for τ, say, τ_0. Now consider the following linear regression,

$$Y_{\tau_0} = \hat{\phi}_{\tau_0} Z + \boldsymbol{X}\hat{\eta}_{\tau_0} + \boldsymbol{W}\hat{\gamma}_{\tau_0} + \hat{\varepsilon}_{\tau_0} \tag{4}$$

Note that, if $\tau = \tau_0$ and if Z is valid instrument conditional on $\boldsymbol{X},\boldsymbol{W}$, then we must have that $Y_{\tau_0} \perp\!\!\!\perp Z|\boldsymbol{X},\boldsymbol{W}$, and thus that $\phi_{\tau_0} = 0$. Following this logic, the AR confidence interval with significance level α is defined as all values of τ_0 such that we cannot reject the null hypothesis $H_0 : \phi_{\tau_0} = 0$ at the chosen significance level. More precisely,

$$\text{CI}_{AR}(\alpha) = \left\{ \tau_0 \,;\, t^2_{\hat{\phi}_{\tau_0}} \leq t^{*2}_{\alpha,\text{df}} \right\} \tag{5}$$

Where $t_{\hat{\phi}_{\tau_0}}$ is the t-value for the null hypothesis $H_0 : \phi_{\tau_0} = 0$ and $t^*_{\alpha,\text{df}}$ is the critical threshold of the t-distribution for a significance level α and df degrees of freedom. This confidence interval can be obtained analytically as a function of the genetic association with the exposure and the genetic association with the outcome, which is now useful to write out explicitly.

By appealing to the Frisch–Waugh–Lovell (FWL) theorem, we can write $\hat{\phi}_{\tau_0}$ as,

$$\begin{aligned}
\hat{\phi}_{\tau_0} &= \frac{\text{cov}(Y^{\perp \boldsymbol{XW}} - \tau_0 D^{\perp \boldsymbol{XW}}, Z^{\perp \boldsymbol{XW}})}{\text{var}(Z^{\perp \boldsymbol{XW}})} \\
&= \frac{\text{cov}(Y^{\perp \boldsymbol{XW}}, Z^{\perp \boldsymbol{XW}})}{\text{var}(Z^{\perp \boldsymbol{XW}})} - \tau_0 \frac{\text{cov}(D^{\perp \boldsymbol{XW}}, Z^{\perp \boldsymbol{XW}})}{\text{var}(Z^{\perp \boldsymbol{XW}})} \\
&= \hat{\beta}_{YZ|\boldsymbol{XW}} - \tau_0 \hat{\beta}_{DZ|\boldsymbol{XW}}
\end{aligned} \tag{6}$$

Where $Y^{\perp \boldsymbol{XW}}$ denotes the variable Y after removing the components linearly explained by $\boldsymbol{X}$ and $\boldsymbol{W}$, and $\hat{\beta}_{YZ|\boldsymbol{XW}}$ denotes the regression coefficient of Z on Y (the genetic association with the outcome) after adjusting for both $\boldsymbol{X}$ and $\boldsymbol{W}$; $\hat{\beta}_{DZ|\boldsymbol{XW}}$ denotes the regression coefficient of Z on D (the genetic association with the exposure) after adjusting for $\boldsymbol{X}$ and $\boldsymbol{W}$.

Likewise, the estimated variance of $\hat{\phi}_{\tau_0}$ can be written as,

$$\begin{aligned}
\widehat{\text{var}}(\hat{\phi}_{\tau_0}) &= \frac{\text{var}(Y^{\perp Z\boldsymbol{XW}} - \tau_0 D^{\perp Z\boldsymbol{XW}})}{\text{var}(Z^{\perp \boldsymbol{XW}})} \times \text{df}^{-1} \\
&= \left(\frac{\text{var}(Y^{\perp Z\boldsymbol{XW}})}{\text{var}(Z^{\perp \boldsymbol{XW}})} + \tau_0^2 \frac{\text{var}(D^{\perp Z\boldsymbol{XW}})}{\text{var}(Z^{\perp \boldsymbol{XW}})} \right. \\
&\quad \left. - 2\tau_0 \frac{\text{cov}(Y^{\perp Z\boldsymbol{XW}}, D^{\perp Z\boldsymbol{XW}})}{\text{var}(Z^{\perp \boldsymbol{XW}})} \right) \times \text{df}^{-1} \\
&= \widehat{\text{var}}(\hat{\beta}_{YZ|\boldsymbol{XW}}) + \tau_0^2 \widehat{\text{var}}(\hat{\beta}_{DZ|\boldsymbol{XW}}) \\
&\quad - 2\tau_0 \widehat{\text{cov}}\left(\hat{\beta}_{YZ|\boldsymbol{XW}},\, \hat{\beta}_{DZ|\boldsymbol{XW}}\right)
\end{aligned} \tag{7}$$

To construct the confidence interval, we need to find all values of τ_0 such that the following inequality holds,

$$\frac{\hat{\phi}^2_{\tau_0}}{\widehat{\text{var}}(\hat{\phi}_{\tau_0})} \leq t^{*2}_{\alpha,\text{df}} \implies \hat{\phi}^2_{\tau_0} - \widehat{\text{var}}(\hat{\phi}_{\tau_0}) t^{*2}_{\alpha,\text{df}} \leq 0 \tag{8}$$

Squaring and rearranging terms we obtain the following quadratic inequality,

$$\begin{aligned}
&\underbrace{\left(\hat{\beta}^2_{DZ|\boldsymbol{XW}} - \widehat{\text{var}}(\hat{\beta}_{DZ|\boldsymbol{XW}}) \times t^{*2}_{\alpha,\text{df}}\right)}_{a} \tau_0^2 \\
&+ \underbrace{2\left(\widehat{\text{cov}}(\hat{\beta}_{YZ|\boldsymbol{XW}}, \hat{\beta}_{DZ|\boldsymbol{XW}}) \times t^{*2}_{\alpha,\text{df}} - \hat{\beta}_{YZ|\boldsymbol{XW}}\hat{\beta}_{DZ|\boldsymbol{XW}}\right)}_{b} \tau_0 \\
&+ \underbrace{\left(\hat{\beta}^2_{YZ|\boldsymbol{XW}} - \widehat{\text{var}}(\hat{\beta}_{YZ|\boldsymbol{XW}}) \times t^{*2}_{\alpha,\text{df}}\right)}_{c} \leq 0
\end{aligned} \tag{9}$$

These conditions are exactly Fieller's solution to the confidence interval of the ratio $\hat{\tau} = \frac{\hat{\beta}_{YZ|\boldsymbol{XW}}}{\hat{\beta}_{DZ|\boldsymbol{XW}}}$.

Our task has thus simplified to find all values of τ_0 that makes the above quadratic equation, with coefficients a, b and c, non-positive. But here we have special interest in two specific cases: (i) when the confidence interval for τ is unbounded; and, (ii) when the confidence interval for τ includes zero.

Let us first consider the case of unbounded confidence intervals. Note this happens when $a < 0$, which means the quadratic curve in Equation 9 will be concave (will have a "$\cap$" shape)—as we increase τ_0 to plus or minus infinity, the inequality is bound to hold and the confidence interval will be unbounded. Also note that $a < 0$ if, and only if,

$$\begin{aligned}
&\hat{\beta}^2_{DZ|\boldsymbol{XW}} - \widehat{\text{var}}(\hat{\beta}_{DZ|\boldsymbol{XW}}) \times t^{*2}_{\alpha,\text{df}} \leq 0 \implies \\
&\frac{\hat{\beta}_{DZ|\boldsymbol{XW}}}{\widehat{\text{se}}(\hat{\beta}_{DZ|\boldsymbol{XW}})} = t_{\hat{\beta}_{DZ|\boldsymbol{XW}}} \leq t^*_{\alpha,\text{df}}
\end{aligned} \tag{10}$$

That is, the confidence interval for τ will be unbounded if and only if we cannot reject that the genetic association with the exposure is zero.

Sensitivity Analysis	Interpretation
Of the genetic association with the **exposure**	The sensitivity of the genetic association with the exposure reveals the *stability* of the MR causal effect estimate. Biases strong enough to result in a failure of rejection that the genetic association with the exposure is zero, also lead to unbounded confidence intervals for the MR causal effect estimate.
Of the genetic association with the **outcome**	The sensitivity of the genetic association with the outcome is equivalent to the sensitivity of the MR causal effect estimate with respect to the *zero* null hypothesis. Biases strong enough to result in a failure of rejection that the genetic association with the outcome is zero equally result in a failure to reject the null hypothesis that the MR causal effect estimate is zero.

Table 4. The sensitivity of the MR causal effect estimate can be decomposed into the sensitivity of its two components: the sensitivity of the genetic association with the exposure and the sensitivity of the genetic association with the outcome.

We now turn our attention to the null hypothesis of zero effect, that is, $H_0 : \tau = 0$. Notice in this case the first two terms of the quadratic equation, a and b, vanish. What we have left is only the term c which will be negative if, and only if,

$$\hat{\beta}^2_{YZ|\boldsymbol{XW}} - \widehat{\text{var}}(\hat{\beta}_{YZ|\boldsymbol{XW}}) \times t^{*2}_{\alpha,\text{df}} \leq 0 \implies \frac{\hat{\beta}_{YZ|\boldsymbol{XW}}}{\widehat{\text{se}}(\hat{\beta}_{YZ|\boldsymbol{XW}})} = t_{\hat{\beta}_{YZ|\boldsymbol{XW}}} \leq t^{*}_{\alpha,\text{df}} \tag{11}$$

In other words, the null hypothesis of zero effect for the causal effect is not rejected if, and only if, the null hypothesis of zero association between the instrument Z with the outcome Y is also not rejected.

We have thus simplified the sensitivity analysis of the MR estimate to the sensitivity analysis of the two genetic associations. If $\boldsymbol{W}$ is strong enough to explain away the genetic association with the exposure, then $\boldsymbol{W}$ is strong enough to make the the causal effect arbitrarily large in either direction. If $\boldsymbol{W}$ is strong enough to explain away the genetic association with the outcome trait, than $\boldsymbol{W}$ is strong enough to explain away the MR estimate. This is summarized in Table 4.

Since we have reduced the problem of sensitivity analysis of MR to the problem of sensitivity analysis of the genetic associations, we can leverage all tools of Cinelli and Hazlett[34] for our problem. Here we thus review the main sensitivity analysis results of Cinelli and Hazlett, in the context of the genetic association with the outcome. All results below, of course, also apply to the genetic association with the treatment, by just replacing Y with D where appropriate.

Sensitivity formulas for the genetic associations. Consider first a univariate W and let $R^2_{Z\sim W|\mathbf{X}}$ denote the partial R^2 of W with the genetic instrument and let $R^2_{Y\sim W|Z,\mathbf{X}}$ denote the partial R^2 of W with the outcome trait. Given the observed genetic association $\widehat{\beta}_{YZ|\boldsymbol{X}}$ and its estimated standard error $\widehat{\text{se}}(\widehat{\beta}_{YZ|\boldsymbol{X}})$, adjusting for $\boldsymbol{X}$ alone, the estimate and standard error we would have obtained further adjusting for W can be recovered with[34] ,

$$\widehat{\beta}_{YZ|\boldsymbol{X}W} = \widehat{\beta}_{YZ|\boldsymbol{X}} \pm \widehat{\text{se}}(\widehat{\beta}_{YZ|\boldsymbol{X}}) \sqrt{\frac{R^2_{Y\sim W|Z,\mathbf{X}}\, R^2_{Z\sim W|\mathbf{X}}}{1 - R^2_{Z\sim W|\mathbf{X}}}(\text{df})} \tag{12}$$

and,

$$\widehat{\text{se}}(\widehat{\beta}_{YZ|\boldsymbol{X}W}) = \widehat{\text{se}}(\widehat{\beta}_{YZ|\boldsymbol{X}}) \sqrt{\frac{1 - R^2_{Y\sim W|Z,\mathbf{X}}}{1 - R^2_{Z\sim W|\mathbf{X}}} \left(\frac{\text{df}}{\text{df} - 1}\right)} \tag{13}$$

Where here now df denote the degrees of freedom of the AR regression actually run. These formulas allow us to investigate how the estimate, standard error, t-values, p-values or confidence intervals would have changed, under a confounder W of any postulated strength. For a singleton W these formulas are exact, and for multivariate $\boldsymbol{W}$, it can further be shown that these formulas are conservative, barring an adjustment on the degrees of freedom[34] (that is, these are the worse biases a multivariate $\boldsymbol{W}$ could cause). These formulas form the basis of the contour plots shown in Figure 2.

Bounds on the partial R^2 of W based on observed covariates. Where investigators are unable to make direct claims on the strength of W, it may be helpful to consider relative claims, by positing, for instance, that W is no stronger than some observed covariate X_j. For that, consider a confounder orthogonal to the observed covariates, ie., $W \perp \mathbf{X}$ and define

$$k_Z := \frac{R^2_{Z\sim W|\mathbf{X}_{-j}}}{R^2_{Z\sim X_j|\mathbf{X}_{-j}}}, \qquad k_Y := \frac{R^2_{Y\sim W|\mathbf{X}_{-j},Z}}{R^2_{Y\sim X_j|\mathbf{X}_{-j},Z}}. \tag{14}$$

where $\mathbf{X}_{-j}$ represents the vector of covariates $\mathbf{X}$ excluding X_j. Then the strength of W can be bounded by[34],

$$R^2_{Z\sim W|\mathbf{X}} = k_Z f^2_{Z\sim X_j|\mathbf{X}_{-j}}, \quad R^2_{Y\sim W|Z,\mathbf{X}} \leq \eta^2 f^2_{Y\sim X_j|\mathbf{X}_{-j},Z} \tag{15}$$

where η is a scalar which depends on k_Y, k_Z and $R^2_{Z\sim X_j|\mathbf{X}_{-j}}$.

Sensitivity statistics for routine reporting. The previous results allow us to perform sensitivity analysis to confounding of any postulated strength. However, widespread adoption of sensitivity analysis benefits from simple metrics that users can report to quickly summarize the robustness of their results. With that in mind, Cinelli and Hazlett[34] introduced two sensitivity statistics for routine reporting: the Robustness Value (RV) and the partial R^2.

Let $f := |f_{Y\sim Z|\mathbf{X}}|$ denote the absolute value of the partial *Cohen's f* of the genetic instrument with the outcome.[1] Now also re-scale the critical threshold, $f^*_\alpha := |t^*_{\alpha,\mathrm{df}-1}|/\sqrt{\mathrm{df}-1}$, and define $f_\alpha := f - f^*_\alpha$. The robustness value RV_α is defined as the minimal strength of association that W must have, both with the genetic instrument Z and the outcome trait Y, in order to make the genetic association with the outcome statistically insignificant. This is given by[34]

$$\mathrm{RV}_\alpha = \begin{cases} 0, & \text{if } f_\alpha < 0 \\ \dfrac{1}{2}\left(\sqrt{f_\alpha^4 + 4f_\alpha^2} - f_\alpha^2\right), & \text{if } f^*_\alpha \le f < f^{*-1}_\alpha \\ \dfrac{f^2 - f^{*2}_\alpha}{1+f^2}, & \text{otherwise.} \end{cases} \tag{16}$$

Any W with both strength of associations below RV_α is not sufficiently strong to make the genetic association with the outcome statistically insignificant, and, thus, also not sufficiently strong to make the MR causal effect estimate statistically insignificant. On the other hand, any W with both strength of associations above RV_α is sufficiently strong to do so.

Moving to the partial R^2, in addition to quantifying how much variation of the outcome trait is explained by the genetic instrument, the partial R^2 also tells us how robust the genetic association with the outcome is to an "extreme sensitivity scenario." More precisely, suppose that the unobserved variable W explained *all* residual variance of the outcome trait. Then, for W to bring the genetic association to zero, it must explain *at least* as much residual variation of the genetic instrument as the residual variation of the outcome trait that the genetic instrument currently explains[34]. Mathematically, if $R_{Y\sim W|Z,\mathbf{X}} = 1$, then for W to make $\hat{\beta}_{YZ|\boldsymbol{X}W} = 0$, we need to have that $R^2_{Z\sim W|\mathbf{X}} \ge R^2_{Y\sim Z|\mathbf{X}}$.

[1] The partial Cohen's f^2 can be written as $f^2_{Y\sim Z|\mathbf{X}} = R^2_{Y\sim Z|\mathbf{X}}/(1-R^2_{Y\sim Z|\mathbf{X}})$.