Detecting and correcting for bias in Mendelian randomization analyses using gene-by-environment interactions

Wes Spiller[1*], David Slichter[2], Jack Bowden[1a] and George Davey Smith[1a]

[a]equal supervisory contribution

[1] School of Social and Community Medicine, University of Bristol, Barley House, Oakfield Grove, Bristol, BS8 2BN, U.K [2] Department of Economics, Binghamton University, State University of New York, 4400 Vestal Parkway East, Binghamton, NY 13902, USA

Abstract

Background: Mendelian randomization has developed into an established method for strengthening causal inference and estimating causal effects, largely as a consequence of the proliferation of genome-wide association studies. However, genetic instruments remain controversial as pleiotropic effects can introduce bias into causal estimates. Recent work has highlighted the potential of gene-environment interactions in correcting for pleiotropic bias in Mendelian randomization analyses.

Methods: We introduce linear Slichter regression (LSR) as a framework capable of identifying and correcting for pleiotropic bias, drawing upon developments in econometrics and epidemiology. If an instrument-covariate interaction induces variation in the association between a genetic instrument and exposure, it is possible to identify and correct for pleiotropic effects. The interpretation of LSR is similar to conventional summary Mendelian randomization approaches. A particular advantage of LSR is the ability to assess pleiotropic effects using individual genetic variants.

Results: We investigate the effect of BMI upon systolic blood pressure (SBP) using data from the UK Biobank and the GIANT consortium using a single instrument (a weighted allelic score), finding evidence of a positive association between BMI and SBP in agreement with two sample summary Mendelian randomization approaches. We assess the performance of LSR with respect to identifying and correcting for horizontal pleiotropy in a simulation setting, highlighting the utility of the approach where the LSR assumptions are violated.

Conclusions: By utilising instrument-covariate interactions within a linear regression framework, it is possible to identify and correct for pleiotropic bias, provided the average magnitude of pleiotropy is constant across interaction covariate subgroups.

Key words: Mendelian randomization, invalid instruments, pleiotropy, Slichter regression, gene-environment interaction.

Key Messages

- Instrument-covariate interactions can be used to identify pleiotropic bias in Mendelian randomization analyses, provided they induce sufficient variation in the association between the genetic instrument and exposure.
- The interpretation of LSR is similar to that of summary MR methods such as MR-Egger regression.
- By regressing the gene-outcome association upon the gene-exposure association across interaction covariate subgroups, it is possible to obtain an estimate of the average pleiotropic effect and a causal effect estimate.
- The approach serves as a valuable test for directional pleiotropy, and can be used to inform instrument selection.

Introduction

Mendelian randomization (MR) has developed into a popular multifaceted approach to assessing causal relationships in epidemiology[1, 2]. In many cases, MR analyses involve employing genetic variants as instrumental variables (IVs) allowing for causal effect estimates to be consistently estimated the presence of unmeasured confounding. This requires candidate variants to be associated with the exposure of interest (IV1), not be associated with confounders of the exposure and outcome (IV2), and not be associated with the outcome through pathways outside of the exposure (IV3)[3]. The extent to which specific genetic variants satisfy these assumptions is often controversial, however, due to uncertainties around the true mechanisms responsible for observed gene-phenotype relationships[4].

One issue of particular concern is potential violation of IV3 through horizontal pleiotropy-occurring when a genetic instrument is associated with a study outcome through biological pathways outside the exposure of interest[5]. This introduces bias into causal effect estimates in the direction of pleiotropic association, and can inflate type I error rates when testing causal null hypotheses[5, 6]. When multiple instruments are available, one potential solution is to adopt a meta-analytic approach[7]. If the set of genetic variants does not exhibit an average non-zero (or 'directional') pleiotropic effect, an inverse variance weighted (IVW) estimate can be used to obtain an effect estimate equivalent to that of two-stage least squares (TSLS) regression[5, 7]. In cases where directional pleiotropy is suspected, MR-Egger, median, and mode-based methods can be used to estimate the magnitude of pleiotropic effects, and provide a corrected causal effect estimate[5, 8, 9]. Such methods are most applicable in two-sample summary MR, particularly MR-Egger regression[10, 11]. In an individual level data setting it is common practice to combine genetic variants into allelic scores to create a powerful instrument, mitigating the effects of both weak instrument and directional pleiotropic bias[12, 13].

In the econometrics literature, Slichter regression has been proposed as a method for evaluating instrument validity within a potential outcomes framework[14, 15]. This involves finding population subgroups where the instrument and exposure are independent, and identifying the association between the instrument and outcome which would arise for such subgroups. For these subgroups, an observed instrument-outcome association serves as evidence of a violation of IV3. Slichter regression builds upon a number of key developments in econometrics, in particular the identification and estimation of local average treatment effects put forward by Imbens and Angrist[15]. The use of baseline covariate interactions in IV analyses has also received some attention, such as in the work of Card[16] estimating returns to schooling utilising an observed interaction between college proximity and IQ. Conley et al[17] emphasise the potential trade-off between instrument strength and degree of IV3 violation in putting forward the notion of plausibly endogeneity, whilst further works underlining the utility of using instrument-covariate interactions have also emerged, such as those of Gennetian et al and Small[18, 19].

In this paper, we introduce Slichter regression within the context of epidemiology, and present linear Slichter regression (LSR) as a statistical framework to identify and correct for pleiotropic bias in MR studies using gene-covariate interactions. The structure of LSR is similar to LD score regression[20], conducted within a linear regression framework, and can be viewed as an analogous approach to MR-Egger regression using only a single genetic instrument. It can be applied in both individual or summary data settings and builds upon previous work utilising instrument-covariate interactions in assessments of pleiotropy and in testing MR assumptions more generally[21-27].

We begin by outlining the LSR framework, highlighting the assumptions and implementation of the approach. With this complete, an applied example is considered examining the effect of body mass index (BMI) upon systolic blood pressure (SBP) using the most recent release

of data from UK Biobank (July 2017) and the GIANT consortium[28]. Initially, a two-sample summary MR analysis is conducted as a frame of reference using derived summary statistics from each sample, after which LSR is implemented using a single allelic score. Using LSR we find evidence suggesting a positive association between BMI and SBP in agreement with two-sample summary MR approaches. The similarity between LSR and MR-Egger pleiotropy estimates is of particular importance, as it highlights the extent to which pleiotropic effect estimates comparable to MR-Egger regression can be estimated using a single instrument. Finally, we conduct a simulation study highlighting the effectiveness of the approach under varying conditions.

Methods

Non-technical intuition

Consider a situation in which the instrument-exposure association is found to vary between subgroups of the target population. We follow Slichter[14] in defining an observed subgroup for which the instrument does not predict the exposure of interest as a *no relevance point*. As a valid genetic instrument can only be associated with the outcome of interest through the exposure, it follows that a valid instrument would also not be associated with the outcome at an observed no relevance point. Any non-zero instrument-outcome association at the no relevance point can therefore be interpreted as evidence of horizontal pleiotropy.

This intuitive approach to pleiotropy assessment has been considered in a number of epidemiological studies, and in particular the approaches of Chen et al[26] and Cho et al[21]. In the work of Chen et al[26], differences in drinking behaviour by gender in East Asian populations were explored through a fixed effects meta-analysis of the *ALDH2* genetic variant and blood pressure. Observing that males are much more likely to consume alcohol than females, gender-stratified drinking behaviour was used to identify female participants as

a no relevance point. This interaction has received further attention in papers such as Cho et al[21] and Taylor et al[29]. Emdin et al[22] used differences in instrument-exposure association by gender to identify males as a no relevance point in examining the relationship between genetically predicted waist-to-hip ratio (WHR), type 2 diabetes, and coronary heart disease (CHD). An association between genetically elevated WHR and CHD for women, and not men, served as evidence against horizontal pleiotropy[22].

Examples of variation in instrument-exposure association across populations extends beyond simple gender differences. Another important example is Tyrrell et al[23], investigating the extent to which genetically predicted BMI is associated with environmental factors through gene-covariate interactions. They identified genetically predicted BMI as a weaker instrument for participants experiencing lower levels of socio-economic deprivation (as quantified by the Townsend deprivation index), and utilised negative controls to examine residual confounding[23]. A further interesting example is stratifying by smoking status, as considered in Freathy et al[30] in their examination of the relationship between genetic instruments used to predict smoking status and adiposity. In their recent work, Robinson et al[31] identify genotype-covariate interactions with respect to the heritability of adult BMI, finding evidence of genotype-age and genotype-smoking interactions. Gene-environment interactions with covariates such as socio-economic status were not identified as having a substantial impact on the distribution of phenotypic effects, though this may be the result of a lack of statistical power or measurement error in self-reported covariates.

A closely related approach to detecting and correcting for pleiotropy has recently been proposed by van Kippersluis and Rietveld[32] on pleiotropy-robust Mendelian randomization (PRMR). Under this framework, in cases where a no relevance point is observed, the degree of association between the instrument and the outcome is equated with the exact pleiotropic effect across the whole population. In this respect the approach is similar to that of Chen et

al[26]. This term is then incorporated as an offset within a standard analysis. Whilst their approach is useful in highlighting the potential of no relevance points in assessing pleiotropy, it can be criticised for ignoring true uncertainty in the pleiotropic effect estimate. Its practical application is also limited by the fact that strict instrument-exposure independence is rare. For example, the authors cite Cho et al's[21] MR analysis using gender specific alcohol consumption as a canonical example[32], but in fact 25% of female participants in this study did consume alcohol[21]. Such a violation would obviously undermine an approach that assumed a strict no relevance point. This serves as motivation for the development of a formal statistical model (LSR) to use variation in gene-exposure associations across a covariate to infer the likely location of a no relevance point whilst properly accounting for its uncertainty, and use this as a basis for detecting and adjusting for pleiotropy.

In presenting LSR we draw attention to similarities with a previous analysis conducted by Cho et al[21]. In this case, a statistical model incorporating a gender-ALDH2 interaction term was fitted to the data assessing the association between alcohol consumption and cardio-vascular risk factors accounting for pleiotropy[21]. The LSR framework shares many common elements with the approach of Cho et al[21], and serves to clarify how it works when individual level data are available. Crucially, however, LSR extends this method so that it can be additionally applied to summary data, thus extending its reach to two sample summary data MR, and also general meta-analysis contexts.

The LSR Framework

Consider an MR study consisting of N participants (indexed by $i = 1, \ldots, N$). For each participant, we record observations of a genetic instrument G_i, an exposure X_i, an outcome Y_i, and a further covariate Z_i which induces variation the association between G_i and X_i through an interaction. This interaction is labelled I_{GZ}. The relationship between each variable is

illustrated in Figure 1, with U representing a set of all unmeasured variables confounding X and Y.

(Figure 1 here)

The exposure X is considered a linear function of G, Z, I_{GZ}, U and an independent error ϵ_X, whilst the outcome Y is a linear function of G, Z, X, U and an independent error ϵ_Y. Using γ and β to denote regression coefficients for the first and second stage models respectively, a two-stage model can be defined as:

$$X_i = \gamma_0 + \gamma_1 G_i + \gamma_2 Z_i + \gamma_3 G_i Z_i + U_i + \epsilon_{Xi} \quad (1)$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 Z_i + U_i + \epsilon_{Yi} \quad (2)$$

The causal effect of X on Y is denoted β_1 and is the parameter we wish to estimate. The pleiotropic effect of the instrument across the sample is β_2. Note that performing an ordinary least squares (OLS) regression of Y upon X would yield a biased estimate of β_1 due to confounding, and two-stage least squares (TSLS) regression of Y on the genetically predicted exposure $\hat{X} = E[X|G]$ would result in biased estimates in cases where $\beta_2 \neq 0$. This can be shown by simulating a two stage IV model with a set of confounding variables U, such that the errors from the first and second stage models are correlated (e.g. with correlation coefficient $\rho = 0.5$), and varying the degree of horizontal pleiotropy via β_2. In this example, the true effect of X upon Y is defined as 1 with results from each method presented in Table 1.

(Table 1 here)

From Table 1 we can see that TSLS provides a more accurate estimate in cases where there is no observed directional pleiotropic effect ($\beta_2 = 0$), but can be substantially biased

when $\beta_2 \neq 0$. Any remaining bias when $\beta_2 = 0$ can be attributed to weak instrument bias or sampling error.

No relevance points can be estimated using model (1) by estimating a value of the covariate $Z = z_X$ at which G and X are independent. This is achieved by calculating the partial effect of G upon X and rearranging such that:

$$\frac{dX}{dG} = \gamma_1 + \gamma_3 Z_i = 0 \tag{3}$$

In our model, this yields the trivial solution

$$z_X = -\left(\frac{\gamma_1}{\gamma_3}\right) \tag{4}$$

In cases where the covariate value z_X is actually observed in the population, regressing Y upon G for the subset of participants with $Z = z_X$ will provide an estimate of horizontal pleiotropy (that is for β_2) as the coefficient of G. Unfortunately, this approach is difficult to implement in practice, either because the value z_X is not observed in the population or the subset of participants is simply too small to provide sufficient power. This provides the motivation for use of LSR in estimating the degree of pleiotropy at a theoretical (or extrapolated) no-relevance point, using differences in instrument-exposure associations across values of Z.

Linear Slichter regression

We begin by constructing the *reduced form* instrumental variable model, that is, models for X given G, and Y given G by re-writing model (1) as

$$X_i = \gamma_0 + (\gamma_1 + \gamma_3 Z_i)G_i + \gamma_2 Z_i + U_i + \epsilon_{Xi} \tag{5}$$

and model (2) as

$$Y_i = \beta_0 + [\beta_1(\gamma_1 + \gamma_3 Z_i) + \beta_2 + \beta_4 Z_i + U_i + \epsilon_{Xi}]G_i + \beta_3 Z_i + U_i + \epsilon_{Yi} \tag{6}$$

Note that we have now also included a possible interaction between the genetic instrument and covariate Z, $\beta_4 G Z_i$, in model (6), which allows the pleiotropic effect to vary across values of Z. To help explain the assumptions that LSR relies upon, we will initially set β_4 to zero so that no such variation can occur. The change in $G - X$ and $G - Y$ associations for a given change in Z can be identified as the coefficient of G in models (5) and (6) respectively (with β_4 set to 0) as

$$G - X \text{ association: } (\gamma_1 + \gamma_3 Z_i)$$

$$G - Y \text{ association: } [\beta_1(\gamma_1 + \gamma_3 Z_i) + \beta_2]$$

The Wald ratio[33] estimand for the causal effect of X on Y (i.e. the true $G - Y$ association divided by the true $G - X$ association) would then be equal to:

$$\frac{\beta_1(\gamma_1+\gamma_3 Z_i)+\beta_2}{\gamma_1+\gamma_3 Z_i} = \beta_1 + \frac{\beta_2}{\gamma_1+\gamma_3 Z_i} \tag{7}$$

That is, the causal effect, β_1, plus a non-zero bias term whenever β_2 is non-zero. In the Cho et al[21] analysis, an estimate for β_1 was obtained by performing TSLS regression using the interaction as the instrument, by fitting models (8) and (9) below:

$$X_i = (\gamma_1 + \gamma_3 Z_i) G_i + \gamma_2 Z_i + \epsilon_{Xi} \tag{8}$$

$$Y_i = \beta_0 + \beta_1 \widehat{X}_i + \beta_2 G_i + \beta_3 Z_i + \epsilon_{Yi}, \tag{9}$$

Where $\widehat{X}_i$ is the fitted value from model (8). In this case, the coefficient β_2 represents the degree of pleiotropy for the genetic instrument G. Cho et al also demonstrate the use of two-stage predictor substitution (TSPS) models of the same structure when considering binary outcome variables[21]. It should be noted that the interaction is assumed to be a valid instrument, and therefore that there is no observed association between the interaction and the outcome in the second stage model($\beta_4 = 0$).

Whilst this approach is useful in not requiring an observed no relevance point, it has two limitations. First, as a consequence of utilising TSLS and TSPS, it is only applicable to cases in which individual level data is available. In utilising genetic data, it has become common to utilise summary data within a meta-analysis context, as individual studies often lack statistical power due to sample size restrictions. A second limitation is that TSLS assumes an underlying linear model, and this may not be the case. For example, in considering adiposity as an exposure, individuals at extreme values could be at greater risk, implying a curved relationship. Care is therefore needed in justifying the assumption of an underlying linear model.

LSR attempts to overcome these limitations by reframing the model within a two-sample summary MR context, delivering a consistent estimate for β_1 by executing the following three step procedure:

1. Estimate $G - X$ and $G - Y$ associations at a range of values of Z.
2. Regress the $G - Y$ associations on the $G - X$ associations within a linear regression.
3. Estimate the causal effect β_1 as the slope of the regression.

Let Z_j denote the jth subgroup of Z ($j = 1, \dots, J$). For each group Z_j, we initially estimate the instrument-exposure association and standard error (step 1) by fitting the following regression model:

$$X_i = \gamma_{j0} + \gamma_{j1} G_i + \epsilon_{jXi} \tag{10}$$

Note that we include a subscript j to distinguish the regression parameters from the first stage model (1). The coefficient γ_{j1} is therefore interpreted as the $G - X$ association for group Z_j. Next, we fit the corresponding instrument-outcome regression model (step 2):

$$Y_i = \delta_{0j} + \delta_{j1} G_i + \epsilon_{jYi} \tag{11}$$

In this case, we use δ_{j1} to denote $G - Y$ association coefficient for group Z_j, distinguishing the model from (6). Thus, from models (10) and (11) we obtain sets of $G - X$ associations (γ_{J1}) and $G - Y$ associations (δ_{J1}) respectively across Z_J subgroups. Finally, we regress the set of $\widehat{\delta_{j1}}$ estimates upon the set of $\widehat{\gamma_{j1}}$ estimates (step 3):

$$\widehat{\delta_{j1}} = \beta_{LSR_0} + \beta_{LSR_1}\widehat{\gamma_{j1}} + \epsilon_{JLSR} \tag{12}$$

In Model (12), β_{LSR_0} is an estimate of directional pleiotropy (β_2), whilst β_{LSR_1} is the causal effect of X upon Y corrected for any directional pleiotropy (β_1). To understand how this is the case, recall that β_2 represents a constant pleiotropic effect across subgroups of Z. Model (12) can be thought of as an average of the ratio estimates across Z_J, with the bias from β_2 estimated as the intercept. This is illustrated in Figure 2.

(Figure 2 here)

To show how the intercept estimates β_2, consider the reduced form model (6) evaluated at the no relevance point z_X. From equation (4), $z_X = -\left(\frac{\gamma_1}{\gamma_3}\right)$. Then, by substitution:

$$\beta_1\gamma_1 + \beta_1\gamma_3\left(\frac{-\gamma_1}{\gamma_3}\right) + \beta_2 = \beta_1\gamma_1 - \beta_1\gamma_1 + \beta_2 = \beta_2 \tag{13}$$

In cases where the intercept estimate passes exactly through the origin, the LSR causal effect estimate would be identical to the inverse-variance weighted (IVW) estimate. This mirrors the equivalence of IVW and MR-Egger regression in the multiple instrument setting when the estimated average pleiotropic effect across all variants is equal to 0. LSR can therefore be viewed as directly analogous to MR-Egger regression: in LSR we simply replace $G - X$ association estimates for multiple variants with subgroup-specific associations for a single variant. R code for implementing LSR is provided in the web appendix.

In implementing LSR, it is important to examine several important features of the analysis. Firstly, it is important to *not* transform effects to be positive using LSR as is the case for MR-Egger regression, as this mischaracterises the interaction term, attenuating causal effect estimates. A simulated example illustrating this issue is presented in the web appendix.

A second consideration is that in cases where instrument-exposure associations are present for all groups in the same direction, the accuracy in extrapolating the regression line towards a theoretical no-relevance point will be a function of the distance from the minimum Z_j instrument-exposure association, and variation in the set of Z_j instrument-exposure associations. This feature of LSR is examined further in the web appendix.

Finally, one benefit of LSR is that the suitability of a linear model can to some extent be examined by visually inspecting the distribution of instrument-exposure and instrument-outcome associations across covariate subgroups of Z. In cases where groups are in ascending or descending order of instrument-exposure associations, and a constant change in the instrument-outcome is observed across the subgroups, a linear model can be argued to be appropriate.

Assumptions of LSR

As LSR relies upon fitting regression estimates within a simple regression model, an initial assumption is that the variance of the instrument-exposure association is negligible, which is referred to as NO Measurement Error (NOME)[5, 34]. In the standard two-sample summary data MR context (with multiple variants but no assumed variant-covariate interactions) violation of NOME, means that causal effect estimates will be attenuated towards zero as a consequence of regression dilution bias. Recent work has highlighted the role of the I^2_{GX} statistic as a means of assessing the degree of attenuation in MR-Egger estimates from NOME violation in this context[34], though at present it is unclear how I^2_{GX} relates to LSR.

As a constant, β_2 contributes to the intercept, and consistent estimates for both β_1 and β_2 are produced in cases where $\beta_4 = 0$. We refer to this as the *CoPE* (Constant Pleiotropic Effect) assumption. That is, pleiotropic effects must remain constant across all values of Z. If $\beta_4 \neq 0$ then CoPE is violated, which in our model would lead to the true pleiotropic effect β_2 being incorrectly equated to $\beta_2 - \frac{\beta_4\gamma_1}{\gamma_3}$ instead. This would, in turn, lead to bias in the causal estimate for β_1 such that:

$$\hat{\beta}_1 = \beta_1 + \frac{\beta_4}{\gamma_3} \tag{14}$$

The derivation of this result is provided in the web appendix. From equation (14) it is clearly possible to mitigate the effect of bias due to CoPE violation when the variation in instrument-exposure association across Z (γ_3) is sufficiently large relative to the variation in pleiotropic effect β_4, as the bias will tend towards zero as γ_3 increases. However, as it is not possible to directly estimate β_4 justifying the relative effect sizes of the first and second stage interactions would likely rely upon a priori knowledge.

The CoPE assumption articulates a fundamental fact about LSR, that we now highlight:

The validity of LSR hinges on treating the interaction as a valid IV.

This places three restrictions on the interaction analogous to the conventional relevance restriction (IV1), exogeneity restriction (IV2) and exclusion restriction (IV3), as illustrated in Figure 3 below.

(Figure 3 here)

First, there must be a non-zero $G - Z$ interaction on X (i.e. $\gamma_3 \neq 0$) . Second, the strength of the interaction should not be associated with a confounder on the $X - Y$ pathway. Third, there should not be a direct association between the interaction and the outcome within the

second stage model. A violation of either the second or third restrictions would result in a non-zero value of β_4, violating the COPE assumptions and consequently introducing bias into causal effect estimates.

The relationships between the interaction and confounders further underscores a key feature of the LSR framework:

Associations between either G or Z and confounders on the $X - Y$ pathway, or direct effects of G or Z on the outcome Y do not invalidate LSR.

To illustrate this, consider equation (15) in which a confounder on the $X - Y$ pathway, denoted U_1, is a linear function of G and Z.

$$U_{1i} = \pi_0 + \pi_1 Z_i + \pi_2 G_i + \pi_3 G_i Z_i + \epsilon_i \tag{15}$$

In equation (15) π is used to denote regression coefficients. Non-zero values of π_1, and π_2 would not induce bias in LSR estimates, although they would result in biased estimates using conventional approaches such as TSLS regression with the instrument G. In this example, the second restriction requires the association π_3 to be zero. This holds for a set of confounding variables with differing configurations of π_1 and π_2 values, provided $\pi_3 = 0$ across the set of confounders. Direct effects of either G or Z with the outcome Y do not invalidate the LSR method, because they can be explicitly modelled in equation (2).

With these considerations in mind, it is important to acknowledge that violations of these restrictions serve as the driving force behind changes in pleiotropic effects across interaction-covariate subgroups, resulting in values of $\beta_4 \neq 0$.

LSR as a sensitivity analysis

In cases where the CoPE assumption is assumed to be violated, LSR can still be used in sensitivity analyses as a means to select a subset of valid instruments. To show how this is the

case, we begin by clarifying that an invalid instrument can be detected in principle whenever $\beta_2 - \frac{\beta_4\gamma_1}{\gamma_3} \neq 0$, due to either $\beta_2 \neq 0$, $\beta_4 \neq 0$, or both. As a consequence, LSR can be used to assess the validity of individual instruments, informing instrument selection and components of allelic scores. There are, however, two important considerations when applying this approach. First, it is not possible to distinguish the average pleiotropic effect across interaction-covariate subgroups from the change in pleiotropic effect between instrument-covariate subgroups. It is therefore a test of invalidity occurring either due to an average non-zero pleiotropic effect across interaction-covariate subgroups, or due to changing pleiotropic effects between interaction-covariate subgroups, and cannot be used to correct LSR estimates directly.

Second, LSR will incorrectly fail to detect invalid instruments in cases where:

$$\beta_2 \neq 0,\ \beta_4 \neq 0,\ \beta_2 = \left(-\frac{\beta_4\gamma_1}{\gamma_3}\right) \tag{16}$$

Causal effect of BMI upon SBP

There exists an extensive literature on the relationship between adiposity and SBP, with both observational[35] and MR[36-38] studies finding evidence of positive association. However, the magnitude of this association has been found to differ markedly between such studies, with observational studies often recording greater effect sizes than those using MR.

As an applied example, we perform two sample summary MR and LSR analyses examining the effect of adiposity (measured using BMI) upon SBP using data from the GIANT consortium[28] and UK Biobank. The motivation in performing both forms of analysis is to highlight the extent to which pleiotropic effect estimates obtained using LSR and a single instrument agree with conventional MR approaches.

In conducting a two-sample summary analysis, effect estimates and standard errors for 95 genetic variants associated with BMI ($p = 5 \times 10^{-8}$) were obtained from Locke AE[28] using the GIANT consortium sample (with a full list of variants presented in the web appendix). Corresponding estimates for each genetic variant with respect to SBP were obtained using UK Biobank. In contrast, LSR was implemented by constructing a weighted allelic score using estimates from the GIANT consortium. The LSR analysis can be viewed as analogous to two-sample summary data MR, using instrument-exposure estimates for BMI as external weights, and individual data from a separate sample to inform instrument-outcome association estimates. In each analysis BMI, SBP, and the weighted allelic score were standardised.

Analysis I: Two-Sample Summary Analysis

We implement several two-sample summary MR methods utilising the mrrobust software package[39] in Stata SE 14.0[40]. Performing IVW provides an estimate comparable to TSLS, and produces estimates with greater precision than alternative summary approaches. However, as IVW estimates can exhibit bias in the presence of horizontal pleiotropy, MR-Egger regression, weighted median, and weighted modal approaches are also considered as sensitivity analyses.

A range of methods are adopted in sensitivity analyses for two key reasons. First, each method relies upon differing assumptions with respect to the underlying distribution of pleiotropic effects in addition to the NOME assumption. MR-Egger regression requires the effect of genetic variants on the exposure to be independent of their pleiotropic effects on the outcome (InSIDE)[5]. The weighted median requires more than 50% of variants to not exhibit pleiotropic effects (with respect to their relative weighting)[8], whilst the modal estimator

assumes that the most frequent value of the pleiotropic bias across the set of genetic variants is zero (ZEMPA)[9].

Estimates using each method are presented in Table 2, with an accompanying plot showing the IVW and MR-Egger estimates in Figure 4.

(Table 2 here)

(Figure 4 here)

With the exception of MR-Egger regression, each of the methods performed above show evidence of a positive association between BMI and SBP. There does not appear to be substantial horizontal pleiotropic effect using either MR-Egger, weighted median, or weighted modal approaches, with the IVW estimate lying within the confidence intervals of both the weighted median and weighted modal estimates. Disagreement in effect estimation between MR-Egger regression and the other methods can in part be explained by regression dilution bias, as evidenced by an I^2_{GX} value of 87.8 (indicative of a relative bias of 12.2% towards the null), as well as by identifying influential outliers. Such outliers can be identified by calculating the degree of heterogeneity utilising Rucker's Q statistic, and estimating the relative contribution of each genetic variant to overall heterogeneity, as outlined in Bowden et al[41]. In doing so, the overall Rucker's Q statistic is 1069.42 ($p =< 0.001$), with rs11191560 contributing almost a quarter of the overall heterogeneity ($Q = 274.4,\ p =< 0.001$).

Analysis II: LSR using Townsend Deprivation Index

In implementing LSR, Townsend Deprivation Index (TDI) was selected as a continuous covariate for which instrument strength was expected to vary, based on findings from

previous studies[23, 42]. TDI is a common derived measure of socio-economic deprivation, using many variables such as car ownership, occupation type and educational attainment[43].

In the UK Biobank, TDI scores were obtained from preceding national census data and calculated for electoral districts ("*wards*" comprised of approximately 5,500 individuals). Participants were assigned a score based upon the area in which they lived, determined using the postcode of their home dwelling. The selection of TDI was based upon previous evidence suggesting genetically determined BMI to be a weaker predictor of BMI for individuals experiencing lower levels of social deprivation[23]. Missing values were considered to be missing completely at random (MCAR), and were removed prior to performing the analysis.

We present observational and TSLS estimates using the weighted allelic score as an instrument and controlling for TDI in Table 3. In both cases, we find evidence of a positive association between BMI and SBP, with a greater magnitude of effect for the observational estimate.

(Table 3 here)

In this case, the estimates agree with findings of the previous studies discussed above. The instrument was also considered to be sufficiently strong to overcome weak instrument bias, with an F statistic of 5112. To perform LSR, we divided the sample on the basis of TDI score into 2, 5, 10, 20, and 50 population subgroups. In each case, a ratio estimate was calculated for each group, after which IVW and LSR estimates were produced. The results of each analysis are presented in Table 4, with IVW referring to an inverse-variance weighted estimate using interaction covariate subgroups.

(Table 4 here)

From Table 4 we see that the IVW estimates are directionally consistent and of a similar magnitude to the TSLS estimates as expected. In each case, there again appears to be limited

evidence of pleiotropy, whilst there appears to be some indication of a positive effect of BMI upon SBP, particularly in the 5-group case. Figure 5 displays both the IVW and LSR estimates for the 5-group case, whilst corresponding plots for other groups are presented in the web appendix.

(Figure 5 here)

Considering Figure 5, a number of key features of the analysis can be identified. Initially, the ordering of the TDI groups supports the assumption that the instrument-exposure association varies across levels of TDI. In particular, the least deprived groups (group 1 and group 2) have the weakest association, suggesting that genetically predicted BMI is a weaker predictor of BMI for participants experiencing lower levels of deprivation. A further observation is that the positioning of each estimate provides some evidence of a linear relationship, suggesting that utilising a linear regression framework is appropriate.

Comparing these estimates to those obtained using two-sample summary MR, there appears to be substantial agreement in the findings of the two approaches. LSR does not detect horizontal pleiotropy, and constraining the LSR model to the intercept yields an effect estimate similar to IVW, weighted median, and weighted modal approaches in the two-sample summary MR analysis. The lack of agreement in detecting a positive effect of BMI upon SBP using conventional thresholds can be attributed to insufficient strength of TDI as in interacting covariate. The observed agreement in the findings between the single instrument and multiple instrument approaches is encouraging, and it would be valuable to repeat LSR using a range of interacting covariates.

One important consideration in performing MR analyses is that causal effect estimates are often uncertain, due to either a lack of precision or doubts regarding the assumptions of the implemented approach. A response to this issue put forward by VanderWeele et al[44], has been

to shift the emphasis from identifying the magnitude of causal effects to identifying the presence of causal effects. Under such a paradigm, estimation using instrument-covariate interactions, such as through LSR, can be particularly insightful in identifying broad effects or associations in epidemiological studies.

Simulations

To illustrate the effectiveness of LSR, and further consider the importance of the CoPE assumption with respect to causal effect estimation, we perform a simulation study within a two-sample MR framework. Considering a realistic case, two sets of simulations are performed, the first using a null causal effect ($\beta_1 = 0$), and the second a positive causal effect ($\beta_1 = 0.05$). Individual level data is generated, from which the necessary summary data estimates are extracted. In each case, a total of 5 population subgroups are considered, with further details provided in the web appendix.

Four distinct cases are considered:

- No pleiotropy and CoPE satisfied
- Directional pleiotropy and CoPE satisfied
- No pleiotropy and CoPE violated
- Directional pleiotropy and CoPE violated

The results for each case represent the mean values for 10 000 simulated datasets.

Results

Results of the simulation analysis are presented in Table 5 and Table 6 representing the null effect and 0.05 causal effect scenarios respectively. The mean F statistic remains the same for each case, with substantial variation in F statistic between interaction covariate groups. This is essential, as the variation in instrument strength can be viewed as variation in instrument

relevance for particular population subgroups. In this case, estimates using IVW and LSR, as well as significance values were taken directly from each regression output without using regression weights, as the variant-outcome associations were found to have the same standard errors.

(Table 5 here)

(Table 6 here)

Initially, cases within the null causal effect scenario are considered. In the valid instrument case, both IVW and LSR provide unbiased causal effect estimates, though the IVW estimate is more accurate. This is similar to comparisons between IVW and MR-Egger regression, supporting use of IVW in cases where directional pleiotropy is absent. Type I error rates remained at approximately 5% for both IVW and LSR, and in testing for directional pleiotropy. In the second case, estimates using IVW are biased in the direction of pleiotropic association, whilst LSR continues to produce unbiased estimates. This bias appears to decrease marginally with sample size increases. As the sample size increases, power to detect directional pleiotropy using LSR increases from 21% to 70%, whilst Type I error rates remain at a nominal 5% level.

The third case represents a situation in which the instrument is not valid, but the degree of pleiotropy changes between population subgroups. In this case, both IVW and LSR produce biased causal effect estimates, though the LSR causal effect estimates exhibit a greater degree of bias than the IVW estimates. This contributes to an increase in Type I error rate relative to IVW, which estimates the causal effect to be smaller in magnitude. In this situation, the LSR test for directional pleiotropy is particularly powerful, rising from 43% to 95% as the sample size increases. This seeming increase in power can be attributed CoPE violation $(\beta_4 \neq 0)$ leading to biased pleiotropy estimates (β_2), in this case overestimating the magnitude of

pleiotropic effects. In the final case, both IVW and LSR produce estimates with similar sizes of bias and precision. A particularly interesting feature of this case is that the LSR test for directional pleiotropy is suggestive of a null pleiotropic effect, remaining at 5%. This represents the circumstances which undermine LSR as a sensitivity analysis. In the positive causal effect scenario, both the LSR and IVW approaches produce estimates exhibiting similar patterns to those in the null causal effect case.

Discussion

In this paper, we have presented a method to identify and correct for pleiotropic bias in MR studies using instrument-covariate interactions. In cases where CoPE is satisfied, individual instruments can be assessed, providing less biased causal effect estimates compared to conventional estimates such as IVW in the presence of directional pleiotropy. Where individual level data are available, and where it is sensible to assume an underlying linear model, the Cho et al[21] approach is appropriate and provides estimates in agreement with LSR.

However, in cases where directional pleiotropy is not present, IVW is a more accurate method and should be preferred. In cases where CoPE is violated, a sensible approach would be to prune invalid variants using the pleiotropy estimates from LSR, and then implement IVW using the set of valid variants. In this sense, LSR can be viewed as a sensitivity analysis in a similar fashion to MR-Egger, which can be applied to a single genetic instrument within the individual level data setting[5, 34].

Comparison with existing methods

LSR represents a synthesis of both Slichter regression[14] and MR-Egger regression. As with Slichter regression, the method can be applied outside of an MR context, provided that an interaction can be identified which induces variation in instrument strength, and conforms with the interaction restrictions previously discussed. In Slichter 2014, an example of

assessing returns to schooling conditioning on IQ levels represents such an application[14, 16]. The similarities with MR-Egger regression are such that, rather than a competing methodology, it is more correctly viewed as a counterpart to the method.

At present, median based approaches do not appear to translate to in the single instrument case. In the summary data setting, such methods rely upon at least 50% of the instruments being valid, or 50% of the instruments being valid with respect to their weighting when implementing weighted median regression. In adapting the approach to LSR, such a method would require at least 50% of the population subgroups to not exhibit pleiotropic bias, as these are analogous to individual instruments in the summary setting. This may prove unlikely using a single invalid instrument, but warrants further study.

Two-Sample Summary LSR

Whilst this paper has focused primarily on the application of LSR to individual level data (albeit by extracting and then meta-analysing summary statistics obtained from it), it clearly applies to cases where subgroup specific summary data on instrument-exposure and instrument-outcome associations are available. An alternative approach would be to meta-analyse summary statistics obtained from many separate studies under the assumption that study-specific estimates relate to a study-specific characteristic. For example, the work of Robinson et al highlights the interaction between age and adult BMI heritability as one potential candidate, given that age is likely to vary naturally across contributing studies.

Limitations of LSR

There are a number of factors which must be considered before implementing LSR. Firstly, the CoPE assumption is essential for causal estimate correction. If there is reason to believe that pleiotropic effects differ between population subgroups, then using the approach will result in misleading causal effect estimates. One useful aspect to this problem, however, is

that provided the first stage interaction is sufficiently strong, bias from changes in pleiotropic effect may be sufficiently small as to be negligible in analyses. This may well be the case in situations such as the Cho et al[21] study, where the difference in instrument effect between gender groups is very strong in comparison to potential variation in pleiotropic effect. As it is not possible to directly measure the change in pleiotropic effect across groups, decisions regarding appropriate instrument-covariate interaction selection require justification.

A second limitation of the approach is that, owing to the limited availability of summary data estimates for particular covariate groups, it may be difficult to implement in a summary data setting. At present researchers may be limited to common groupings such as gender, unless further information is made available upon request. A second complication using the summary LSR approach focuses upon the use of study heterogeneity. In many cases, the degree to which such heterogeneity is present with respect to the instrument-covariate interaction may be insufficient to perform a meaningful analysis. A related concern is that studies exhibiting such heterogeneity may undermine the extent to which homogeneity in remaining effects can be assumed. This can introduce confounding and undermine subsequent inference.

A further consideration pertaining to the majority of methods, including LSR, is the extent to which the study sample is representative of the sample of interest. In cases where the sample is not representative, selection bias can have a substantial impact on resulting estimates. This is illustrated in the difference between the estimated effect of BMI upon SBP using the interim UK Biobank release, in which the inclusion of a disproportionate number of heavy smokers may have resulted in a 3-fold increase in the magnitude of the estimated effect (as given in the web appendix). As a number of previous studies have found evidence suggesting an interaction between genetically predicted BMI and smoking, this could highlight smoking as a confounder which violates the interaction-covariate restrictions outlined in this paper. It

may also be possible that composite phenotypes (such as scores derived from several measures) may have differing contributions to the outcome at differing interaction-covariate levels.

This limitation can somewhat be mitigated by performing multiple iterations of LSR using a set of interaction covariates. Provided that the instrument-covariate interaction of sufficient strength, it would be expected that resulting estimates would be in agreement. In cases where substantial disagreement is observed, such disagreement could be indicative of violation of the CoPE assumption, or characteristics of the underlying confounding structure. The work of Emdin et al[22] and Krishna et al[45] follow this reasoning. Further work will consider the implications of interaction-covariate selection, and role of confounding within the context of LSR.

Conclusion

This paper formalises an intuitive method for assessing pleiotropic bias, which has gained increasing traction in recent years. At present, LSR serves as a valuable test for directional pleiotropy, and can provide causal effect estimates robust to directional pleiotropy in cases where CoPE is satisfied. It is therefore most appropriate for use in studies using individual level data, and in informing instrument selection and allelic score construction.

Supplementary Material

A web appendix containing supplementary materials can be found at:

Funding

This work was supported by the Medical Research Council (MRC) and the University of Bristol fund the MRC Integrative Epidemiology Unit [MC UU 12013/1, MC UU 12013/9].

Jack Bowden is supported by an MRC Methodology Research Fellowship (grant MR/N501906/1)

Acknowledgements

The authors would like to thank Gibran Hemani and James Staley for assistance in data formatting, as well as Yoonsu Cho, Neil Davies, Fernando Hartwig, and Tom Palmer for useful comments and suggestions on this work.

This research has been conducted using the UK Biobank Resource under Application Number 8786.

References

1. Davey Smith, G. and S. Ebrahim, *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* Int J Epidemiol, 2003. **32**(1): p. 1-22.
2. Davey Smith, G. and G. Hemani, *Mendelian randomization: genetic anchors for causal inference in epidemiological studies.* Hum Mol Genet, 2014. **23**(R1): p. R89-98.
3. Burgess, S., D.S. Small, and S.G. Thompson, *A review of instrumental variable estimators for Mendelian randomization.* Statistical Methods in Medical Research, 2015: p. 0962280215597579.
4. Glymour, M.M., E.J. Tchetgen Tchetgen, and J.M. Robins, *Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions.* Am J Epidemiol, 2012. **175**(4): p. 332-9.
5. Bowden, J., G. Davey Smith, and S. Burgess, *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* Int J Epidemiol, 2015. **44**(2): p. 512-25.
6. Burgess, S. and S.G. Thompson, *Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects.* Am J Epidemiol, 2015. **181**(4): p. 251-60.
7. Burgess, S., A. Butterworth, and S.G. Thompson, *Mendelian randomization analysis with multiple genetic variants using summarized data.* Genet Epidemiol, 2013. **37**(7): p. 658-65.
8. Bowden, J., et al., *Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator.* Genet Epidemiol, 2016. **40**(4): p. 304-14.
9. Hartwig, F.P., G. Davey Smith, and J. Bowden, *Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption.* International Journal of Epidemiology, 2017.

10. Hartwig, F.P. and N.M. Davies, *Why internal weights should be avoided (not only) in MR-Egger regression.* Int J Epidemiol, 2016. **45**(5): p. 1676-1678.

11. Bowden, J., S. Burgess, and G. Davey Smith, *Response to Hartwig and Davies.* Int J Epidemiol, 2016. **45**(5): p. 1679-1680.

12. Burgess, S. and S.G. Thompson, *Use of allele scores as instrumental variables for Mendelian randomization.* Int J Epidemiol, 2013. **42**(4): p. 1134-44.

13. Davies, N.M., et al., *The many weak instruments problem and Mendelian randomization.* Stat Med, 2015. **34**(3): p. 454-68.

14. Slichter, D., *Testing Instrument Validity and Identification with Invalid Instruments.* 2014: University of Rochester: Rochester: USA.

15. Imbens, G.W. and J.D. Angrist, *Identification and Estimation of Local Average Treatment Effects.* Econometrica, 1994. **62**(2): p. 467-475.

16. Card, D., *Using Geographic Variation in College Proximity to Estimate the Return to Schooling*. 1993, National Bureau of Economic Research, Inc.

17. Conley, T.G., C.B. Hansen, and P.E. Rossi, *Plausibly Exogenous.* Review of Economics and Statistics, 2012. **94**(1): p. 260-272.

18. Gennetian, L., Bos, J.M, Morris, P.A, *Using Instrumental Variables Analysis to Learn More from Social Policy Experiments.* MDRC Working Papers on Research Methodology, 2002.

19. Small, D.S., *Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables.* Journal of Statistical Research, 2012.

20. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* Nature Genetics, 2015. **47**(3): p. 291-+.

21. Cho, Y., et al., *Alcohol intake and cardiovascular risk factors: A Mendelian randomisation study.* Scientific Reports, 2015. **5**.

22. Emdin, C.A., et al., *Genetic Association of Waist-to-Hip Ratio With Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease.* Jama-Journal of the American Medical Association, 2017. **317**(6): p. 626-634.

23. Tyrrell, J., et al., *Gene-obesogenic environment interactions in the UK Biobank study.* Diabetologia, 2016. **59**: p. S51-S51.

24. Davey Smith, G., *Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health.* Genes and Nutrition, 2011. **6**(1): p. 27-43.

25. Lewis, S.J. and G. Davey Smith, *Alcohol, ALDH2, and esophageal cancer: A meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach.* Cancer Epidemiology Biomarkers & Prevention, 2005. **14**(8): p. 1967-1971.

26. Chen, L., et al., *Alcohol intake and blood pressure: A systematic review implementing a Mendelian Randomization approach.* Plos Medicine, 2008. **5**(3): p. 461-471.

27. Davey Smith, G., *Mendelian Randomization for Strengthening Causal Inference in Observational Studies: Application to Gene x Environment Interactions.* Perspectives on Psychological Science, 2010. **5**(5): p. 527-545.

28. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.

29. Taylor, A.E., et al., *Exploring causal associations of alcohol with cardiovascular and metabolic risk factors in a Chinese population using Mendelian randomization analysis.* Sci Rep, 2015. **5**: p. 14005.

30. Freathy, R.M., et al., *Genetic variation at CHRNA5-CHRNA3-CHRNB4 interacts with smoking status to influence body mass index.* Int J Epidemiol, 2011. **40**(6): p. 1617-28.

31. Robinson, M.R., et al., *Genotype-covariate interaction effects and the heritability of adult body mass index.* Nat Genet, 2017. **49**(8): p. 1174-1181.

32. van Kippersluis, H. and C.A. Rietveld, *Pleiotropy-robust Mendelian Randomization.* bioRxiv, 2016.

33. Wald, A., *The Fitting of Straight Lines if Both Variables are Subject to Error.* 1940: p. 284-300.

34. Bowden, J., et al., *Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic.* Int J Epidemiol, 2016.

35. Droyvold, W.B., et al., *Change in body mass index and its impact on blood pressure: a prospective population study.* Int J Obes (Lond), 2005. **29**(6): p. 650-5.

36. Li, L.J., et al., *Assessing the Causality between Blood Pressure and Retinal Vascular Caliber through Mendelian Randomisation.* Sci Rep, 2016. **6**: p. 22031.

37. Timpson, N.J., et al., *Does greater adiposity increase blood pressure and hypertension risk?: Mendelian randomization using the FTO/MC4R genotype.* Hypertension, 2009. **54**(1): p. 84-90.

38. Holmes, M.V., et al., *Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis.* Am J Hum Genet, 2014. **94**(2): p. 198-208.

39. Spiller, W., N.M. Davies, and T.M. Palmer, *Software Application Profile: mrrobust - A Tool For Performing Two-Sample Summary Mendelian Randomization Analyses.* bioRxiv, 2017.

40. StataCorp, *Stata Statistical Software: Release 14*. 2015: College Station, TX: StataCorp LP.

41. Bowden, J., et al., *Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NOME assumption.* bioRxiv, 2017.

42. Rask-Andersen, M., et al., *Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status.* PLOS Genetics, 2017. **13**(9): p. e1006977.

43. Mackenbach, J.P., *Health and Deprivation - Inequality and the North - Townsend,P, Phillimore,P, Beattie,A.* Health Policy, 1988. **10**(2): p. 207-207.

44. VanderWeele, T.J., et al., *Methodological challenges in mendelian randomization.* Epidemiology, 2014. **25**(3): p. 427-35.

45. Krishna, A., et al., *Trends in group inequalities and interindividual inequalities in BMI in the United States, 1993-2012.* Am J Clin Nutr, 2015. **101**(3): p. 598-605.

Figures

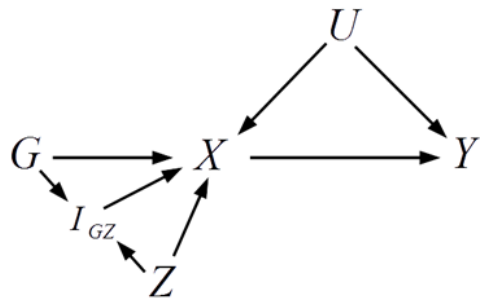

Figure 1: A directed acyclic graph (DAG) showing the assumed relationship between each variable in LSR.

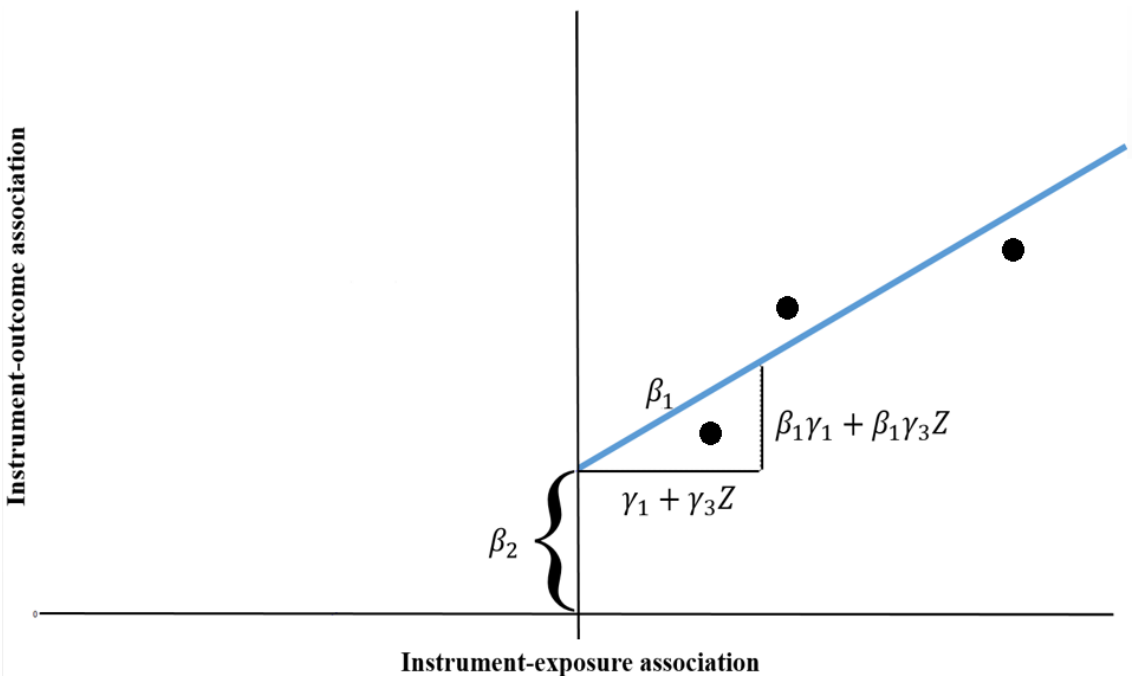

Figure 2: Hypothetical plot showing components of LSR. For a set of Z groups represented as solid points, the x-axis represents the association between the genetic instrument and the exposure, whilst the y-axis shows the association between the genetic instrument and the outcome. The point at which $x = 0$ is an estimate of the theoretical no relevance point, with a remaining (pleiotropic) association between the instrument and the outcome given as the intercept (β_2).

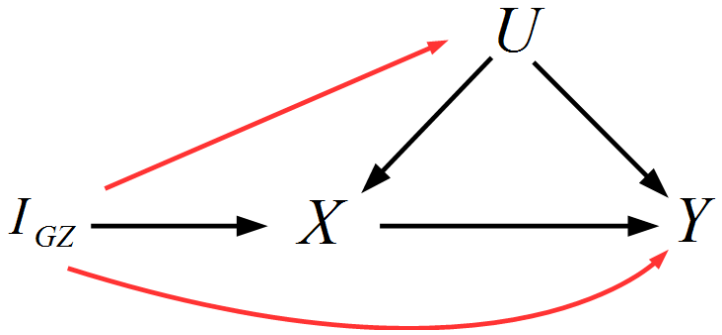

Figure 3: A DAG illustrating the role of the instrument-covariate interaction as an instrument within LSR. In this case, red arrows indicate associations which would invalidate the LSR method, whilst the black arrow from to is necessary to satisfy the relevance assumption analogue.

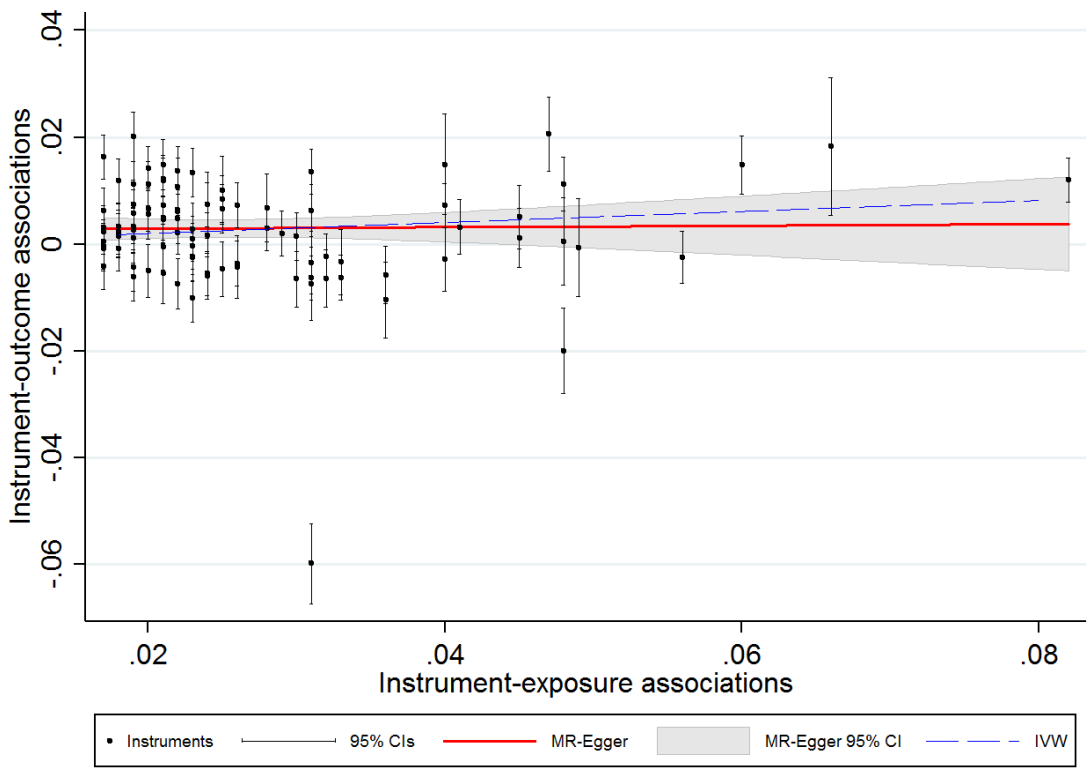

Figure 4: Scatter plot showing IVW and MR-Egger estimates for the effect of BMI upon SBP

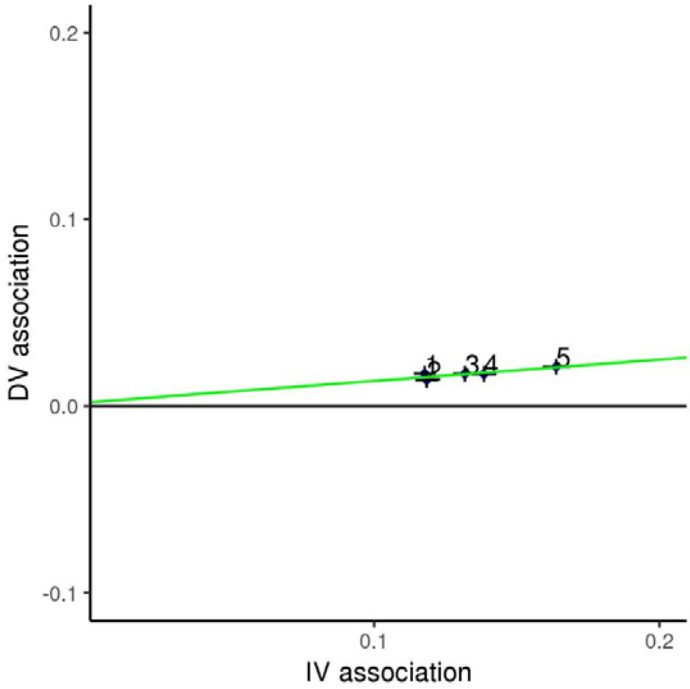

Figure 5: IV estimates using IVW (blue) and LSR (green), with points numbered in ascending order of social deprivation.

Tables

Table 1: OLS and TSLS estimates under differing degrees of pleiotropy

Induced Pleiotropy	***Model***	***Exposure Estimate*** $\boldsymbol{\beta_1 = 1}$	***95% CI***
$\boldsymbol{\beta_2 = 0}$	OLS	1.15	(1.14, 1.16)
	TSLS	1.02	(0.98, 1.06)
$\boldsymbol{\beta_2 = 1}$	OLS	1.22	(1.21, 1.23)
	TSLS	1.97	(1.90, 2.04)

Table 2: Two sample summary MR estimates for the effect of BMI upon SBP

Method	*Estimate*	*SE*	*95% CI*	*p.value*
IVW	0.103	0.030	(0.04, 0.16)	<0.001
MR-Egger (intercept)	0.003	0.002	(-0.00, 0.01)	0.212
MR-Egger (effect) [1]	0.014	0.077	(-0.14, 0.16)	0.851
Weighted Median	0.130	0.020	(0.09, 0.17)	<0.001
Modal Estimator [2]	0.133	0.027	(0.08, 0.19)	<0.001

[1] $I^2_{GX} = 87.8$ [2] Smoothing parameter $\phi = 1$

Table 3: OLS and TSLS effect estimates

	Estimate	*SE*	*95% CI*	*p.value*
OLS				
Intercept	<0.0001	0.002	(-0.004, 0.004)	>0.999
BMI	0.192	0.002	(0.19, 0.20)	<0.001
TDI	-0.056	0.002	(-0.06, -0.05)	<0.001
TSLS				
Intercept	<0.0001	0.002	(-0.004, 0.004)	>0.999
BMI	0.129	0.014	(0.10, 0.16)	<0.001
TDI	-0.050	0.002	(-0.054, -0.047)	<0.001

Table 4: IVW and LSR Estimates using different TDI grouping

Number of groups	*Method*	*Estimate*	*SE*	*95% CI*	*p.value*
2[*]	LSR (intercept)	-0.009	-	-	-
	LSR (effect)	0.197	-	-	-
	IVW	0.132	0.006	(0.05, 0.21)	0.031
5	LSR (intercept)	0.002	0.005	(-0.02, 0.02)	0.728
	LSR (effect)	0.114	0.040	(-0.02, 0.24)	0.065
	IVW	0.129	0.004	(0.12, 0.14)	<0.001
10	LSR (intercept)	0.006	0.013	(-0.02, 0.04)	0.670
	LSR (effect)	0.087	0.096	(-0.01, 0.31)	0.391
	IVW	0.129	0.012	(0.10, 0.16)	<0.001
20	LSR (intercept)	0.002	0.015	(-0.03, 0.03)	0.915
	LSR (effect)	0.117	0.110	(-0.11, 0.35)	0.300
	IVW	0.129	0.015	(0.10, 0.16)	<0.001
50	LSR (intercept)	-0.008	0.013	(-0.03, 0.02)	0.514
	LSR (effect)	0.190	0.094	(0.00, 0.38)	0.050
	IVW	0.129	0.012	(0.10, 0.16)	<0.001

* Note that standard errors using LSR are not defined for the 2-group case.

Table 5: Performance of IVW and LSR methods in simulation setting with null causal effect $\beta_1 = 0$

Case	***N***	***N Per Decile***	***Mean F Statistic***	***IVW Mean Estimate (mean SE)***	***IVW Type I Error Rate***	***LSR Mean Estimate (mean SE)***	***LSR Power of Pleiotropy Test***	***LSR Effect Type I Error Rate***
Case 1:	10000	2000	80.1	0.003 (0.046)	0.049	0.004 (0.066)	0.047	0.050
$\beta_1 = 0$	20000	4000	159.3	0.002 (0.033)	0.050	0.003 (0.047)	0.048	0.050
$\beta_2 = 0$	30000	6000	238.3	0.001 (0.027)	0.052	0.001 (0.038)	0.050	0.050
$\beta_4 = 0$	40000	8000	317.3	0.001 (0.023)	0.052	0.001 (0.033)	0.049	0.049
	50000	10000	396.2	0.000 (0.021)	0.050	0.001 (0.030)	0.049	0.052
Case 2:	10000	2000	80.1	0.091 (0.060)	0.153	0.004 (0.066)	0.209	0.050
$\beta_1 = 0$	20000	4000	159.3	0.089 (0.051)	0.168	0.003 (0.047)	0.365	0.050
$\beta_2 = 0.05$	30000	6000	238.3	0.089 (0.048)	0.157	0.001 (0.038)	0.504	0.050
$\beta_4 = 0$	40000	8000	317.3	0.088 (0.047)	0.150	0.001 (0.033)	0.612	0.049
	50000	10000	396.2	0.088 (0.046)	0.135	0.001 (0.030)	0.708	0.052
Case 3:	10000	2000	80.1	0.082 (0.059)	0.130	0.169 (0.061)	0.262	0.429
$\beta_1 = 0$	20000	4000	159.3	0.081 (0.051)	0.128	0.169 (0.043)	0.442	0.673
$\beta_2 = 0$	30000	6000	238.3	0.080 (0.048)	0.121	0.167 (0.035)	0.579	0.816
$\beta_4 = 0.05$	40000	8000	317.3	0.080 (0.047)	0.100	0.167 (0.031)	0.676	0.895
	50000	10000	396.2	0.079 (0.046)	0.091	0.167 (0.027)	0.770	0.945
Case 4:	10000	2000	80.1	0.169 (0.043)	0.783	0.169 (0.061)	0.048	0.429
$\beta_1 = 0$	20000	4000	159.3	0.168 (0.030)	0.962	0.169 (0.043)	0.051	0.673
$\beta_2 = 0.05$	30000	6000	238.3	0.167 (0.025)	0.994	0.167 (0.035)	0.051	0.816
$\beta_4 = 0.05$	40000	8000	317.3	0.167 (0.022)	0.999	0.167 (0.031)	0.048	0.895
	50000	10000	396.2	0.167 (0.019)	1.000	0.167 (0.027)	0.047	0.945

Table 6: Performance of IVW and LSR methods in simulation setting with positive causal effect $\beta_1 = 0.05$

Case	***N***	***N Per Decile***	***Mean F Statistic***	***IVW Mean Estimate (mean SE)***	***IVW Power to detect causal effect***	***LSR Mean Estimate (mean SE)***	***LSR Power of pleiotropy Test***	***LSR Power to detect causal effect***
Case 5:	10000	2000	80.1	0.053 (0.046)	0.145	0.054 (0.066)	0.047	0.093
$\beta_1 = 0.05$	20000	4000	159.3	0.052 (0.033)	0.219	0.053 (0.047)	0.048	0.123
$\beta_2 = 0$	30000	6000	238.3	0.051 (0.027)	0.288	0.051 (0.038)	0.050	0.148
$\beta_4 = 0$	40000	8000	317.3	0.051 (0.023)	0.355	0.049 (0.033)	0.049	0.181
	50000	10000	396.2	0.050 (0.021)	0.423	0.051 (0.030)	0.049	0.207
Case 6:	10000	2000	80.1	0.141 (0.060)	0.381	0.054 (0.066)	0.209	0.093
$\beta_1 = 0.05$	20000	4000	159.3	0.139 (0.051)	0.495	0.053 (0.047)	0.365	0.123
$\beta_2 = 0.05$	30000	6000	238.3	0.139 (0.048)	0.553	0.051 (0.038)	0.504	0.146
$\beta_4 = 0$	40000	8000	317.3	0.138 (0.047)	0.597	0.051 (0.033)	0.612	0.181
	50000	10000	396.2	0.138 (0.046)	0.630	0.051 (0.030)	0.708	0.207
Case 7:	10000	2000	80.1	0.132 (0.059)	0.341	0.219 (0.061)	0.262	0.615
$\beta_1 = 0.05$	20000	4000	159.3	0.131 (0.051)	0.417	0.219 (0.043)	0.442	0.854
$\beta_2 = 0$	30000	6000	238.3	0.130 (0.048)	0.474	0.217 (0.035)	0.579	0.945
$\beta_4 = 0.05$	40000	8000	317.3	0.129 (0.047)	0.510	0.217 (0.031)	0.676	0.981
	50000	10000	396.2	0.129 (0.046)	0.534	0.217 (0.027)	0.770	0.993
Case 8:	10000	2000	80.1	0.219 (0.043)	0.933	0.219 (0.061)	0.048	0.615
$\beta_1 = 0.05$	20000	4000	159.3	0.218 (0.030)	0.997	0.219 (0.043)	0.051	0.854
$\beta_2 = 0.05$	30000	6000	238.3	0.217 (0.025)	1.000	0.217 (0.035)	0.051	0.945
$\beta_4 = 0.05$	40000	8000	317.3	0.217 (0.022)	1.000	0.217 (0.031)	0.048	0.981
	50000	10000	396.2	0.217 (0.019)	1.000	0.217 (0.027)	0.047	0.993