ISER Working Paper Series

A Comparison of Robust Methods for Mendelian Randomization Using Multiple Genetic Variants

Yanchun Bao
ISER, University of Essex

Paul Clarke
ISER, University of Essex

Melissa C Smart
ISER, University of Essex

Meena Kumari
ISER, University of Essex

No. 2018-08
June 2018

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

University of Essex

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

www.iser.essex.ac.uk

Non-Technical Summary

The causal relationship of an exposure, such as adiposity, to a social outcome, such as income or wellbeing, is often of interest to researchers. However, when there are unobserved confounders (that is, variables related to the exposure and the outcome), the estimated association obtained using conventional regression models is not a causal effect. A causal effect is the effect on the outcome of changing an individual's exposure. Mendelian Randomization (MR) studies are now widely used in epidemiology; these involving using genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure. An IV identifies the causal effect if the IV affects the outcome only through its effect on the exposure. MR studies are thus vulnerable to bias if the genetic variant is pleiotropic and affects the outcome through a pathway other than that through the exposure. MR studies involving many genetic variants have more power to detect true causal effects, but are more prone to pleiotropy bias. Several methods have been proposed to provide robustness when genetic variants are pleiotropic. In this paper, we review, apply and conduct a comprehensive simulation study to access the performance of these methods for *Understanding Society*-like data.

The standard way to use IVs in economics research is to use two-stage least squares (2SLS). The first stage of 2SLS involves using an IV to obtain a prediction of the exposure; the second stage involves replacing the actual exposure with this prediction, and then regressing the outcome on it in the usual way. If we have many genetic variants, the approach used for MR studies is typically to combine these variants into a single polygenic/polygenetic score. If any of the genetic variants are pleiotropic, MR-Egger can be used to produce pleiotropy-robust estimates provided that the size of the direct effect of the genetic variant on the outcome is independent of its association with the exposure: the so-called InSIDE condition. The MR-median estimate is simply the median of the 2SLS estimates obtained using each genetic variant in turn as the only IV (rather than in combination); it is pleiotropy-robust if less than half of the genetic variants are pleiotropic. The last method that we evaluate is the recently-developed some invalid some valid instrumental variable estimator (sisVIVE). This is method that fits 2SLS subject to the constraint that less than half of the genetic variants are pleiotropic. We investigate the performance of these methods by simulating data under various scenarios involving pleiotropic SNPs.

We found, as expected, that 2SLS and polygenetic score-based methods will be biased when SNPs are pleiotropic. Among the robust methods, we found that sisVIVE outperformed MR-Median and MR-Egger across a range of scenarios. However, its performance could be poor in absolute terms, and particularly in the presence of 'indirect' pleiotropy where the genetic variants were related to omitted variables linked to both exposure and outcome. This is known to lead to failure of the key 'InSIDE' condition for MR-Egger, but we found it also affects sisVIVE despite not being formally required for identification. We argue that this is because the consistency criterion for sisVIVE cannot identify the true causal effect if there is indirect pleiotropy. In the application to *Understanding Society*, we found no evidence for pleiotropic bias, and the negative effect of body mass index (BMI) on personal income to be

around five times larger than the observational association. However, this conclusion depended on the unverifiable assumption that InSIDE holds.

It is very important to, as far as possible, reduce the bias of estimates when seeking to understand the causal relationship between two variables like BMI and income. A MR study would estimate unbiased causal relationship between exposure and outcome if SNPs are not pleiotropic. Our comparisons provide guidance about in which case these estimates are trustable and in which they are not. Our work is focusing on methodological discussion and has no direct effect on policy making but would help MR researchers to better understand the conclusions they made and therefore could have downstream impact on policy making.

A comparison of robust methods for Mendelian Randomization using multiple genetic variants

Yanchun Bao,* Paul S Clarke, Melissa C Smart, Meena Kumari.

Institute for Social & Economic Research, University of Essex, UK

* Correspondence to: Yanchun Bao, Institute for Social & Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ UK, email: ybaoa@essex.ac.uk, Tel: 01206874780

Abstract

We report the results of a Mendelian randomization study in which multiple genetic variants are used as instrumental variables to estimate the causal effect of body mass index on personal income in the presence of unobserved confounding. The data come from *Understanding Society*, a large-scale longitudinal household survey, and the GIANT consortium study. Mendelian randomization studies are known to be affected by both weak instrument bias and the pleiotropic bias that arises when some genetic variants are invalid instrument variables. We review and compare some of the recently developed techniques for using multiple genetic variants as instrumental variables. Our principal focus, however, is to assess the 'some invalid some valid instrumental variable estimator' (sisVIVE) developed by Kang et al. (2016). We conduct a comprehensive simulation study to assess sisVIVE for *Understanding Society*-like data, and find that it outperforms alternative methods across a range of scenarios. However, its performance is poor in absolute terms when the presence of indirect pleiotropy leads to failure of the key 'InSIDE' condition, despite this not being explicitly required for identification. We argue that this is because the consistency criterion for sisVIVE does not identify the true causal effect if InSIDE fails. In the application to *Understanding Society*, we find no evidence for pleiotropic bias, and the negative effect of body mass index on income to be around five times larger than the observational association. However, this conclusion depends on the unverifiable assumption that InSIDE holds.

Key words: Mendelian randomization, instrumental variables, pleiotropic bias, MR-Egger, MR-Median, sisVIVE

1 Introduction

Mendelian randomization studies are now widely used in epidemiology [1] [2] and social and economic research [3] [4, 5]. These studies involve the use of genetic variants as instrumental variables to estimate the causal effects of modifiable exposures on outcomes from observational data. Instrumental variable (IV) methods can potentially overcome the problems of unobserved confounding and reverse causation when estimating causal effects using observational data. The genetic variant is a valid IV if a) it is associated with the exposure, b) it has no direct effect on the outcome, and c) it has no indirect effect on the outcome mediated by the unobserved confounders.

The genetic variants used in Mendelian randomization studies are called single nucleotide polymorphisms (SNPs). SNPs are variations in the DNA sequence where individuals differ from each other in terms of a single nucleotide. Each variant form is called an allele and common SNPs usually have two alleles. An individual can have none, one or two copies of a specific allele. The expression of a selected SNP, called its phenotype, should be the exposure or an observable trait associated with this exposure. If only one SNP is used, the causal effect of the exposure on the outcome can be estimated using the ratio of the estimated coefficient of the SNP - from the regression of the outcome on this SNP - and the corresponding coefficient from the regression of the exposure on the SNP [6]. If the SNP is a valid IV, and the causal relationship between exposure and outcome is linear, the ratio estimator will be consistent but not unbiased for the true causal effect.

A well-known problem with the ratio estimator is 'weak instrument' bias, that is, the bias that arises when the instrumental variables are insufficiently predictive of the exposure [7] [8]. In practice, this problem tends to arise for IVs that are weakly correlated with the exposure, which is typically the case with Mendelian randomization studies. One strategy for avoiding weak instrument bias is to use more than one SNP. The causal exposure effect can be estimated by, for example, combining the SNPs into what are sometimes called allele, or genetic risk, scores [9], which we herein refer to as polygenic risk scores. An alternative strategy is to use two-stage least squares (2SLS) with the SNPs as multiple IVs [10]. The rationale for using more than one SNP is that additional SNPs contain extra information which, it is hoped, can be combined to predict exposures more accurately and so alleviate weak instrument bias.

In this paper, we use Mendelian randomization to estimate the causal effect of body mass index (BMI) on personal income using data from *Understanding Society*: *The UK Longitudinal Household Study* (UKHLS). UKHLS contains rich genetic data from which we obtain 71 out of the 97 common genetic variants found to be associated with BMI in genome-wide association studies (GWAS) at the genome-wide level of significance [11]. We adopt the multiple SNPs strategy here because the single genetic marker most strongly associated with BMI from FTO gene (rs1558902) explains only 0.27 percent of the variation.

However, the use of multiple SNPs presents two further challenges. The first is that, even if all the SNPs were valid IVs, the combined explanatory power of the additional SNPs could still be small enough to lead to *many* weak instruments bias [12]. This is certainly possible here because even combining all 97 BMI-associated SNPs from the GIANT study would have explained only 2.7 percent of the total variation in BMI [11]. While many weak instruments bias can be considerably reduced by using the limited information maximum likelihood estimator, this comes at the cost of a substantial loss of power. An alternative is to use a 'two-sample' approach by taking estimates of the associations between the SNPs and exposure from another, ideally much larger, data set [13]. Provided the second sample is drawn from a comparable population to the first, and the estimates are precise, this approach can considerably reduce weak instrument bias without any major loss of power. Hence, we also use a two-sample approach for our analysis, with SNP-BMI estimates taken from the GIANT study [11].

Whether a one or two-sample approach is used, the major challenge facing all Mendelian randomization studies is that one or more of the SNPs is not a valid IV satisfying conditions a) to c) given above. Such SNPs are related to multiple traits and so said to be 'pleiotropic'. There has been recent work on methods to adjust for pleiotropic bias. These methods are based on a joint model which relates the multiple SNP-exposure and SNP-outcome associations. Two widely used and relatively simple techniques which offer robustness against invalid IVs are 'MR-Egger' regression [14] and 'MR-Median' regression [15]. MR-Egger is regarded as being robust to invalid IVs if condition c) holds, while MR-Median is apparently robust if less than half of the SNPs are invalid IVs. However, the principal focus of this paper is on the effectiveness of the recently developed 'some invalid, some valid instrumental variable estimator' (sisVIVE) [16]. Like MR-Median, sisVIVE is theoretically robust to invalid IVs if more than half of the SNPs are valid instrumental variables.

We use the methods above to estimate the causal effect of BMI on personal income. Our main methodological aim is to evaluate the performance of sisVIVE relative to the more established approaches, and in absolute terms, for scenarios like our data example. In doing this, we contribute to the understanding of, and good practice for, the use of these methods in epidemiological research. The remainder of the paper is structured as follows: in Section 2, we provide model assumptions for identification of causal effects; in Section 3, we briefly review and compare the methods we consider for using multiple SNPs; in Section 4, we present the results from a simulation study that mimics data to assess the relative performance of these methods, and discuss the identification of sisVIVE; in Section 5, we present our real-data application of these methods to estimate the causal effect of BMI on personal income; and we conclude by discussing our findings and potential avenues for future work in the future at Section 6.

2 Modelling for Mendelian randomization studies

2.1 Choosing SNPs

A single-SNP Mendelian randomization study will typically involve a SNP for which there is robust evidence that it is a genotype for the exposure. This evidence is usually obtained from a dedicated genome-wide association study (GWAS). A GWAS involves estimating the associations between a genome-wide set of genetic variants (typically SNPs) and biological traits to identify those variants which are associated with each trait. GWAS estimates are adjusted for errors due to multiple testing, potential confounders of these associations and for population groups with different genotype distributions (so-called population stratification). The genome-wide level of significance is determined statistically; a p-value significance threshold of 5×10^{-8} has become a widely accepted.

The accuracy of a GWAS in determining which SNPs are associated with which traits depends on the sample size and, particularly, on the adequacy of the confounding and population-stratification adjustments. There is also a risk that the SNP will be pleiotropic so that either condition b) or c) does not hold, which would lead to biased estimates even if the association between the SNP and exposure were not weak [17]. In considering the impact of pleiotropic bias, we distinguish between 'direct' pleiotropy resulting from failure of condition b), and 'indirect' pleiotropy resulting from failure of condition c).

As explained above, the UKHLS contains 71 of the 97 SNPs identified by the GIANT consortium as being associated with body mass index [11]. The full list of these SNPs is given in Table S1 in the Supplementary Information; the distributions of each SNP in UKHLS and GIANT are given in the table.

2.2 Modelling Assumptions

We denote the outcome variable as Y and the exposure as X, and consider scenarios in which Y can be treated as a continuous variable that is causally related to X by the linear model

$$Y = \gamma_0 + \gamma_X X + \epsilon_Y, \quad (1)$$

where γ_X is the causal exposure effect, and ϵ_Y is the model error comprising the combined effect of every influence (apart from X) on outcome Y. Any adjustments for observed confounding variables associated with exposure and outcome may be included but, for notational simplicity, we have omitted these from (1). It is supposed either way that there remains substantial unobserved confounding because we have omitted important confounding variables, which are absorbed into ϵ_Y and thus induce an association between ϵ_Y and X. In such cases, standard regression estimation of (1) using ordinary least squares (OLS) or generalised least squares would be inconsistent and biased for causal parameters like γ_X.

The precise interpretation of γ_X depends on the assumptions we make about exposure-effect heterogeneity, that is, between-person variation in the causal effect of BMI on personal income. For two-stage least squares (to be introduced below), the estimate of γ_X can be interpreted as the average causal effect of BMI on personal income, if the effect of BMI is the same for everyone , or its between-person variation is independent of the exposure given the genetic IVs (and any covariates included to adjust for observed confounding); see Chapter 5.2 in [18].

Mendelian randomization studies involve choosing one or more SNPs to use as IVs. Suppose that we use GWAS studies to identify J SNPs which we denote by $\mathbf{G} = (G_1, \dots, G_J)'$. Each SNP takes values $G_j \in \{0,1,2\}$ to indicate the number of times that the specific the allele associated with increased exposure was found at this gene location. This allele is referred to as the 'risk' or 'effect' allele, and the other as the 'base' or 'non-effect' allele. The effect allele for each of the 71 SNPs used here can be found in Table S1 of the Supplementary Information.

Using this notation, we can rewrite conditions a) to c) for a single-SNP, G_j, in a Mendelian randomization study as follows: a) $\text{Cov}(X, G_j) \neq 0$; (b) $\text{Cov}(Y, G_j|X) = 0$ (no direct pleiotripy); and (iii) $\text{Cov}(\epsilon_Y, G_j) = 0$ (no indirect pleiotropy); conditioning on observed confounding variables is implicit. Conditions a) to c) are sometimes referred to as the core conditions [10], and G_j is a valid IV only if it satisfies these conditions.

The final point to make here is that the SNPs will be assumed to be drawn from distinct gene regions such that the G_j are mutually independent at the population level.

3 Estimating causal exposure effects

The simplest IV estimator is the ratio estimator based on candidate IV, Z, which we can write as

$$\hat{\gamma}_{X(Z)} = \frac{\hat{\Gamma}_{(Y:Z)}}{\hat{b}_{(X:Z)}}, \quad (3)$$

where $\hat{\Gamma}_{(Y:Z)}$ is the OLS estimate of the coefficient of Z from the simple linear regression of Y on Z, and $\hat{b}_{(X:Z)}$ is the OLS estimate of Z from the simple linear regression of X on Z. If observed confounding variables are included in (1) as covariates, the outcome and exposure above are respectively replaced by the residuals obtained from regressing each in turn on these covariates.

We denote the resulting ratio estimator for putative IV $Z = G_j$ by $\hat{\gamma}_{X;j} = \hat{\gamma}_{X(G_j)}$, its numerator by $\hat{\Gamma}_j = \hat{\Gamma}_{(Y:G_j)}$ and its denominator by $\hat{b}_j = \hat{b}_{(X:G_j)}$. If G_j is a valid IV then $\hat{\gamma}_{X;j}$ is consistent for the causal exposure effect γ_X but, as discussed above, it may be subject to considerable weak instrument bias.

3.1 Estimation with multiple SNPs that are all valid IVs

Polygenic risk scores: The properties of polygenic risk scores (also known as allele scores and genetic risk scores) in Mendelian randomization studies are reviewed in detail elsewhere [9]. To summarise, polygenic risk scores have the general form

$$G = \sum_{j=1}^{J} w_j G_j, \quad (4)$$

where w_j is a user-specified weight for G_j. If w_j takes the same value for every SNP then G is called the simple polygenic risk score (SPRS). The ratio estimator (3) using SPRS as the instrumental variable can be written as

$$\hat{\gamma}^{SPRS} = \sum_{j=1}^{J} \hat{\gamma}_{X;j} \left\{ \frac{\hat{b}_j \hat{v}_j}{\sum_{j'} \hat{b}_{j'} \hat{v}_{j'}} \right\}, \quad (5)$$

where $\hat{v}_j$ is a consistent estimate of $v_j = \mathrm{Var}(G_j)$. Note that the $\hat{v}_j$ in (5) can be equivalently replaced with estimates of $\sigma_{Y;j}^{-2} = 1/\mathrm{Var}(\hat{\Gamma}_j)$, that is, the inverse of the estimated standard error of $\hat{\Gamma}_j$ [14]. The term in parenthesis on the right-hand side of (5) can be viewed as the weight for $\hat{\gamma}_{X;j}$ in a weighted sum of SNP-specific ratio estimates. This weight ensures that any SNP weakly correlated with the exposure, or which varies little between individuals (relative to other SNPs), makes only a small contribution to $\hat{\gamma}^{SPRS}$.

An alternative to the SPRS is the internally weighted polygenic risk score (IPRS) with $w_j = \hat{\beta}_j$. The IPRS can be written

$$\hat{\gamma}^{IPRS} = \sum_{j=1}^{J} \hat{\gamma}_{X;j} \left\{ \frac{\hat{b}_j^2 \hat{v}_j}{\sum_{j'} \hat{b}_{j'}^2 \hat{v}_{j'}} \right\}, \quad (6)$$

which is equal to (5) but with $\hat{b}_j^2$ rather than $\hat{b}_j$ appearing in the weight. In fact, this is a very good approximation of the 2SLS estimator (see below) obtained using the set of SNPs **G** as multiple IVs which would hold exactly if the sample SNPs were perfectly uncorrelated [14]. If (6) is used with summarised rather than individual-level data then $\hat{\gamma}^{IPRS}$ is known as the inverse weighted (IVW) estimator [14, 19].

Comparing (5) with (6) reveals that the two estimators differ in terms of how each SNP-specific ratio estimate is weighted. The numerator and denominator of the SPRS weight depend on the signs of $\hat{b}_j$ so it makes sense that G_j is always coded as the number of effect alleles to ensure that the sign of $\hat{b}_j$ is always positive and the numerator is non-zero. Because IPRS is equivalent to two-stage least squares if the SNPs are independent, an advantage of IPRS (6) is that it combines the SNP-specific estimates efficiently; the combination is

efficient in that the standard error of $\hat{\gamma}^{IPRS}$ is as small as possible (among all consistent and asymptotically normal estimators) provided that ε_Y in (1) is homoscedastic; the standard error of $\hat{\gamma}^{SPRS}$ will be similarly small only if $b_j = b$ for all j.

Two-stage Least Squares: The two-stage least squares (2SLS) estimator can be described as follows. Stage one involves fitting the so-called reduced-form model

$$X = \mathbf{z}'\mathbf{b} + \epsilon_X, \quad (7)$$

where $\mathbf{z}' = (1, \mathbf{G}')$ comprises a constant term and the vector of SNPs, $\mathbf{b} = (b_0, \ldots, b_J)'$ are the SNP-exposure effects, and ϵ_X is the model residual satisfying $E(\epsilon_X|\mathbf{G}) = 0$ such that $b_j = \mathrm{Cov}(X, G_j)/\mathrm{Var}(G_j)$ for $j \geq 1$ if the SNPs are independent. Stage two involves calculating $\hat{X} = \mathbf{z}'\hat{\mathbf{b}}$ for each individual, and regressing Y on $\hat{X}$ using OLS; the 2SLS estimator $\hat{\gamma}_X^{2SLS}$ is the estimated coefficient of $\hat{X}$ obtained from the stage-two regression.

The problem of many weak instruments bias for the 2SLS estimator is explored in detail by Davies et al. (2015) [12]. Using a similar notation to theirs, the 2SLS estimator can be written as

$$\hat{\boldsymbol{\gamma}}^{2SLS} = \underset{\boldsymbol{\gamma}}{\mathrm{argmin}}[\{\boldsymbol{\varepsilon}_Y(\boldsymbol{\gamma})Z\}W_N\{\boldsymbol{\varepsilon}_Y(\boldsymbol{\gamma})Z\}'], \quad (8)$$

where the observed data on the sample individuals are $\{\mathbf{z}_i, X_i, Y_i : i = 1, \ldots, N\}$, Z is the $N \times (J+1)$ design matrix with row i given by $\mathbf{z}_i'$, $W_N = (Z'Z)^{-1}$ is a $(J+1) \times (J+1)$ weight matrix, $\boldsymbol{\varepsilon}_Y(\boldsymbol{\gamma}) = (\varepsilon_{Y;1}, \ldots, \varepsilon_{Y;N})$, $\varepsilon_{Y;i} = Y_i - \gamma_0 - \gamma_X X_i$ is the vector of residuals from model (1) for individual i, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_X)'$. The bias of $\hat{\gamma}_X^{2SLS}$ has been shown to depend multiplicatively on the ratio of the covariance between ϵ_Y and ϵ_X, the inverse variance of $\hat{X}$, and the number of SNPs so that, the greater the number of SNPs, the greater the bias [12].

Limited information maximum likelihood and the generalized method of moments: Limited information maximum likelihood (LIML) is an alternative to 2SLS that can be used to reduce the impact of many weak instruments bias. It remains biased but is considerably less biased than 2SLS because its bias does not increase as further SNPs are included in the analysis [12].

Davies et al. (2015) [12] also showed that 2SLS is a 'one-step' generalized method of moments (GMM) estimator (Hansen, 1982) [20], and that different choices of weight matrix

W_N result in different types of GMM estimator. Specifically, they showed that the 'continuously updated estimator' (CUE) involves using (8) but with the weight matrix

$$W_N(\boldsymbol{\gamma}) = [\{\boldsymbol{\varepsilon}_Y(\boldsymbol{\gamma})Z\}'\{\boldsymbol{\varepsilon}_Y(\boldsymbol{\gamma})Z\}]^{-\mathbf{1}}. \quad (9)$$

They also showed that CUE is effectively an extension of LIML to allow for heteroscedastic errors in (1) and so potentially more efficient than LIML. If only one SNP were available then 2SLS, LIML and CUE would be identical [10] but, with multiple SNPs, LIML and CUE reduce bias avoid overfitting to the sample data and thus reduce bias. The conventional standard error estimates produced by GMM for both LIML and CUE can be corrected for negative bias if the instruments are weak [12].

However, both LIML and CUE are less numerically stable estimators than 2SLS, so it is recommended that estimation is repeated from multiple starting values to check that the initial solution is the global maximum. We focus on LIML in the subsequent simulation study involving only valid IVs (Section 4.1) to demonstrate how more advanced GMM estimators improve on 2SLS; the reader is referred to [12] for a full comparison of 2SLS, LIML and CUE.

3.2 Estimation with multiple SNPs where some are invalid IVs

If any of the chosen SNPs were invalid IVs then every estimator discussed in Section 3.1 would be biased, regardless of whether the instruments were weak or not, because pleiotropy would lead to model (1) being incorrectly specified. Bowden et al. (2015) [14] proposed the following model to incorporate the impact of pleiotropy:

$$Y = \pi_0 + \gamma_X X + \sum_{j=1}^{J}(\alpha_j + \theta_j)G_j + \epsilon_Y, \quad (10)$$

where $\alpha_j \neq 0$ if there is a direct pleiotropic bias (that is, the effect of G_j leads to failure of core condition b)), and $\theta_j \neq 0$ if there is indirect pleiotropic bias (that is, failure of core condition c)). The model error ϵ_Y satisfies $E(\epsilon_Y|\mathbf{G}) = 0$.

Under model (10), the relationship between the true numerators and denominators of the ratio estimators is

$$\Gamma_j = \pi_j + \gamma_X b_j, \quad (11)$$

where $\pi_j = \alpha_j + \theta_j$ is the sum of the pleiotropic errors related to SNP $j = 1, \dots, J$. This relationship drives the 'consistency criterion' that identifies the causal exposure effect when the identity of the invalid-IV SNPs is unknown [16].

MR-Egger regression: Bowden et al. (2015) [14] developed MR-Egger regression by adapting the Egger regression technique from the meta-analysis literature to Mendelian randomization. MR-Egger involves treating estimates $\{\hat{\Gamma}_j, \hat{b}_j : j = 1, \dots, J\}$ as data and fitting the simple linear regression

$$\hat{\Gamma}_j = \gamma_0^{\text{Egg}} + \gamma_X^{\text{Egg}} \hat{b}_j + \epsilon_j^{\text{Egg}}, \quad (12)$$

where ϵ_j^{Egg} is a residual that must satisfy $E\left(\epsilon_j^{\text{Egg}} \middle| b_j\right) = 0$ (noting that expectation is with respect to the set of chosen SNPs). Weighted least-squares regression can alternatively be used with weights $\hat{\upsilon}_j$ or $\hat{\sigma}_{Y;j}^{-2}$ to account for differential minor allele frequencies (MAFs) such that SNPs with low MAFs contribute little to the estimate of (12). If InSIDE holds, the target parameter is $\gamma_X^{\text{Egg}} = \gamma_X$ and the true intercept term γ_0^{Egg} is the average pleiotropic effect $E(\pi_j)$. If all SNPs in the analysis were valid IVs then $\gamma_0^{\text{Egg}} = 0$, but this is unlikely to be the case in practice.

The requirement that $E\left(\epsilon_j^{\text{Egg}} \middle| b_j\right) = 0$ is called the 'InSIDE' condition and is key to identifying $\left(\gamma_0^{\text{Egg}}, \gamma_X^{\text{Egg}}\right)$ [14]. InSIDE is thought to be plausible if (the multiple-SNPs equivalent of) core condition b) fails, but is generally unrealistic if condition c) fails because failure leads to both $\hat{b}_j$ and ϵ_j^{Egg} depending on θ_j [14]. MR-Egger also requires that $\hat{b}_j$ is precisely estimated because otherwise γ_X^{Egg} will be biased towards zero (see the discussion of the 'no measurement error assumption' in Section 3.4).

MR-Median: Bowden et al. (2016) [15] proposed MR-Median as an alternative to MR-Egger that, in theory, does not require the InSIDE condition to hold. The authors view MR-Median as a practicable alternative to the sisVIVE estimator to be introduced next. The main strengths of MR-Median are its simplicity and that it can be used if only summary data are available.

The MR-Median estimator is simply the median of the SNP-specific ratio estimates. In other words, if $\{\hat{\gamma}_{X(j)} : j = 1, \dots, J\}$ is the ordered set of ratio estimates (i.e. such that $\hat{\gamma}_{X(j+1)} \geq \hat{\gamma}_{X(j)}$

for all $j = 1, \ldots, J-1$) then $\hat{\gamma}_X^{\text{Med}} = \hat{\gamma}_{X(J/2)}$ If J is even, or $\hat{\gamma}_X^{\text{Med}} = \{\hat{\gamma}_{X(J/2)} + \hat{\gamma}_{X(J/2+1)}\}/2$ if J is odd. The inverse variance weighted version of MR-Median is $\hat{\gamma}_X^{\text{Med}} = \hat{\gamma}_{X(j)} + (\hat{\gamma}_{X(j+1)} - \hat{\gamma}_{X(j)}) \times (0.5 - s_j)/(s_{j+1} - s_j)$, where $s_j = \sum_{k=1}^{j} \hat{\sigma}_{\hat{\gamma}_{X(k)}}^{-2}$ and j is the largest integer such that $s_j < 0.5$.

MR-Median is consistent for γ_X if less than $J/2$, or 50 percent, of the SNPs are invalid instrumental variables (the consistency requirement for the weighted MR-Median is that at least 50 percent of the weight comes from valid instrumental variables). Like MR-Egger, it can be used with summary as well as individual-level data.

3.3 Some invalid, some valid instrumental variables estimator (sisVIVE)

The next estimator is the main focus of our study. Kang et al. (2016) [16] developed the 'some invalid some valid instrumental variable estimator' (sisVIVE) by adapting the LASSO for 2SLS. SisVIVE works by penalizing SNPs which are inconsistent with model (1) (that is, have pleiotropic effects) so that any invalid instruments are removed when fitting model (10). The parameters of (10) are shown to be identified if less than half of the SNPs are not valid IVs, whichever of conditions b) or c) fail [16].

Essentially, sisVIVE is an extension of 2SLS (8) under model (2) with the addition of a penalisation term. We write it as follows:

$$\begin{pmatrix} \hat{\gamma}_{X;\lambda} \\ \hat{\boldsymbol{\pi}}_\lambda \end{pmatrix} = \underset{\boldsymbol{\gamma},\boldsymbol{\pi}}{\operatorname{argmin}} \left[\frac{1}{2} \{\boldsymbol{\varepsilon}_Y(\gamma_X, \boldsymbol{\pi}) Z\} W_N \{\boldsymbol{\varepsilon}_Y(\gamma_X, \boldsymbol{\pi}) Z\}' - \lambda \parallel \boldsymbol{\pi} \parallel_1 \right], \quad (13)$$

where λ is the pre-specified scalar tuning parameter, $\parallel \boldsymbol{\pi} \parallel_1 = \sum_{j=1}^{J} |\pi_j|$ is the standard ℓ_1-norm, W_N is the same weight matrix as was used in (8), $\boldsymbol{\varepsilon}_Y(\gamma_X, \boldsymbol{\pi}) = (\varepsilon_{Y;1}, \ldots, \varepsilon_{Y;N})$, $\varepsilon_{Y;i} = Y_i - \gamma_X X_i - \boldsymbol{\pi}' \mathbf{z}_i$ is the value of the residuals in model (10) for individual i, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)'$. If both core conditions b) and c) hold then (13) reduces to 2SLS (8) because $\boldsymbol{\pi}$ equals zero. Note that, in practice, the outcome and instrumental variable are mean centred so that $\mathbf{z}_i = \mathbf{G}_i$ and π_0 from model (10) equals zero.

The resulting procedure is closely related to the ℓ_1-penalization used by the LASSO [21]. The difference is that sisVIVE only penalises $\boldsymbol{\pi}$ and not γ_X as it searches for solutions consistent with (11). The choice of tuning parameter is crucial: a large value of λ would force

heavy penalization of $\boldsymbol{\pi}$ towards zero so that all the SNPs were treated as valid instruments; conversely, a low value of λ would lead to all the SNPs being treated as invalid instruments. The optimal λ is obtained using the same cross-validation method coupled with the 'one standard error' rule as used for the LASSO [16, 21]. Windmeijer et al. (2017) [22] have recently detailed further properties of the sisVIVE estimator including conditions under which it will not be consistent.

The estimated standard errors produced by penalized regression methods generally perform poorly, so we propose to use sisVIVE in conjunction with other valid-IV methods. The resulting approach thus consists of two steps in which step one involves using sisVIVE to select a set of valid IVs, and step two involves applying the chosen valid-IV method from Section 3.1 to the selected IVs. Post-estimation approaches like this have been found to have better finite-sample properties than $\hat{\gamma}_{X;\lambda}$ [23-25].

Finally, we make the following note about sisVIVE in the presence of heterogeneous causal effects. In Section 2.2, we noted for 2SLS that γ_X can be interpreted as the average causal effect if the effect of BMI is homogenous, or there is heterogeneity but the between-individual variation in BMI effects is independent of the exposure given the genetic IVs. Kang et al. [16] showed that the same assumptions are required if we wish to interpret γ_X as the average causal effect when using sisVIVE. However, we have shown that potential dependence of the mean treatment effect on the SNPs can result in $\pi_j \neq 0$ even if SNP j is a valid IV, but that this can be prevented if we can further assume that the between-individual variation in BMI causal effects is mean independent of both exposure and the SNPs (see Section S2.1 in the Supplementary Information).

3.4 Two-sample strategies

In the discussion so far, it has been assumed that the analysis will be based on one individual-level data set. It is, however, also possible to adopt a two-sample strategy if estimates of the SNP-exposure estimates are available from another, preferably larger, study [26-28]. We denote the estimates obtained from this second sample as $\{\widetilde{b_j}: j = 1, \ldots, J\}$.

Estimates from a second sample can be incorporated into the approaches described above as follows:

- Simple polygenic risk scores: as before.

- Externally weighted polygenic risk score: choose $w_j = \widetilde{\beta}_j$ to give externally weighted polygenic risk score (EPRS) $\hat{\gamma}^{EPRS}$. This is the two-sample version of IPRS; the resulting estimator is equal to (6) but with $\hat{b}_j\tilde{b}_j$ rather than $\hat{b}_j^2$ in the weights for the SNP-specific contributions.
- Two-stage least squares: Use $\tilde{X} = \mathbf{z}'\widetilde{\boldsymbol{\beta}}$ in stage two rather than $\hat{X}$.
- MR-Egger: Fit $\hat{\Gamma}_j = \gamma_0^{\text{Egg}} + \gamma_X^{\text{Egg}}\tilde{b}_j + \epsilon_j^{\text{Egg}}$ rather than (12).
- MR-Median: Use the (possibly weighted) median of $\hat{\tilde{\gamma}}_j = \hat{\Gamma}_j/\widetilde{b_j}$ from $\{\hat{\tilde{\gamma}}_{(j)}: j = 1, \ldots, J\}$.
- sisVIVE: Use $\varepsilon_{Y;i} = Y_i - \gamma_X\tilde{X}_i - \boldsymbol{\pi}'\mathbf{z}_i$, where $\tilde{X}_i = \mathbf{z}_i'\tilde{\mathbf{b}}$, and apply (13).

If the IVs were all valid, we could additionally use $\hat{\tilde{\gamma}}_j = \hat{\Gamma}_j/\widetilde{b_j}$ rather than $\hat{\gamma}_j$ in SPRS or EPRS to further reduce weak instrument bias because, in contrast to $\hat{b}_j$ and $\hat{\Gamma}_j$, there would be no association between $\tilde{b}_j$ and $\hat{\Gamma}_j$ [13]; in fact, if $\tilde{b}_j = b_j$ then $\hat{\tilde{\gamma}}_j$ would be unbiased because it would reduce to an OLS estimate from the regression of Y on the known and unconfounded variable $G_j b_j$. More realistically, we require that the standard error of $\tilde{b}_j$ to be small, and that both samples are drawn from the same population (or at least from two populations that were similar in terms of known characteristics like ethnic group related to population stratification), for $\tilde{b}_j \simeq b_j$ to hold [29, 30]. Such scenarios satisfy what is called the no measurement error (NOME) condition [13]. However, if $\tilde{b}_j$ is not precisely estimated then NOME will not hold. NOME is so-called because the imprecision means that $\tilde{b}_j = b_j + \mu_j$, where μ_j behaves like mean-zero measurement error if $\tilde{b}_j$ is unbiased; the estimate is hence subject to an attenuation bias towards zero if NOME fails. This second form of weak instrument bias can be viewed as a kind of conservative shrinkage, which is less harmful than the first because it makes it more difficult to reject the null hypothesis of no causal effect.

In the same spirit as the second stage of 2SLS, sisVIVE can also be implemented as part of a two-sample strategy by simply replacing X in (13) with $\widetilde{X}_i = \mathbf{z}_i'\tilde{\mathbf{b}}$; this is equivalent to performing a standard sisVIVE under the constraint $Z\{(Z'Z)^{-1}Z'X\} = Z\tilde{\mathbf{b}}$. Like MR-Egger, neither MR-Median nor sisVIVE are robust to failure of NOME.

4 Simulation results

We now carry out a simulation study to explore the effectiveness of sisVIVE for the types of scenario we believe to be plausible for UKHLS. Two sets of scenarios are considered: in the first, the SNPS are all valid instrument variables; and in the second, we allow some SNPs to be invalid instrumental variables. The design of the study is based on those elsewhere (e.g. Bowden et al. 2015) [14], but the data are simulated in such a way as to mimic key characteristics of the UKHLS data. For example, the 71 SNPs $G_1, \ldots, G_{71}$ are generated independently from a trinomial distribution in which the probabilities of G_j being 0, 1, 2 are respectively equal to the proportions of G_j being 0, 1 and 2 in the UKHLS data (see Table S1 in the Supplementary Information). The causal association of BMI and SNP j is denoted by β_j and set to a value of which is taken from the GIANT consortium study [11]. The true value of the causal effect of BMI on income is $\gamma_X = -0.2$. All the results presented below are based on 1000 generated samples each of size $N = 10\,000$.

We use the following *Stata* functions to implement the methods described above: ivreg2 to implement 2SLS or LIML with single IV/multiple IVs; mregger from the mrrobust package to implement IVW/MR-Egger method; and mrmedian from the mrrobust package to perform MR-Median. The R package sisVIVE (downloaded from CRAN at https://cran.r-project.org/web/packages/sisVIVE.html) is used to implement sisVIVE.

4.1 Scenarios where all SNPs are valid IVs

To simulate data subject to unobserved confounding, the error terms in model (1) and model (7) are respectively decomposed as $\epsilon_Y = U\gamma_U + \dot{\epsilon}_Y$ and $\epsilon_X = U + \dot{\epsilon}_X$, where U is a zero-mean variable representing unobserved confounding, and $\dot{\epsilon}_Y$ and $\dot{\epsilon}_X$ are not only mutually independent but jointly independent of $(Y, X, \mathbf{G}', U)$. Parameter $\gamma_U = 1$ indexes the extent of unobserved confounding by controlling the strength and sign of the correlation between ϵ_Y and ϵ_X, such that there is no unobserved confounding if $\gamma_U = 0$. Values of the outcome and exposure are respectively generated under model (1) and model (7), with $U, \dot{\epsilon}_X$ and $\dot{\epsilon}_Y$ all independently generated from standard normal distributions. The average of F-statistics of the SNPs with BMI exposure X is 2.5, which indicates that these are weak instruments.

Using a one-sample strategy, we compare the performance of sisVIVE on the generated data with that of SPRS, IPRS, LIML, MR-Egger and MR-Median. As discussed in Section 3.3,

rather than use sisVIVE alone we use it in conjunction with valid-IV methods. We choose SPRS and IPRS to partner sisVIVE and respectively refer to the resulting estimates as sisVIVE-SPRS and sisVIVE-IPRS (noting that IPRS is the equivalent of 2SLS in these scenarios, and inverse-variance weighting is used for both MR-Egger and MR-Median). Using a two-sample strategy, we compare sisVIVE-SPRS and sisVIVE-2SLS with EPRS, 2SLS, MR-Egger and MR-Median under different levels of precision for $\tilde{b}_j$. The first precision level ("True") is the gold standard in which $\tilde{b}_j$ is taken to equal the true causal effect β_j; the second precision "Precise") with $\tilde{b}_j \sim N(\beta_j, 0.01^2)$ represents situations where the estimates are fairly accurate; and the third level ("Imprecise") with $\tilde{b}_j \sim N(\beta_j, 0.05^2)$ represents imprecise estimates from the second sample.

(Table 1 is here)

As expected, we found the valid-IV methods SPRS, IPRS, EPRS and LIML to perform at least as well as the robust methods in every scenario. Some numerical results from this study are presented in Table 1; we present only the gold standard two-sample results here, but the results for all three precision levels can be found in Table S2 in the Supplementary Information. Table 1 contains an assessment of each estimator's sampling distribution (bias, standard error and mean square error) and the performance of its normal-based confidence intervals (coverage and power). In addition, the first row of Table 2 (below) contains the average proportion of valid-IV SNPs falsely selected out (FSO) by sisVIVE from the set of valid IVs (see below for the definition of FSO).

In summary, we find, as expected, that the one-sample versions of the valid-IV estimators, except for SPRS and LIML, are affected by many weak instruments bias [12]. SPRS outperforms IPRS in terms of bias and overall in terms of MSE, despite IPRS being more efficient. Only SPRS and LIML offer close-to-nominal coverage and are the most powerful at detecting the true causal effects. The MSEs of MR-Egger and MR-Median, and particularly MR-Egger, are inferior to those of the valid-IV estimators by dint of having much larger standard errors; both also perform poorly in terms of coverage and power.

Using a two-sample strategy, every estimator is nearly unbiased when the true SNP-exposure association is known, or at least precisely estimated. On average, sisVIVE correctly selects 99 percent of the SNPs as valid IVs, which means the performance of sisVIVE-SPRS and sisVIVE-2SLS is respectively identical to that of SPRS and 2SLS. The inefficiency of both

MR-estimators considerably reduces the coverage and power of both. However, it can be seen in Table S2 (in the Supplementary Information) that every method is increasingly biased towards zero due as the precision of $\hat{b}_j$ decreases, due to failure of NOME, with a consequent impact on coverage and power. The bias of EPRS is less affected by NOME than the others, but its standard error is inflated and power very low.

4.2 Where some SNPs are invalid IVs

The outcomes are now generated under model (10) under three different pleiotropy scenarios:

1. Direct pleiotropy under which the InSIDE condition holds so that $\theta_j = 0$ for all $j = 1, \dots, J$, where the direct pleiotropy is balanced by generating $\sum_{s=1}^{S} \alpha_s = 0$, $\alpha_s \sim U(-0.2, 0.2)$, $s = 1, \dots, S$, where $S \leq J/2$ is the number of invalid SNPs, and $U(a, b)$ indicates the continuous uniform distribution on real interval (a, b). In this scenario, individual SNPs lead to direct pleiotropy but the full set of SNPs does not. The true SNP-exposure association equals the causal effect of the SNP: $b_j = \beta_j$.
2. As in Scenario 1 (including that $\theta_j = 0$ for all $j = 1, \dots, J$, and $b_j = \beta_j$) but the direct pleiotropy is now unbalanced with $\sum_{s=1}^{S} \alpha_s > 0$ and $\alpha_s \sim U(0, 0.2)$, $s = 1, \dots, S$ for the pleiotropic SNPs.
3. As in Scenario 2, except there is indirect pleiotropy with positive $\theta_s \sim U(0, 0.4)$ for pleiotropic SNP $s = 1, \dots, S$, and InSIDE fails because the true association between G_s and exposure is $b_s = \beta_s + \theta_s$ rather than β_s; for the valid-IV SNPs, $b_j = \beta_j$ as before.

Values of the outcome and exposure are respectively generated under model (10) and model (7) with true SNP-exposure association for invalid SNP G_j is $\beta_j + \theta_j$ for Scenario 3. The error terms of (10) and (7) are generated the same way as described in in Section 4.1 (see also Section S1 in the Supplementary Information).

For each of these scenarios, the performance of each estimator is assessed for $S = 10, 20$ and 30 invalid IVs, which corresponds respectively to 14, 28 and 42 percent of the SNPs; these scenarios all satisfy the requirement that less than 50 percent of the SNPs are invalid. Details of simulation design are given in Section S1 in the Supplementary information.

(Table 2 is here)

We first assess the performance of sisVIVE in terms of mean False Select In (FSI) and mean False Select Out (FSO). For each simulated data set, FSI is the ratio of number of pleiotropic SNPs which sisVIVE has not identified divided by total number of pleiotropic SNPs; and FSO is the ratio of valid SNPs which have been identified divided by the total number of valid SNPs. The results are shown in Table 2 are the mean of FSI and FSO cross 1000 simulation sets for all three pleiotropy scenarios defined above. When the number of pleiotropic SNPs is 30, sisVIVE incorrectly discards (on average) up to 71 percent of the valid-IV SNPs, but it performs well (no more than 5 percent of valid-IV SNPs incorrectly discarded) when there are only 10 invalid IVs. Continuing to focus on FSO, the performance of the two-sample version of sisVIVE is very similar to the one-sample version if only direct pleiotropic effects are present. It does slightly better than the one-sample version if the true SNP effects are known, but less well if there is indirect pleiotropy and InSIDE fails, even in the true precision scenario. Across all the presented scenarios, FSI is as high as 39 percent and as low as 11 percent; sisVIVE performs best in the Imprecise precision scenario with both types of pleiotropy and InSIDE fails, but this is offset by it performing worst in terms of FSO.

Table 3 contains the results for scenarios with 10 invalid-IV SNPs using one- and two-sample strategies. Estimates for SPRS, IPRS/EPRS, 2SLS, MR-Egger, MR-Median, sisVIVE-SPRS and sisVIVE-IPRS/EPRS/2SLS are presented under pleiotropy Scenarios 2 and 3. The two-sample results presented are, again, those for True precision scenario in which the true SNP-exposure associations are known. The full set of results across all three pleiotropy scenarios for 10, 20 and 30 invalid IVs are presented in Tables S3 to S11 in the Supplementary Information.

(Table 3 is here)

The picture that emerges from these results is that, across these scenarios, sisVIVE-SPRS has the best performance in terms of bias and MSE, closely followed by sisVIVE-IPRS/2SLS. However, every method is subject to major bias, and this is particularly true if a) a one-sample strategy is used (in any circumstances except for the least plausible pleiotropy scenario in which only direct balanced pleiotropy is present), and b) the third pleiotropy scenario where InSIDE has failed due to the presence of both types of pleiotropy. This

conclusion also holds if performance is judged in terms of coverage and power of the normal-based confidence intervals.

If InSIDE holds and only direct pleiotropy is present, MR-Egger performs best in terms of bias and confidence-interval coverage using a two-sample strategy, but even here it performs less well than sisVIVE-SPRS and sisVIVE-2SLS in terms of MSE; this is also the case in the Precise and Imprecise precision scenarios (see Tables S6-S8 in the Supplementary Information).

Overall, the same patterns are found for scenarios with 20 (see Tables S4, S7 and S10 in the Supplementary Information) and 30 (see Tables S5, S8 and S11) invalid-IV SNPs, except that the magnitude of the biases and MSEs, and the extent to which the confidence intervals fail to achieve nominal coverage, becomes worse as the number of invalid-IV SNPs increases.

The performance of sisVIVE can sometimes be explained in terms of its FSO and FSI: it tends to indicate a high proportion of the invalid-IV SNPs are valid IVs, and incorrectly lead us to discard a high proportion of valid IVs. This is also true when using two-sample strategies where we have knowledge of the true SNP-exposure associations.

The poor performance of MR-Median and sisVIVE in the third pleiotropy scenario is surprising because both methods, in theory, only require that less than half of the SNPs are invalid. We investigated this further by rerunning the simulations for the third pleiotropy scenario but, when using a two-sample strategy, taking the true causal effects β_j of the SNPs on the exposure to be known to us for the pleiotropic SNPs rather than $b_j = \beta_j + \theta_j$. The results (not presented) show the bias, MSE and coverage of the sisVIVE-based approaches are at a similar level to those for the scenarios in which there is only direct pleiotropy; the same is true for MR-Median. Taken together, these results indicate that indirect pleiotropy and the failure of InSIDE have a detrimental effect.

Proposition 2 of [22] contains conditions under which sisVIVE will not satisfy the 'irrepresentable condition' of [31] and hence cannot be consistent. These conditions roughly imply that sisVIVE will not be consistent if the strength of the invalid-IV SNPs is greater than those of the valid-IV ones. This is a possible explanation for the poor performance of sisVIVE because our scenario 3 simulations satisfy this condition. However, using their Corollary 1[22], we set up a new version of our third scenario in which InSIDE fails but

Proposition 2 is satisfied. The set up of the study is given in Section S2.2, and the results for FSI 10, 20 and 30 invalid-IV SNPs respectively presented in Tables S12-S14 in the Supplementary Information. It can be seen that the sisVIVE performs almost as badly as the original Scenario 3 so we conclude [31] failure of the irrepresentable condition does not explain our results.

We argue that this is because the consistency criterion underpinning the identification of MR-Median and sisVIVE (set out by Kang et al. (2016) [16] in their Theorem 1), is not satisfied if there is indirect pleiotropy and InSIDE fails. In short, identification is not possible because the effects of direct pleiotropy are generally confounded with the causal effects on the exposure of the invalid-IV SNPs, so the consistency criterion cannot hold. However, the moment condition is identified if there is no indirect confounding or we known β_j (the causal effect of SNP on exposure) which explains the improved performance of sisVIVE and MR-Median in our further simulations. We provide a more detailed argument in Section S2.3 in the Supplementary Information.

5. Analysis using UKHLS data

To estimate the causal effect of BMI on personal income using UKHLS, we use 71 out of the 97 common genetic variants identified as being associated with BMI at a genome-wide level of significance [11]. We excluded 16 of the 97 SNPs because the inclusion of these in original GWAS study [11] was based on secondary analyses involving only men, only women or non-Europeans (6 SNPs), or because the SNP had an imputation quality less than the 0.9 threshold (10 SNPs). The remaining 71 variants explain 1.6 percent of the variation in BMI among UKHLS participants; the effect alleles frequencies from both UKHLS and the GIANT study [11] are listed in Supplementary Information Table S1.

UKHLS is an annual household-based panel study which started collecting information about the social, economic and health status of its participants in 2009. The data set for our analysis is drawn from the General Population Sample (GPS) and the British Household Panel Survey (BHPS) arms of UKHLS; BHPS merged with UKHLS in 2010 at the start of UKHLS wave two. UKHLS collected additional health information, including BMI and blood samples, at wave two (for GPS) and wave three (for BHPS). In total, 10 480 individuals were genotyped on the Infinium Human Core Exome Beadchip. After quality-

control steps, a sample of 9944 individuals was obtained from which a further 1104 individuals were excluded based on the following criteria: genetic relatedness larger than 0.05 percent (707); BMI greater than 60 kg/m^2 (8); and aged under 25 (389). The number of cases with at least one wave of personal income and no missing on BMI, SNPs and covariates is N = 8047.

The outcome is the average annual personal income (API) taken over three consecutive waves starting from the wave at which the individual's health information was recorded. For individuals with one or two of these observations missing, we took the average API over the available waves. We standardised API separately for the GPS and BHPS samples because average API is higher in the BHPS than the GPS (possibly due to inflation). We also standardized BMI, and included the following baseline covariates $\mathbf{C}$: age (at which BMI was measured), gender, and the first 20 genetic principal components where genome-wide principal components function as ancestry markers [30]. Controlling for population stratification and focusing on the white population would go further to ensure that the core condition c) is not violated.

The reported causal effect estimates to follow are all conditional on these covariates. For a one-sample strategy, we estimate b_j and Γ_j from the UKHLS data controlling for $\mathbf{C}$. Following the guidelines for sisVIVE with covariates, we use the residual obtained from regressing API on $\mathbf{C}$ rather than raw API, and the residual obtained from regressing BMI on $\mathbf{C}$ rather than raw BMI. Following a two-sample strategy, the exposure-SNP estimate $\tilde{b}_j$ was taken from the GIANT study [11]. These estimates are also adjusted for $\mathbf{C}$ and so comparable with those from UKHLS if we assume the underlying populations of UKHLS and the GIANT consortium are similar.

(Table 4 is here)

It can be seen from Table 4 that BMI has a significant negative observational association with personal income. The one-sample Mendelian Randomization estimates are even larger negative numbers, but not significant: MR-Egger test finds no evidence to reject the null hypothesis that there is no pleiotropic effect ($\gamma_0^{\text{Egg}} = 0$); and sisVIVE indicates that all 71 SNPs are valid IVs, so sisVIVE-SPRS and sisVIVE-IPRS would respectively give the same estimate as SPRS and IPRS. Using a two-sample strategy, we must believe the population of external GIANT study is comparable to that of UKHLS. The estimated causal effects of BMI

on income are all negative, statistically significant and much larger in magnitude than the one-sample results. The two-sample Egger test again shows no significant average pleiotropy effect, and two-sample sisVIVE again indicates there are no invalid IVs. The underlying assumption that both samples are drawn from equivalent populations cannot be directly tested, but we can say that the effect-allele frequencies for UKHLS and GIANT are similar, which supports this assumption. As neither sisVIVE nor MR-Egger shows any evidence of pleiotropy, we should be able to trust EPRS which indicates that higher BMI cause the lower personal income.

Finally, we rerun the analysis with the additional inclusion of the 707 people originally excluded due to genetic relatedness. These people can be included if the possible correlation between genetically related people is accounted for using cluster-robust standard error estimation (e.g. using the cluster option in *Stata*). The results (not shown) do not significantly vary from those presented in Table 4.

6. Discussion

Our investigation has revealed that, while it outperforms the alternative approaches we considered, sisVIVE performed poorly if the InSIDE condition failed due to the presence of indirect pleiotropy. This is despite these scenarios satisfying the requirement for Kang et al.'s consistency criterion that more than 50 percent of the SNPs are valid instrumental variables. We argue that this is due to non-identification of leading to failure of the consistency criterion; this would explain the similarly poor performance of MR-Median which also depends on the consistency criterion. The conclusion from our data analysis, which indicates the true causal effect of body mass index could be under-estimated by a factor of five, is thus dependent on an assumption that there is no indirect pleiotropy through which the SNPs are related to the confounders. In general, such an occurrence cannot be ruled out because genes are determined at conception, and confounders determined at any point up to the time at which an individual's exposure is determined. The exception to this would be if there was strong scientific evidence that the phenotypical trait of each SNP was functionally related to the exposure of interest.

The performance of sisVIVE in scenarios where InSIDE holds and there is no indirect pleiotropy are more encouraging, but indicate that any Mendelian randomization study should

be based on a sensitivity analysis involving robust MR-Egger and 'valid IV' methods to capture whether differences between the estimates point to the presence of pleiotropic SNPs or even unobserved confounding. Our conclusions are also line with others who suggest using unweighted polygenic risk scores or a two-sample strategy [32-34]. In one-sample strategies, it appears that imprecision in the estimated weights of an internally weighted polygenic risk score, while improving efficiency, can lead to substantial bias. While simple polygenic risk scores perform well here, it is important to note that in further simulations we found (results not shown) severe bias was introduced if the effect-allele coding of the SNPs led to $\hat{\beta}_j$ (or $\tilde{\beta}_j$) being positive when true β_j would have led us to code it the other way. Such 'flip flopping' is possible [35] even in the absence of population stratification, but remains a potential source of bias for SPRS despite its being discounted elsewhere (e.g. [36], page 1883). The LIML or CUE estimators would, however, be unaffected by flip-flopping and, unlike 2SLS, not subject to large biases.

Acknowledgement: This work was supported by the Economic and Social Research Council (Grant Number: ES/M008592/1). YB and PC were also supported by the Economic and Social Research Council (ESRC) through Research Centre on Micro-Social Change (MiSoC) at the University of Essex, grant number ES/L009153/1. The UK Household Longitudinal Study is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council (Grant Number: ES/M008592/1). Data were collected by NatCen and the genome wide scan data were analysed by the Wellcome Trust Sanger Institute. Information on how to access the data can be found at https://www.understandingsociety.ac.uk/.

References

1. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Smith GD. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine.* 2008; **27**(8): 1133-1163.
2. Burgess S, Thompson SG. *Mendelian randomization— Methods for using genetic variants in causal estimation.* Chapman &Hall/CRC; 2014.
3. Tyrrell J, Jones SE, Beaumont R, Astley CM, Lovell R, Yaghootkar H, Tuke M, Ruth KS, Freathy RM, Hirschhorn JN, others. Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *BM.* 2016; **352**: i582.
4. Scholder SVK, Smith GD, Lawlor DA, Propper C, Windmeijer F. Child height, health and

human capital: Evidence using genetic markers. *European Economic Review*. 2013; **57**: 1-22.
5. Tillmann T, Vaucher J, Okbay A, Pikhart H, Peasey A, Kubinova R, Pajak A, Tamosiunas A, Malyutina S, Hartwig FP, others. Education and coronary heart disease: mendelian randomisation study. *BMJ*. 2017; **358**: j3542.
6. Didelez V, Sheehan N, Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*. 2007; **16**(4): 309-330.
7. Burgess S, Thompson SG, and C.C.G. Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*. 2011; **40**(3): 755-764.
8. Stock J, Wright J, and Yogo M. A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics.* 2002; **20**: 518–529.
9. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology.* 2013; **42**(4): 1134-1144.
10. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, 2015.
11. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, Powell C, Vedantam S. Buchkovich ML, Yang J, Croteau-Chonka DC, others. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015; **518**(7538): 197-206.
12. Davies NM, Scholder SVK, Earbmacher H, Burgess S, Windmeijer F, Smith GD. The many weak instruments problem and Mendelian randomization. *Statistics in Medicine*. 2015; **34**(3): 454-468.
13. Bowden J, Del Greco FM, Minelli C, Smith GD, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic. *International Journal of Epidemiology.* 2016; **45**(6): 1961–1974.
14. Bowden J, Smith GD, Burgess S, Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 2015; **44**(2): 512-525.
15. Bowden J, Smith GD, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology.* 2016; **40**(4): 304-314.
16. Kang H, Zhang A, Cai TT, Small DS. Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization. *Journal of the American Statistical Association*. 2016; **111**(513): 132-144.
17. Del Greco MF, Minelli C, Sheehanc NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in Medicine.* 2015; **34**(21): 2926-2940.
18. Wooldridge JM. *Econometric Analysis of Cross-sectional and Panel Data (first edition).* Cambridge, MA: MIT Press; 2002.
19. Burgess S, Bowden J. Integrating summarized data from multiple genetic variants in Mendelian randomization: bias and coverage properties of inverse-variance weighted methods. 2015; arXiv:1512.04486.
20. Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Societ.* 1982; **50**(4): 1029-1054.
21. Natarajan BK. Sparse approximate solutions to linear systems. *SIAM Journal on Computing.* 1995; **24**:227-234.
22. Windmeijer F, Farbmacher H, Davies N, Smith GD. On the use of the Lasso for instrumental variables estimation with some invalid instruments. 2017. Available: http://www.efm.bris.ac.uk/economics/working_papers/pdffiles/dp16674.pdf
23. Belloni A, Chen D, Chernozhukov V, Hansen C. Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica.* 2012; **80**(6):2369-2429.
24. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Annals of Statistics*. 2004; **32**(2): 407-451.

25. Pierce B, Burgess S. Efficient design for Mendelian randomization studies: subsample and two-sample instrumental variable estimators. *American Journal of Epidemiology*. 2013; **178**(7): 1177-1184.
26. Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association* 1995; **90**(430): 431-442.
27. Inoue A and Solon G. Two-sample instrumental variables estimators. *Review of Economics and Statistics*. 2010; **92**: 557-561.
28. Burgess S, Butterworth A, Thompson SG. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology*. 2013; **37**(7): 658-665.
29. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadjck NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006; **38**(8): 904-909.
30. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 2006. **101**(476): 1418-1429.
31. Hartwig F, Davies N. Why internal weights should be avoided (not only) in MR Egger regression. *International Journal of Epidemiology*. 2016; **45**(5): 1676-1678.
32. Bowden J, Burgess S, Smith GD, Response to Hartwig and Davies. *International Journal of Epidemiology*. 2016; **45**(5): 1679-1680.
33. Kemp FJ, Sayers S, Smith GD, Tobias JH, Evans DM. Authors' response to Hartwig and Davies. *International Journal of Epidemiology*. 2016; **45**(5): 1678-1679.
34. Lin PI, Vance JM, Pericak-Vance MA, Martin ER. No gene is an island: The flip-flop phenomenon. *American Journal of Human Genetics*. 2007; **80**(3): 531-538
35. Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in Medicine*. 2016; **35**(11): 1880-1906.

Methods	Mean (SD)	Mean SE	MSE	Coverage %	Power %
True value	-0.2	-	-	-	-
One sample methods					
-SPRS	-0.202 (0.104)	0.104	0.011	96.0	49.2
IPRS	-0.008 (0.068)	0.071	0.041	23.5	3.8
LIML	-0.206 (0.121)	0.096	0.015	89.0	57.9

Weighted Egger	0.054 (0.119)	0.122	0.079	43.7	7.3
Weighted Median	0.003 (0.090)	0.104	0.049	49.5	2.6
sisVIVE-SPRS	-0.202 (0.104)	0.104	0.011	95.9	48.8
sisVIVE-IPRS	-0.008 (0.068)	0.071	0.042	23.4	3.8
Two samples methods (True, $\widetilde{\boldsymbol{b}}_{\boldsymbol{j}} = \boldsymbol{\beta}_{\boldsymbol{j}}$)					
EPRS	-0.203 (0.095)	0.095	0.009	95.8	58.6
2SLS	-0.199 (0.085)	0.089	0.007	95.8	61.3
Weighted Egger	-0.208 (0.218)	0.220	0.047	95.5	16.5
Weighted Median	-0.200 (0.126)	0.142	0.016	97.4	26.2
sisVIVE-SPRS	-0.202 (0.104)	0.104	0.011	96.0	49.2
sisVIVE-2SLS	-0.199 (0.085)	0.089	0.007	95.8	61.3

Table 1 Simulation results for multiple instruments, all SNPs are valid, U~N(0,1), ε_{Xi} ~N(0,1), ε_{Yi} ~N(0,1), MC step=1000 and sample size=10,000

Table 2 Mean FSI% and FSO% of sisVIVE for 1000 simulation data sets

		One sample		**Two sample**					
				True: $\widetilde{\boldsymbol{b}}_{\boldsymbol{j}} = \boldsymbol{\beta}_{\boldsymbol{j}}$		**Precise: $\widetilde{\boldsymbol{b}}_{\boldsymbol{j}} \sim \boldsymbol{N}(\boldsymbol{\beta}_{\boldsymbol{j}}, \mathbf{0.01^2})$**		**Imprecise: $\widetilde{\boldsymbol{b}}_{\boldsymbol{j}} \sim \boldsymbol{N}(\boldsymbol{\beta}_{\boldsymbol{j}}, \mathbf{0.05^2})$**	
Scenarios	No. IVs Invalid	MFSI (%)	MFSO (%)	MFSI (%)	MFSO (%)	MFSI (%)	MFSO (%)	MFSI (%)	MFSO (%)

All IVs are valid	0	-	6.9	-	0	-	0	-	0.42
InSIDE holds and Direct Pleiotropy (balanced)	10	30.1	2.9	29.9	1.7	30.3	1.8	30.0	2.1
	20	20.6	9.2	21.2	6.8	21.3	6.7	20.8	7.9
	30	17.0	13.3	17.3	13.3	17.2	13.4	17.3	14.7
InSIDE holds and Direct Pleiotropy (positive)	10	36.1	1.9	30.5	2.4	31.0	2.2	31.6	2.2
	20	25.3	12.8	24.9	13.0	24.8	12.5	23.7	8.9
	30	21.2	31.9	23.3	32.3	23.0	30.4	21.0	18.6
InSIDE fails and Direct Pleiotropy (positive)	10	31.9	41.3	35.7	42.7	34.4	43.9	10.6	40.8
	20	39.2	42.1	32.5	54.2	32.6	54.2	23.1	66.3
	30	37.6	44.5	29.5	59.5	29.1	59.3	22.4	70.6

Table 3 Simulation results for multiple instruments when 10 SNPs are invalid, U~N(0,1), ε_{Xi} ~N(0,1), ε_{Yi} ~N(0,1), MC step=1000 and sample size=10,000

		10 pleiotropic SNPs (Positive direct pleiotropy, InSIDE holds)
True value	-0.2	

	Mean (SD)	Mean SE	MSE	Coverage %	Power %
One-sample strategy					
SPRS	0.342 (0.145)	0.092	0.315	9.0	86.4
IPRS	0.270 (0.124)	0.150	0.236	7.0	42.5
Weighted Egger	0.199 (0.258)	0.257	0.225	64.8	13.3
Weighted Median	0.076 (0.110)	0.114	0.088	30.6	9.2
sisVIVE-SPRS	-0.106 (0.115)	0.106	0.022	81.2	15.1
sisVIVE-IPRS	0.042 (0.083)	0.078	0.065	14.5	11.6
Two-sample strategy, True precision $\widetilde{b}_j = \beta_j$					
EPRS	0.251 (0.136)	0.084	0.222	2.0	75.4
2SLS	0.255 (0.139)	0.192	0.226	24.7	15.3
Weighted Egger	-0.214 (0.483)	0.473	0.233	93.7	6.4
Weighted Median	-0.099 (0.154)	0.152	0.034	90.3	10.6
sisVIVE-SPRS	-0.130 (0.118)	0.108	0.019	85.3	20.0
sisVIVE-2SLS	-0.128 (0.109)	0.096	0.017	84.2	29.5

10 pleiotropic SNPs
(Positive direct and indirect pleiotropy, InSIDE fails)

	Mean (SD)	Mean SE	MSE	Coverage %	Power %
One-sample strategy					
SPRS	0.574 (0.079)	0.047	0.606	0	100
IPRS	0.938 (0.078)	0.056	1.301	0	100
Weighted Egger	1.093 (0.097)	0.061	1.682	0	100
Weighted Median	0.998 (0.115)	0.065	1.449	0	100
sisVIVE-SPRS	0.517 (0.145)	0.115	0.535	2.3	91.5
sisVIVE-IPRS	0.761 (0.140)	0.058	0.944	0.1	99.4
Two-sample strategy, True precision $\widetilde{b}_j = \beta_j$					
EPRS	0.968 (0.080)	0.031	1.371	0	100
2SLS	0.970 (0.079)	0.057	1.376	0	100
Weighted Egger	1.149 (0.104)	0.057	1.830	0	100
Weighted Median	1.016 (0.116)	0.066	1.492	0	100
sisVIVE-SPRS	0.550 (0.149)	0.077	0.585	0.1	98.8
sisVIVE-2SLS	0.814 (0.133)	0.051	1.062	0	99.7

Table 4 Real data analysis, the association/causal effect estimates of BMI on personal income using n = 8047 Understand Society (UKHLS) data

	Estimate	Std. Error	P-value
Observational study	-0.032	0.011	0.002***
One-sample strategy			

SPRS	-0.112	0.082	0.173
IPRS	-0.049	0.066	0.458
2SLS	-0.048	0.064	0.449
LIML	-0.058	0.081	0.470
Weighted MR-Egger	-0.126	0.115	0.272
Egger test	0.003	0.004	0.411
Weighted MR-Median	-0.137	0.097	0.157
sisVIVE	-0.048	-	-
Number of invalid IVs detected by sisVIVE	0		
Two-sample strategy			
EPRS	-0.155	0.076	0.041**
2SLS	-0.154	0.071	0.030**
Weighted MR-Egger	-0.311	0.179	0.082*
Egger test	0.005	0.005	0.302
Weighted MR-Median	-0.292	0.106	0.006***
sisVIVE	-0.141	-	-
No. invalid IVs detected by sisVIVE	0		