Genetic Instrumental Variable (GIV) regression: Explaining socioeconomic and health outcomes in non-experimental data

Supporting Information

Thomas A. DiPrete, Casper A.P. Burik and Philipp D. Koellinger

April 11, 2018

Contents

www.pnas.org/cgi/doi/10.1073/pnas.1707388115

The Supporting Information (SI) for this article consists of six sections. In section 1, we provide technical details of the method for estimating narrow-sense SNP heritability from polygenic scores. In section 2, we discuss the details of GIV regression and why it provides more accurate estimates than OLS or MR for the case where measured SNPs have direct pleiotropic effects on the exposure and the outcome. In section 3, we discuss estimating the effects of an exposure in the presence of pleiotropy combined with other sources of endogeneity that are related to the observed genotypes (e.g. unobserved genetic variants, epistasis, genetic nurturing). Section 4 extends the possible sources of endogeneity further to cases that are unrelated to genetics (e.g. purely environmental unobserved confounds).

For each of these sections, we provide evidence from detailed simulations under a varying set of assumptions that cover a range of empirically-likely situations. Each of these simulations is generated at the level of individual SNPs. The SNP level simulations are used to generate data for the exposure and outcome variable in both the simulated GWAS samples and the replication sample and the simulated data are then used to estimate the parameters of interest using alternative methods.

Section 5 describes the data and methods used for our empirical examples, and we provide additional information about the empirical examples described in the article. The last section of the SI provides some practical guidelines for the usage of GIV regression.

1 Estimating narrow-sense SNP heritability from polygenic scores.

1.1 Technical details.

We begin by showing that consistent estimates of the chip heritability of a trait (i.e. the proportion of variance in a trait that is due to linear effects of currently measurable SNPs) can be obtained from polygenic scores. If y is the outcome variable, X is a vector of control variables including a constant, and $S^*_{y|X}$ is a summary measure of genetic tendency for y in the presence of controls for X, then one can write

$$y = X\beta + \gamma S^*_{y|X} + \epsilon \tag{1}$$

where, for example, y is educational attainment. Typical variables in X would be age, gender, and the first twenty principal components in the genetic data as controls for population structure. If the heritability of y is caused by a large number of genetic loci, each with a very small effect [1], we call y a "genetically complex trait." In this situation, the genetic liability for y cannot be adequately represented by just one gene. Rather, it is preferable to approximate the genetic liability $S^*_{y|X}$ with a polygenic score (PGS). The weights of each SNP that are summed up in the PGS are obtained from a GWAS on y in an independent sample [2, 3]. In a GWAS, y is regressed on each SNP separately, typically including a set of control variables such age, sex, and the first few principal components of the genetic data to control for population structure [4]. Thus, the obtained estimates for each SNP do not account for correlation between SNPs (a.k.a. linkage disequilibrium – LD), which may bias the PGS. In practice, several solutions are available to deal with this challenge, including pruning SNPs for LD prior to constructing the score [5] or using a method that explicitly takes the LD structure between SNPs into account (e.g. LDpred, see [6]). The scores themselves ($S_{y|X}$) are linear combination of the elements in G weighted by the estimated coefficients, $\hat{\zeta}_{y|X}$ obtained from

$$y = X\beta + G\hat{\zeta}_{y|X} + \varepsilon \tag{2}$$

where G is an $n \times m$ matrix of genetic markers, and $\hat{\zeta}_{y|X}$ is the $m \times 1$ vector of LD-adjusted estimated effect sizes, where the number of SNPs (the size of m in equation 2) is typically in the millions. If the true effects of each SNP on the outcome were known, the true genetic tendency $S^*_{y|X}$ would be expressed by the PGS for y, and the marginal R^2 of $S^*_{y|X}$ in equation 1 would be the chip heritability of the trait. In practice, GWAS results are obtained from finite sample sizes that only yield noisy estimates of the true effects of each SNP. Thus, a PGS constructed from GWAS results typically captures far less of the variation in y than suggested by the chip heritability of the trait ([2]; [7]; [8]). We refer to the estimate of the PGS from available GWAS data as $S_{y|X}$, and substitute $S_{y|X}$ for $S^*_{y|X}$ in equation 1. The variance of a trait that is captured by its available PGS increases with the available GWAS sample size to estimate ζ and converges to the true narrow-sense heritability of the trait at the limit if all relevant genetic markers were included in the GWAS and if the GWAS sample size were sufficiently large [8].

As reported in [2] and [9], the explained variance in a regression of a phenotype on its PGS can be expressed as

$$R^2_{y,S_y} = \frac{(n/m)h^4}{(n/m)h^2 + 1} \tag{3}$$

where y is standardized, σ^2_g is the genetic variance of y (i.e., the proportion of the variance in y explained by G), n is the sample size, and m is the number of genetic markers. For example, a PGS for EA based on a GWAS sample of 100,000 individuals would be expected to explain about 4% of the variance of EA in a hold-out sample (assuming there are 70,000 effective loci, all of them included in the GWAS, and a chip heritability of 20% [9]), even though the estimated total heritability of EA in family studies is roughly 40% [10].

It has long been understood that multiple indicators can, under certain conditions, provide a strategy to correct regression estimates for attenuation from measurement error ([11]; [12]). Instrumental variables (IV) regression using estimation strategies such as two stage least squares (2SLS) and limited information maximum likelihood (LIML) will provide a consistent estimate for the regression coefficient of a variable that is measured with error if certain assumptions are satisfied ([13]; [14]): (1) The IV is correlated with the problem regressor, and (2) conditional on the variables included in the regression, the IV does not directly cause the outcome variable, and it is not correlated with any of the unobserved variables that cause the outcome variable [13]. In general, these assumptions are difficult to satisfy. In the present case, however, GWAS summary statistics can be used in a way that comes close enough to meeting these conditions to measurably improve results obtainable from standard OLS regression and from standard Mendelian Randomization (MR) [15].

Multiple indicators of the PGS provide a theoretical solution to the problem of attenuation bias, and, we argue, a practical solution as well. The most straightforward solution to the problem is to split the GWAS discovery sample for y into two mutually exclusive subsamples. This produces noisier estimates of $S^*_{y|X}$, with lower predictive accuracy. However, it also produces an IV for $S_{y|X}$ that has desirable properties. Formally, we let $\hat{\zeta}_{y1|X}$ be the estimated coefficient vector for $\zeta_{y|X}$ in equation 2 from the first training sample, and $\hat{\zeta}_{y2|X}$ be the coefficient vector estimated from the second training sample. It follows then that

$$\begin{aligned}\hat{\zeta}_{y1j|X} &= \zeta_{yj|X} + u_{y1j|X}\\ \hat{\zeta}_{y2j|X} &= \zeta_{yj|X} + u_{y2j|X}\end{aligned}$$

for the j-th genetic marker, where $u_{y1|X}$ and $u_{y2|X}$ are asymptotically normally distributed errors with $E(u_{y1j|X}) = E(u_{y2j|X}) = 0$ and $V(u_{y1j|X}) = V(u_{y2j|X}) = \sigma^2_\varepsilon n^{-1}/var(x_j)$, and where x_j is the observed number of reference alleles for location j. In practice, the SNPs in $\hat{\zeta}_{y1|X}$ and $\hat{\zeta}_{y2|X}$ do not need to be exactly identical. Our derivations and results hold if the SNPs in both scores capture a sufficiently large amount of the SNP heritability of y, even if they are not the same SNPs. This is feasible because SNPs that are close to each on the same chromosome are often correlated with each other (a phenomenon referred to as linkage disequilibrium or LD), but the coefficient vectors $\hat{\zeta}_{y1|X}$ and $\hat{\zeta}_{y2|X}$ typically come from GWAS analyses that regress the outcome on one SNP at a time, ignoring the correlation structure between SNPs. Thus, neighboring SNPs that are correlated typically carry similar information about their contribution to y via $\hat{\zeta}_{y1|X}$ and $\hat{\zeta}_{y2|X}$ and can therefore we substituted with each other in the construction of the PGS.
Because the two discovery samples are non-overlapping, $u_{y1|X}$ and $u_{y2|X}$ would be independent of each other if the PGS model is correctly specified (we return to this point below). By applying the two vectors of estimated coefficients, we obtain two PGS,

$$\begin{aligned}S_{y1|X} &= S^*_{y|X} + v_1 = G\zeta_{y|X} + Gu_{y1|X} = S^*_{y|X} + Gu_{y1|X}\\ S_{y2|X} &= S^*_{y|X} + v_2 = G\zeta_{y|X} + Gu_{y2|X} = S^*_{y|X} + Gu_{y2|X}\end{aligned} \tag{4}$$

where G is the matrix of genetic markers for the analytical sample. We then rewrite equation

1 in terms of the observed first PGS as

$$\begin{aligned} y &= X\beta + \gamma S^{*}_{y|X} + \epsilon \\ &= X\beta + \gamma\left(S_{y1|X} - Gu_{y1|X}\right) + \epsilon \\ &= X\beta + \gamma S_{y1|X} + \left(\epsilon - Gu_{y1|X}\right) \end{aligned} \tag{5}$$

As can be seen from equation 5, the PGS $S_{y1|X}$ is correlated with the error term via its correlation with $Gu_{y1|X}$ from equation 4. However, under the assumptions that equation (2) accurately describes the relationship between G and y and that the genetic architecture of the trait is identical across GWAS and prediction samples, then $S_{y2|X}$ would meet the two requirements to be a valid instrument for $S_{y1|X}$, if it is correlated with $S_{y1|X}$ (through their mutual dependence on $S^{*}_{y|X}$) and if it is uncorrelated with the disturbance term. Clearly, the first requirement is met. Also, clearly $S^{*}_{y|X}$ ($= G\zeta_{y|X}$) is not correlated with $Gu_{y1|X}$. The remaining question, then is whether $Gu_{y2|X}$ is correlated with $Gu_{y1|X}$. The covariance of $Gu_{y1|X}$ and $Gu_{y2|X}$ is

$$\begin{aligned} Cov(Gu_{y1|X}, Gu_{y2|X}) &= E([Gu_{y1|X}][Gu_{y2|X}]) - (E([Gu_{y1|X}])(E[Gu_{y2|X}]) \\ &= E([Gu_{y1|X}][Gu_{y2|X}]) \end{aligned}$$

This follows because each term of $Gu_{y1|X}$ has the form $g_j u_j$ and the expectation of each of these terms is zero by virtue of the properties of OLS regression, namely that the residual has mean zero and is orthogonal to the regressors. Now,

$$\begin{aligned} E([Gu_{y1|X}][Gu_{y2|X}]) &= E\left\{\sum_{j=1}^{m} g_j^2 u_{y1j|X} u_{y2j|X} + \sum_{j=1}^{m}\sum_{k\neq j}^{m} g_j g_k u_{y1j|X} u_{y2k|X}\right\} \\ &= \sum_{j=1}^{m} E(g_j^2) E(u_{y1j|X} u_{y2j|X}) + \sum_{j=1}^{m}\sum_{k\neq j}^{m} E(g_j g_k) E(u_{y1j|X} u_{y2k|X}) \\ &= \sum_{j=1}^{m} E(g_j^2) E(u_{y1j|X}) E(u_{y2j|X}) + \sum_{j=1}^{m}\sum_{k\neq j}^{m} E(g_j g_k) E(u_{y1j|X}) E(u_{y2k|X}) \\ &= 0 \end{aligned} \tag{6}$$

where the third row follows because the coefficient errors for any given genetic marker from one sample will be independent of their value in a second independent sample. Now IV regression will be valid if the IV $S_{y2|X}$ is uncorrelated with the error term in equation 5, i.e., if

$$plim\,\frac{1}{n}\sum_i (S_{y2|X})_i(\epsilon_i - (Gu_{y1|X})_i) = plim\,\frac{1}{n}\sum_i (S^*_{y|X} + Gu_{y2|X})_i(\epsilon_i - (Gu_{y1|X})_i)$$

$$= plim\,\frac{1}{n}\sum_i \Big[(S^*_{y|X})_i\epsilon_i + (S^*_{y|X})_i(Gu_{y1|X})_i + (Gu_{y2|X})_i\epsilon_i + (Gu_{y1|X})_i(Gu_{y2|X})_i\Big]$$

$$= plim\,\frac{1}{n}\sum_i (Gu_{y1|X})_i(Gu_{y2|X})_i =$$

$$=0 \quad (7)$$

A complexity in the present situation is that the condition in equation 7 does not automatically follow from equation 6, because the correlation in the sample is computed on the given coefficient errors that were generated via the regressions in the two GWAS samples. This is readily appreciated if the number of markers was very small. If this number m equalled one, for example, then clearly the sample average of the square of each person's genetic marker multiplied by two given coefficient errors would not be zero even though the coefficient errors themselves were independent random draws from a distribution with mean zero.

However, as we show through SNP-level simulations below, this condition will generally hold for genetically complex traits that have been investigated in large-scale GWAS. In particular, assuming that all measured SNPs are causal and independent and their effect sizes are drawn from a normal distribution, we find that even when the GWAS sample is smaller than is the number of SNPs, IV estimation with $S_{y2|X}$ as the instrument for $S_{y1|X}$ does a very good job of recovering the true coefficient for S^*_y across a range of scenarios. In practice, SNPs are not independent because of linkage disequilibrium. However, there are more than 1,000,000 approximately independent loci (i.e. groups of SNPs that vary together) in 1000 Genomes imputed data that might potentially affect traits [16]. And even after stringent quality control and filtering of GWAS summary statistics, typically at least 200,000 LD-independent loci remain [17]. If only independent loci are used in the construction of the PGS, it is reasonable to assume that the independence assumption holds so long as the polygenic score is not dominated by a relatively small number of loci. If we then assume that genetic effects on y stem from both correlated and uncorrelated markers, the situation becomes only slightly more complicated. As mentioned above, the practical challenge is that the coefficient vector $\zeta_{y|X}$ typically comes from GWAS analyses that regress the outcome on one SNP at a time, ignoring the correlation structure between SNPs. The statistical dependence among SNPs in the construction of PGS is then dealt with in one of various ways. One obviously suboptimal solution is to ignore LD structure entirely and to construct the PGS using all available SNPs and their univariate coefficients. In practice, this naive solution often performs relatively well, although not as good as more sophisticated approaches. A second solution is to use LD-pruning. In this approach, only the most strongly associated SNP in each independent locus is used to construct the score, and the score consists of tens or even hundreds of thousands of approximately independent SNPs[18]. Finally, there are algorithm such as LDpred [6] that infer the LD-corrected, multivariate coefficients of each SNP from the original GWAS results taking all SNPs and their actual correlation structure into account. LDpred is the current best practice solution to construct PGS because it yields slightly better predictive performance than ignoring LD-structure or LD-pruning.

Our formal derivations until now assumed that the true coefficients of the genetic markers in G do not vary in the population. More generally, we might assume that the population consists of a finite number of (possibly latent) groups, $k = 1, ..., K$ with the kth group having the polygenic score $S^*_{yk|X}$. Absent information about the specific number of groups and the group memberships of individuals in any specific population, the polygenic score that would be estimated from a sufficiently large sample from that population would be a weighted average of the scores for each group, with the weights dependent on the proportion each group is of the total population [13]. Any population P therefore can be characterized in terms of its group composition, $p_1, p_2, ..., p_K$. The above results apply straightforwardly when the PGS are estimated and analyzed using samples from a single group. When they are instead estimated on a population that is a mixture of groups, the situation is more complicated. The true PGS for any individual who is in group k can be expressed as

$$S^*_{yk|X} = \bar{S}^*_{yP|X} + \Delta_{yk|X}$$

where $P = \{p_1, p_2, ..., p_K\}$ is the group composition that defines population P and $\Delta_{yk|X}$ is the deviation between the group k specific PGS for trait y and the population average (for population P). Under this elaboration, equation 5 can be written as

$$\begin{aligned}
y_{ik} &= X_i\beta + \gamma S^*_{yik|X} + \epsilon_i \\
&= X_i\beta + \gamma \left(\bar{S}^*_{yiP|X} + \Delta_{yik|X} \right) + \epsilon_i \\
&= X_i\beta + \gamma \bar{S}_{y1iP|X} + \left(\epsilon_i + \gamma \Delta_{yik|X} - \gamma v_{1i} \right)
\end{aligned}$$

where $S^*_{yik|X}$ is the true PGS for trait y for individual i in group k, and where $\bar{S}_{y1iP|X}$ is the first polygenic score estimated using coefficients from the GWAS sample drawn from population P. Variation in true PGS by group creates the possibility that the exclusion restriction will be violated. If $\bar{S}_{y2P|X}$ is the IV, then $\bar{S}_{y2P|X}$ is correlated with $\Delta_{yk|X}$ to the extent that the true PGS differ by group and to the extent that the weighted average deviation of the true PGS estimated from each individual's group and the true PGS estimated from the other groups correlates with the PGS for the population P. If the two PGS scores were estimated on one "pure" group and the analysis sample was for a second "pure" group, then the deviation between the two PGS would of course correlate with the PGS for one of the groups, and the exclusion restriction would be violated unless the SNP coefficients of the PGS for the one group were the same as the beta coefficients of the PGS for the other group. If the analysis sample and the GWAS samples are drawn from the same population (i.e., the same mixture of groups), we would expect the correlation between the deviations for analysis sample members (drawn from each of the groups in the same proportion as the GWAS sample) and the true PGS for the GWAS sample to be very small. If the population consists only of a single group or, equivalently, if all groups have the same SNP coefficients in their PGS for trait y, then the issue of group-specific heterogeneity in PGS disappears.[1]

When PGS for y are used that were constructed with a different set of control variables than are used in the regression, the above results need to be modified. Let us assume that variables χ were controlled in the GWAS and variables X are controlled in the regression model. Then

$$\begin{aligned} y &= X\beta + \gamma S^*_{y|\chi} + \{S^*_{y|X} - S^*_{y|\chi} + \varepsilon\} \\ &= X\beta + \gamma S^*_{y|\chi} + \{Gd_{yX\chi} + \varepsilon\} \end{aligned}$$

where $d_{yX\chi}$ is the vector of differences in the effects of genetic markers on y when X is controlled and when χ is controlled. If a finite sample PGS of y is constructed using χ as controls, i.e., $S_{y1|\chi}$, and this finite sample PGS is used in place of $S_{y1|X}$ as a proxy for $S^*_{y|X}$ in model 1, one obtains

$$y = X\beta + \gamma S_{y1|\chi} + \left(Gd_{yX\chi} - Gu_{y1|\chi} + \epsilon\right)$$

where

$$S_{y1|\chi} = S^*_{y|\chi} + Gd_{yX\chi} + Gu_{y1|\chi}$$

The problem now is that using $S_{y2|\chi}$ as an IV would violate the exclusion restriction to the extent that $d_{yX\chi}$ differs from zero, because $Gd_{yX\chi}$ is both in $S_{y2|\chi}$ and in the error, and because $S^*_{y|\chi}$ would generally be correlated with $Gd_{yX\chi}$. The extent of bias would depend on the extent to which the effects of the genetic markers on y differ when X and when χ are controlled.

Once a consistent estimate for $\hat{\gamma}$ has been obtained, it is possible to derive an estimate of the narrow-sense SNP (or chip) heritability of y. In a univariate linear regression model with standardized variables, the squared regression coefficient is equal to R^2. This follows directly from the definition of R^2 as the variance of y explained by X as a fraction of total variance of y. Thus, γ^2 in 1 can be thought of as the narrow-sense chip heritability of y if both y and $S^*_{y|X}$ are standardized variables with mean zero and a standard deviation of one (assuming the controls included in X are not correlated with genotype G). In practice,

[1] This issue is similar to the attenuation of predictive accuracy of a PGS that results from an imperfect genetic correlation between the GWAS summary statistics in the hold-out sample and the GWAS summary statistics in the discovery sample [19].

however, the estimate $\hat{\gamma^2}$ originates from a regression on a PGS that contain measurement error ($S_{y1|X}$ or $S_{y2|X}$) rather than on the true PGS $S^*_{y|X}$. In particular, the obtained regression coefficient $\hat{\gamma^2}$ will be standardized using the variance of $S_{y1|X}$ or $S_{y2|X}$ instead of the variance of $S^*_{y|X}$. It turns out that this implies that the heritability estimate $\hat{\gamma^2}$ is biased by a factor equal to $var(S_{y|X})/var(S^*_{y|X})$, which simplifies to $1/var(S^*_{y|X})$ if the observed score was standardized.[2] However, it is possible to derive a simple error correction because one can estimate the variance of $S^*_{y|X}$ by estimating the covariance of $S_{y1|X}$ and $S_{y2|X}$:

$$cov(S_{y1|X}, S_{y2|X}) = cov(S^*_{y|X} + e_{y1}, S^*_{y|X} + e_{y2}) = \rho(S_{y1|X}, S_{y2|X}) = var(S^*_{y|X}).$$

With an estimate of $var(S^*_{y|X})$ at hand, we can back out an unbiased heritability estimate:

$$h^2_y = \hat{\gamma^2} var(S^*_{y|X})/var(y).$$

When y is standardized, $var(y) = 1$, the error correction simplifies to

$$h^2_y = \hat{\gamma^2} \rho(S_{y1|X}, S_{y2|X}).$$

An estimate of the standard error of h^2_y can be obtained using the Delta method [20].[3]

1.2 Simulations

Our first set of simulations are based on the following model for y:

$$y = \gamma_1 + \gamma_2 S^*(y) + \epsilon$$

We generate S^* using varying numbers of independently drawn genetic markers from 1,000 to 300,000 − up to the memory limits of our processor nodes (512 GB) − with a minor allele frequency of 0.5 and coefficients for these genetic markers.[4] The constant γ_1 is set to zero and the coefficients for the genetic markers are drawn from a normal distribution. We also draw ϵ from a normal distribution. The variance of the distributions for ϵ and the coefficients of the genetic markers are set such that the heritability is correct and the variance of y is equal to 1 (i.e. y is standardized). We use this data generating process to produce two independent samples, which together constitute the GWAS sample. We specify varying sizes of the total GWAS sample from 50,000 to 500,000 observations. We generate these data under three different assumptions about the SNP heritability of y, namely that h^2 is alternately set to 0.1, 0.3, and 0.5. We then use the two independent GWAS samples to estimate the effect of each marker twice, using bivariate regressions of y on each of the

[2] We thank Elliot Tucker-Drob for pointing this out to us.

[3] See [21] for an alternative correction method.

[4] Assuming a MAF of 0.5 for all markers is unlikely to affect our results beyond statistical power.

individual markers. In a third independent sample ($N = 10,000$) we construct the PGS for y, which we designate as $S(y_1)$ and $S(y_2)$, using the two GWAS estimates.

We then estimate the effect $\hat{\gamma_2}$ of the PGS for y on y. We do this using an IV regression with $S(y_2)$ as the IV for $S(y_1)$. In other words, we use OLS to estimate the second stage model

$$y = \gamma_1 + \gamma_2 \hat{S}(y_1) + \epsilon$$

where the predicted value of $S(y_1)$ is obtained via estimates from a first stage regression of $S(y_1)$ on $S(y_2)$, i.e.,

$$\hat{S}(y_1) = \hat{\beta}_1 + \hat{\beta}_2 S(y_2).$$

The standardized coefficient estimate $\hat{\gamma_2}$ from the second stage regression is used to obtain an estimate for h^2 via the equation

$$\hat{h}^2 = \hat{\gamma_2}^2 (corr(S(y_1)S(y_2))).$$

Table 1 shows the results of these simulations where we do 20 simulations for each condition and report average results in the table. Panel (a) presents simulations where the SNP heritability is set to 0.1. As can be seen in panel (a), the estimated heritability is very close to the true heritability so long as the GWAS sample size is as large or larger than the number of SNPs that are included in the computation of the PGS. Also, for all simulations where the GWAS sample exceeds the number of SNPs, the standard errors are small relative to the estimate. In panel (b), we simulate using a heritability of 0.3 and we obtain an accurate estimate with a relatively small standard error when the GWAS sample is as large or larger than the number of SNPs. The same result is obtained when the data are generated with a heritability of 0.5. Generally speaking, we observe that the sample size needed for an accurate estimate of heritability has an inverse relationship with the size of the heritability. Thus, 50,000 cases is not sufficient to estimate heritability precisely when the true SNP heritability is 0.1 and the number of SNPs is 100,000, and 100,000 cases produces an accurate estimate but a fairly large standard error. The precision of the estimates increases considerably for both of these cases, however, when the true SNP heritability is 0.3, and even 50,000 cases is sufficient to produce a precise and accurate estimate of heritability when the true SNP heritability is 0.5 and the number of SNPs is 100,000 or fewer. As mentioned above, most practical applications will be based on more than 100,000 independent SNPs, although many of them may actually have a true effect of zero. Hence, the remaining causal loci for y will tend to have slightly larger true effects than we simulated here under the assumption that all SNPs are causal. Slightly larger SNP effects imply better statistical power in GWAS analyses and a more favorable ratio of estimated effect sizes to their standard errors. Thus, our simulation results are likely to be conservative lower bounds for the accuracy that our method can achieve for estimating heritability in real data.

Table 1: Estimating the SNP heritability of y

	Number of SNPs	Total GWAS sample size 50,000	100,000	300,000	500,000
$h^2 = 0.1$	1000	0.1002	0.1008	0.0999	0.1016
		(0.00725)	(0.00666)	(0.00621)	(0.00618)
	10000	0.1039	0.0969	0.0988	0.0995
		(0.0165)	(0.0112)	(0.0080)	(0.0073)
	100000	0.1247	0.0964	0.0972	x
		(0.1343)	(0.0475)	(0.0215)	
	300000	0.1822	0.09197*	x	x
		(8.3243)	(0.1512)		
$h^2 = 0.3$	1000	0.2937	0.2968	0.2961	0.2952
		(0.0100)	(0.0095)	(0.0093)	(0.0092)
	10000	0.2999	0.3016	0.2976	0.2982
		(0.0175)	(0.0134)	(0.0106)	(0.0100)
	100000	0.2873	0.3087	0.2999	x
		(0.0889)	(0.0522)	(0.0232)	
	300000	0.2811	0.3558	x	x
		(0.2586)	(0.1713)		
$h^2 = 0.5$	1000	0.4969	0.4951	0.5001	0.5103
		(0.0108)	(0.0103)	(0.0101)	(0.0102)
	10000	0.5039	0.4991	0.5008	0.4974
		(0.0181)	(0.0140)	(0.0114)	(0.0107)
	100000	0.5167	0.5110	0.5024	x
		(0.0988)	(0.0519)	(0.0234)	
	300000	0.6080	0.5460	x	x
		(0.4702)	(0.1536)		

Mean of heritability estimates of twenty simulations for several GWAS sample sizes, varying the number of SNPs and the heritability (h^2) of y. Standard errors (in parentheses) are calculated via the delta method. The size of the replication sample is 10,000.

* Mean of nineteen simulations, due to one extreme outlier.

x Unable to simulate due to memory constraints on the high memory nodes of the high performance computer.

2 Reducing bias due to direct pleiotropic effects on exposure and outcome

We next address situations where the question of interest is not the SNP heritability of y per se, but rather the influence of some non-randomized exposure T on y (e.g. a behavioral or environmental variable, or a non-randomized treatment due to policy or medical interventions). We rewrite equation 1 such that

$$
\begin{aligned}
y &= \delta T + X\beta_y + \gamma S^*_{y|XT} + \epsilon_y \\
&= \delta T + X\beta_y + G\zeta_{y|XT} + \epsilon_y
\end{aligned} \tag{8}
$$

where

$$
\begin{aligned}
T &= \alpha S^*_{T|X} + X\beta_T + \epsilon_T \\
&= G\xi_{T|X} + X\beta_T + \epsilon_T
\end{aligned} \tag{9}
$$

where, for example, y could be educational attainment and T could be body height and where we assume that the disturbance term is uncorrelated with genetic variables. We drop the subscript on the coefficients on the exogenous control variables X below when it would not lead to confusion. In each case, it is presumed that the outcome variable is to some extent caused by genetic factors, and the concern is that the genetic propensity for the outcome variable is also correlated with the treatment represented by T in equation (8). We now use $S^*_{y|XT}$ rather than $S^*_{y|X}$ in the equation, where $S^*_{y|XT}$ is the linear combination of the effects of SNPs on y when T is controlled. Given that T is in the model, the effect of individual SNPs on y will generally involve a direct effect net of T (ζ) and an indirect effect stemming from the combination of their effect on T (ξ) and the effect of T on y.

Adding the true conditional score ($S^*_{T|X}$) as a control variable to a regression of Y on T would eliminate bias arising from direct pleiotropy. Pleiotropy leads to omitted variable bias from the failure to control for the (possibly tens of thousands of) individual SNPs in the structural model that influence both Y and T directly. So imagine a model that contained tens of thousands of variables for the SNPs and a sample large enough and computers capable enough of estimating the coefficients of this model using OLS. Aside from the enormous number of regressors, this is a standard regression problem. Under standard conditions and if uncontrolled pleiotropy is the only source of bias, then the the coefficients of the SNPs converge in probability to the true coefficients while the coefficient on T converges in probability to its true value. In other words, the sum of the SNPs multiplied by their coefficients converges in probability to the true conditional PGS $S^*_{y|XT}$. Thus, controlling for tens of thousands of SNP variables in the structural model becomes closer and closer to controlling for the true conditional PGS in sufficiently large samples. Of course, in sample sizes that are currently available or that might be available in the foreseeable future, we are very far from being able to use direct controls for the individual SNPs in order to address the pleiotropy problem. However, proxies for $S^*_{y|XT}$ are already available.

If the true $S^*_{y|XT}$ is not observed and cannot be explicitly controlled in equation 8, it is part of the disturbance term. If so-called Type 1 pleiotropy is present [15], then T itself is a function of the same genetic markers that have other effects on y that do not operate through T, and the coefficients of these markers on T (ξ), which represent the indirect effects of the markers on y that operate through T, are correlated with the direct effects of the markers on y (ζ) when T is controlled in equation 8. Because of the correlation between T and $S^*_{y|XT}$ (which along with ϵ_y is in the disturbance term in equation 8), $\hat{\delta}$ will be a biased estimate of the effects of T.

While the true $S^*_{y|XT}$ is unknown, we may be able to obtain a proxy $S_{y|XT}$ for it from GWAS in finite sample sizes. While it is not guaranteed, the general conclusion of the literature is that the use of proxy variables such as $S_{y|XT}$ is an improvement over omitting the variable being proxied [22, 23]. However, if the proxy is measured with error, some bias will remain. More specifically, if $S_{y|XT}$ is used instead of $S^*_{y|XT}$ in equation 8, we get

$$S_{y|XT} = S^*_{y|XT} + Gu_{y|XT} = S^*_{y|XT} + v$$

which yields

$$y = \delta T + X\beta + \gamma S_{y|XT} + (\epsilon_y - \gamma v) \tag{10}$$

The problem now is that $S_{y|XT}$ is constructed from a large number of regressions of one genetic marker at a time along with the control variable T. The presence of T in the GWAS regressions for y produces estimated coefficients for the markers, G, that are functions of T. The error in the conditional PGS for y is a function of the coefficient estimates for the individual genetic markers and therefore is correlated with T. The estimation error plus pleiotropy produces a correlation between T and v, which still induces bias in OLS estimates of δ. Thus, while the use of a proxy control such as $S_{y|XT}$ will generally reduce bias in the estimated effect of δ, some bias will remain as long as $S_{y|XT}$ is measured with error.

Because of its relevance later on, we also note that the problem cannot be solved by constructing a PGS for y that is unconditional on T, i.e., $S_{y|X}$. The use of $S_{y|X}$ instead of $S_{y|XT}$ produces an over-control, where an estimate of the total effect of each genetic marker is controlled instead of just the direct effect. This over-control would produce a severe downward bias in the estimate of δ. To see this, imagine that there are only two genetic markers, g_1 and g_2, where

$$T = \xi_1 g_1 + \xi_2 g_2 + e_T$$

and therefore, where

$$\begin{aligned} y &= \delta(\xi_1 g_1 + \xi_2 g_2 + e_T) + \zeta_1 g_1 + \zeta_2 g_2 + e_y \\ &= (\delta\xi_1 + \zeta_1)g_1 + (\delta\xi + \zeta_2)g_2 + \delta e_T + e_y \\ &= (0S^*_T + \delta e_T) + \gamma S^*_y + e_y \end{aligned} \tag{11}$$

Note that the effects $\delta\xi + \zeta_k$ represent the total effect of g_k on y and provide the reduced form for the structural model in equation 8 if T is omitted from the model. As can be seen in equation 11, a control for the true unconditional PGS for y (S_y^*) would be expected to produce an estimated effect of T that biased towards zero. Substituting the proxy S_y for S_y^* would not eliminate the downward bias entirely.

In standard MR, a measure of genetic tendency ($S_{T|X}$) for a behavior of interest (T in equation 8) is used as an IV in an effort to purge $\hat{\delta}$ of bias that arises from correlation between T and unobservable variables in the disturbance term under the argument that the genetic tendency variable, e.g., the measured PGS $S_{T|X}$, is exogenous ([14]; [24]). This approach would generally be successful if the endogeneity in the error term is from non-genetic sources and, consequently, if the genetic information in the IV for T is uncorrelated with the error term. In the absence of pleiotropy and other forms of genetic endogeneity (e.g., genetic nurturing), MR should be an effective strategy if the IV is strong enough to provide reasonable precision in the estimator.

However, MR becomes problematic when genetic variables in the error term are affecting y net of T while at the same time are correlated with the genetic variables in the equation for T, in other words, when ξ is correlated with ζ in equations 8 and 9. An example of this situation would be the use of a PGS for height as an instrument for height in a regression of the effect of height on educational attainment. The second stage regression in MR, then, takes the form

$$y = \delta\hat{T} + X\beta + \{\epsilon_y + \gamma S^*_{y|XT} + \delta(T - \hat{T})\} \tag{12}$$

The problem with this approach is that the PGS for height will typically fail to satisfy the exclusion restriction because of pleiotropy: the genetic variants that predispose individuals to be tall may also directly increase the predisposition for higher educational attainment [25, 26] (e.g. via healthy cell growth and metabolism). Because ξ is correlated with ζ, $S^*_{T|X}$ is correlated with $S^*_{y|XT}$. This problem is not solved even if we could use the true PGS S_T^* as the IV, because the genetic effects in S_T^* are correlated with the genetic effects in $S^*_{y|XT}$. Whether the endogeneity bias from pleiotropy is big enough to offset MR's potential advantages for addressing the endogeneity from non-genetic sources is an empirical question that depends on the specific situation. The problems that pleiotropy creates for MR could be solved if the true genetic propensity for y, net of T could be directly controlled in the regression.[5] Unfortunately, this is not possible because the best we can do is to use $S_{y|XT}$ as a proxy for $S^*_{y|XT}$.

Endogeneity bias that stems purely from non-genetic sources can sometimes be addressed through the use of non-genetic IVs that are available from randomized clinical trials (RCTs) or from natural experiments. It can also sometimes be addressed through the

[5]Note that this approach would not solve problems caused by other sources of genetic endogeneity such as environmental effects in the error term that were correlated with parental genes, which themselves are correlated with the genetic information in $S_{T|X}$.

use of fixed effects strategies when data on siblings or dizygotic twins is available. It can also often be reduced by controlling for observable variables that affect both the "assignment" to T and also the outcome variable. However, in the absence of data from RCTs, endogeneity bias from genetic sources is a difficult problem, and certainly one that MR does not directly address. We therefore first discuss the cases where all the endogeneity bias is from genetic sources, whether pleiotropy alone or pleiotropy in combination with other genetic confounds. We subsequently address the implications of endogeneity bias that emerge from both genetic and non-genetic sources.

2.1 Reducing bias from pleiotropy.

First, we assume that the *only* source of endogeneity in equation 8 is pleiotropy, and we examine the performance of a set of estimators intended to reduce its impact. The first strategy is to reduce the correlation between the instrument S_T in MR and the error by controlling for a proxy of $S^*_{y|XT}$, namely $S_{y1|XT}$. We refer to the combined use of $S_{y|XT}$ as a control and $S_{T|X}$ as an IV as "enhanced Mendelian Randomization" (EMR). Controlling for $S_{y|XT}$ as a proxy for $S^*_{y|XT}$ is not fully adequate because the error in $S_{y|XT}$ (i.e., $S^*_{y|XT} - S_{y|XT}$) is correlated with $S^*_{T|X}$. As noted previously, the use of a proxy control should improve the quality of the estimate for δ. However, as we show below, the pleiotropy bias at levels that would be expected to occur for real-world applications creates serious problems for MR as an effective strategy for obtaining accurate estimates of δ even if $S_{y|XT}$ is included as a control variable.

The second set of strategies drop the use of S_T as an instrument because it is not a valid instrument in the presence of pleiotropy and is not of practical utility in this situation either, as we show below. Instead, we start with the well-known formula for endogeneity bias for a generic OLS with dependent variable y, included covariates X, and coefficients of these covariates contained in the vector β, namely

$$\begin{aligned}\hat{\beta} &= \left(X'X\right)^{-1} X'y \\ &= \left(X'X\right)^{-1} X' \left(X\beta + \epsilon\right) \\ &= \beta + \left(X'X\right)^{-1} X'\epsilon\end{aligned}$$

So

$$E[\hat{\beta}|X] = \beta + E[\left(X'X\right)^{-1} X'\epsilon|X] \tag{13}$$

In other words, the coefficient bias from OLS is the expected regression coefficient of the error on the included variables in the regression. If ϵ is the sum of an omitted variable that we can label as z, which is correlated with the regressors and additional variables that are uncorrelated with the regressors, then the bias for each coefficient β_k in the vector β in equation 13 becomes the product of the regression coefficient for x_k in the regression of z on all the omitted variables multiplied by the effect of z on the outcome.

For simplicity, we assume that the only variables in the regression are T and a potential proxy for $S^*_{y|XT}$, which we call $S_{y|XT}$. For any given proxy, $S_{y|XT}$, the bias in the estimate of δ (the coefficient for T in equation 8) comes from the expected coefficient on T in the regression of $\gamma S^*_{y|XT} - \tilde{\gamma} S_{y|XT}$ on T and $S_{y|XT}$. We consider three alternatives as proxies for $S^*_{y|XT}$. First, we use $S_{y|XT}$, i.e. the conditional PGS for y from the full GWAS sample, in a simple OLS regression.[6] Second, we attempt to adjust for measurement error in $S_{y|XT}$ by constructing the predicted conditional PGS for y, called $\hat{S}_{y1|XT}$,by using $S_{y2|X}$ (the unconditional PGS for y) as an IV for $S_{y1|XT}$ where GWAS coefficients for $\hat{S}_{y1|XT}$ and $S_{y2|X}$ are obtained from non-overlapping GWAS samples of the same population .We call this approach, where $\hat{S}_{y1|XT}$ (the predicted conditional PGS for y) is used as the regressor in the second stage, "conditional GIV regression" (GIV-C).

We generally expect the use of GIV-C to perform better than the use of the proxy $S_{y|XT}$ in simple OLS. Recall that where the true effect of T on y is positive and in the presence of positive pleiotropy, the estimated effect of T on y will have positive bias. This follows from the positive correlation between $S^*_{y|XT}$ and T and from the positive effect of $S^*_{y|XT}$ (which is a component of the error) on y. The presence of the proxy $S_{y|XT}$ with simple OLS adds a partially offsetting negative bias, because the correlation between $S_{y|XT}$ and T is positive and the effect $\tilde{\gamma}_{OLS}$ is also positive, but $\hat{\gamma}_{OLS} S_{y1|XT}$ is being subtracted, which causes the offsetting bias to be negative. The net bias is expected to be smaller with the inclusion of the proxy than with no proxy at all, but we still expect it to be positive both because the correlation between T and $S_{y|XT}$ would be lower than between T and $S^*_{y|XT}$, and because we expect $\hat{\gamma}_{OLS}$ to be attenuated relative to γ.

When GIV-C is used instead of simple OLS, the term in the error becomes

$$\gamma S^*_{y|XT} - \hat{\gamma}_{IVC} \hat{S}_{y1|XT} \tag{14}$$

The presence of T as a regressor in the first and second stages of GIV-C, which is correlated with $S^*_{y|XT}$, prevents the IV strategy from obtaining a consistent estimate of γ. Nonetheless, we would generally expect $\hat{\gamma}_{IVc} > \hat{\gamma}_{OLS}$, and therefore we expect the positive bias for the estimate of δ to be smaller when using GIV-C than when estimating δ using simple OLS with the proxy $S_{y|XT}$.

We also employ a third estimator that substitutes the unconditional PGS for y (i.e., substitutes $S_{y1|X}$ for the conditional PGS $S_{y1|XT}$) as the proxy control in the structural model in equation 8. We then use $S_{y2|X}$ (the same IV as with GIV-C) to predict $S_{y1|X}$, obtaining $\hat{S}_{y1|X}$ as the regressor in the second stage. We call this third approach "unconditional GIV regression" (GIV-U). With GIV-U, the problem term in the error is

$$\gamma S^*_{y|XT} - \hat{\gamma}_{IVU} \hat{S}_{y1|X} \tag{15}$$

[6] We also estimate versions of this model using only the first half of the GWAS sample to be able to compare results across methods while holding GWAS sample size constant. We call the resulting score $S_{y1|XT}$ and we present both sets of estimates in our simulation results.

As before, the presence of the first term produces a positive bias in the estimate of δ, while the second term produces an offsetting negative bias. The offset will be stronger when $S_{y1|X}$ is used as the covariate in the structural model than when $S_{y1|XT}$ is used because the coefficients of the genetic markers in $S_{y1|X}$ are $\hat{\delta}\hat{\xi}+\hat{\zeta}$, where ξ is the effect of the genetic marker on T. The presence of $\hat{\delta}G\hat{\xi}$ in the second term in the error produces a stronger downward bias. This downward bias is made still stronger by the use of $\hat{\gamma}_{IVU}$ instead of $\hat{\gamma}_{OLS}$ as the coefficient, because we expect the first stage regression to reduce the downward bias of $\hat{\gamma}_{OLS}$.

To summarize, we expect these three proxies to behave differently in the simulations, and, as we will see, this expectation is met in practice. It turns out to be the case that GIV-C and GIV-U provide upper and lower bounds for the effect of T across a range of plausible scenarios for pleiotropy and for heritability.

2.2 Evidence from simulations

To address the utility of these estimators, we conducted a set of simulations. We first discuss simulations under various assumptions about endogeneity and heritability for the case where the data generation model includes an effect of T on y. After discussing each of the relevant scenarios, we will then revisit each of these scenarios and examine the performance of the alternative estimators using a data generation process in which there is no effect of T on y.

We simulated data for two independent GWAS samples and for an independent prediction sample. The data generating process for the pleiotropy analysis is as follows:

$$T = \alpha S_T^* + \epsilon_T, \qquad \epsilon_T \sim \mathcal{N}(0,\sigma_\eta^2)$$
$$y = \gamma S_{y|T}^* + \delta T + \epsilon_y, \qquad \epsilon_y \sim \mathcal{N}(0,\sigma_\epsilon^2)$$

$S_{y|T}^*$ and S_T^* were constructed from the simulation of 10,000 independent genetic markers and coefficients for these genetic markers. The coefficients for these markers are drawn from a joint multivariate normal distribution, where the correlation between the zeta's for $S_{y|T}^*$ and the xi's for S_T^* (see equations 8 and 9) is varied in order to simulate varying degrees of pleiotropy (this genetic correlation is labelled as ρ in Tables 2-14). Each simulation is based on a GWAS combined sample of 100,000 with 10,000 SNPs, a third independent prediction sample of 10,000 and twenty repetitions. We use these values because they are large enough to reveal the essential properties of the estimators under the alternative conditions considered in the tables.

T is standardized and has a mean of 0 and a variance of 1. The variance of the true polygenic scores (S_T^* and $S_{y|T}^*$) are simulated to match the heritability of the two traits. Furthermore, y is standardized when T has no causal effect ($\delta = 0$) and in absence of pleiotropy. When δ is 1 and in the absence of pleiotropy, the variance of y is equal to 2. When pleiotropy increases, the scale of the coefficients of the markers is kept constant to minimize the parameter changes across simulations and thus the variance of y increases. See table 16 for a list of parameters with the matching variance and heritability of y, and the standardized effect size of T. We do not include any X variables in the simulations, because they are not needed in order to analyze the essential issues.

In these simulations, we vary the amount of genetic correlation as well as the heritability for both y and T.[7] In the tables below, we report average coefficient estimates and standard errors across 20 repetitions for each model. We also limit the simulations below to the case of positive pleiotropy. In practice, this corresponds to a state of knowledge where the analyst either knows the sign of the pleiotropy correlation, or knows that it is weak but is uncertain whether it is weak positive or weak negative.

Tables 2, 3, 4, and 5 show the results of this set of simulations where we vary both the extent of heritability (h^2) for T and for y and the strength of the genetic correlation (ρ) between the effects of SNPs on T and their effect on y, net of T (i.e., between ζ and ξ). Each of these columns has four panels across the columns:

- The first panel reports the OLS estimates of y on T with no additional controls.
- The second panel reports estimates that are based on MR, i.e., that use S_T as an instrument for T. The column labelled MR only includes T as the regressor and uses S_T as the as the IV. Column EMR-1 is a version of enhanced MR that uses S_T as an IV along with a control for S_{y1} in an effort to reduce pleiotropy. Column EMR-2 uses $S_{y1|T}$ as the control and uses two IVs, namely S_T and S_{y2}.
- The third panel, which is labelled "Conditional Proxy PGS", uses versions of the conditional PGS for y as a proxy control for $S^*_{y|T}$. The column labelled as "OLS S(y|T)" uses the conditional PGS for y from the entire GWAS sample as the proxy control. The column labelled as "OLS S(y1|T)" uses the conditional PGS for y from the first half of the split GWAS sample as the proxy control. The column labelled as GIV-C uses the conditional PGS for y from the first GWAS sample as the proxy control but it uses S_{y2} as the IV. S_T is not used as an IV in any of these models.
- The fourth panel, which is labelled "Unconditional Proxy PGS," uses versions of the unconditional PGS for y as a proxy control. The first column (OLS S(y)) uses the unconditional PGS from the full GWAS sample (S_y) as the proxy control. The second column, which is labelled as "OLS S(y1)" uses the unconditional PGS for y from the first half of the split GWAS sample (S_{y1}) as the proxy control. The column labelled as GIV-U uses the unconditional PGS for y from the first GWAS sample (S_{y1}) as the proxy control but it uses S_{y2} as its IV. S_T is not used as an IV in any of these models.

[7] There is a logical relationship between the level of heritability for T and y , the strength of the correlation between the effects of genetic markers on T and on y, the size of the effect of T on y in the structural model for y, and the error variance in the equations for T and y. These logical relationships make some combinations of heritability and genetic correlation impossible, but we explore a wide range of the possible values in the simulations below.

Each of the tables has six sets of rows. The first three panels down the rows present simulations where the true effect of T on y is 1.0. These rows show the ability of the various estimators to recover an accurate estimate of T when T actually has an effect on y. The second three panels present simulations where T is specified to have no effect on y. It is worth pointing out that T can be correlated with y (e.g., via a pleiotropic correlation between ξ and ζ) without it necessarily being the case that T has a causal effect on y. It could be the case that T and y are correlated (partly) because y is a cause of T. It could also be the case that T and y are correlated with neither variable causing the other. These rows show the extent to which an estimator will erroneously report that T affects y when in reality it has no effect.

Table 2: Endogeneity between y and T due to Pleiotropic Effects, $h^2 = 0.2$ for both y and T

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1,\ \rho = 0.2$	1.0405	1.2018	0.8247	1.0522	1.0190	1.0249	1.0081	0.9525	0.9744	0.9100
	(0.0001)	(0.0010)	(0.0017)	(0.0011)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1,\ \rho = 0.5$	1.1004	1.5040	1.1240	1.1155	1.0473	1.0617	1.0131	0.9891	1.0150	0.9419
	(0.0001)	(0.0012)	(0.0023)	(0.0015)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1,\ \rho = 0.8$	1.1602	1.8058	1.5649	1.1880	1.0764	1.1011	1.0094	1.0275	1.0567	0.9761
	(0.0001)	(0.0014)	(0.0040)	(0.0051)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.2$	0.0405	0.2018	0.1083	0.0522	0.0190	0.0249	0.0092	0.0139	0.0127	0.0015
	(0.0001)	(0.0010)	(0.0009)	(0.0011)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.5$	0.1004	0.5040	0.2922	0.1155	0.0473	0.0617	0.0180	0.0349	0.0519	0.0016
	(0.0001)	(0.0012)	(0.0012)	(0.0015)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.8$	0.1602	0.8058	0.5956	0.1880	0.0764	0.1011	0.0164	0.0585	0.0857	0.0015
	(0.0001)	(0.0014)	(0.0021)	(0.0051)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parenthesis) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. See the supplementary text for details.

Table 2 shows the results under the conditions of modest heritability for both y and T, where we vary the genetic correlation (ρ) between ζ and ξ (i.e., between the effect of SNPs on T and on y, net of T) between 0.2 and 0.8. As can be seen in the first three row panels for Table 2, the MR estimate for T is upwardly biased, and the bias gets worse as the pleiotropy gets stronger. Indeed, MR seriously underperforms simple OLS (i.e., with no proxy control) in obtaining an accurate estimate for the effect of T on y when T is specified to have an actual effect. In the presence of positive pleiotropy, OLS of course overestimates the effect of T; it attributes the direct effect of SNPs on y to the indirect effect through T. The amount of over-estimation also, as expected, grows with the size of the genetic correlation between the effect of markers on T and their effect on y, net of T. GIV-C, in contrast, provides highly accurate estimates of the effect of T on y even in the case of very strong pleiotropy. Interestingly, GIV-U also produces rather accurate estimates of the effect of T on y, though at a heritability of 0.2, GIV-U underestimates the size of δ. We note that GIV-C and GIV-U are together providing bounds for the true answer at these specifications for the simulation. The use of an unconditional or a conditional proxy in OLS also performs well; it is only slightly less accurate than GIV-C, but sometimes it underestimates and sometimes it overestimates the true answer.

The bottom panels of Table 2 show the performance of the estimators when T has no true effect. MR erroneously finds that T has a significant effect on y and the size of this estimated effect grows with the strength of the pleiotropy. Simple OLS is more accurate, and both GIV-C and GIV-U are more accurate still. They both estimate a very small effect of T on y. This makes sense, because if the true effect of T on y is zero, this means that T should have a very weak relationship in a finite sample with the coefficient errors for the genetic markers in the unconditional PGS for y (i..e, $G(\zeta - \hat{\zeta}_1) = v_1$) hence with the PGS error that is part of the error term in 10. Similarly, S_{y2} will also have a very weak correlation with v_1. Therefore, T and S_{y2} are valid instruments for the case where δ =0 and where there is no non-genetic endogeneity, and the GIV-U estimates are very close to the true answer in this case.

Table 3: Endogeneity between y and T due to Pleiotropic Effects, $h^2 = 0.4$ for both y and T

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1$, $\rho = 0.2$	1.0810	1.2002	0.6374	1.0874	1.0425	1.0497	1.0313	0.8563	0.8983	0.7900
	(0.0001)	(0.0004)	(0.0007)	(0.0003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.5$	1.2011	1.5024	0.9270	1.2086	1.1044	1.1230	1.0604	0.9337	0.9825	0.8575
	(0.0001)	(0.0004)	(0.0009)	(0.0004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.8$	1.3210	1.8040	1.4201	1.3322	1.1640	1.2016	1.0533	1.0228	1.0773	0.9376
	(0.0001)	(0.0004)	(0.0015)	(0.0008)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.2$	0.0810	0.2002	0.0738	0.0874	0.0425	0.0497	0.0336	0.0176	0.0288	0.0025
	(0.0001)	(0.0004)	(0.0003)	(0.0003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.5$	0.2011	0.5024	0.2053	0.2086	0.1044	0.1230	0.0736	0.0447	0.0729	0.0024
	(0.0001)	(0.0004)	(0.0004)	(0.0004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.8$	0.3210	0.8040	0.4707	0.3322	0.1640	0.2016	0.0800	0.0813	0.1295	0.0018
	(0.0001)	(0.0004)	(0.0007)	(0.0008)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. See the supplementary text for details.

Table 3 has the same layout as Table 2, but in Table 3 the heritability for both T and for y is increased from 0.2 to 0.4. Higher heritability slightly increases the positive bias of

GIV-C, and it also increases the negative bias of GIV-U. Even though GIV-C has positive bias, it is always more accurate than MR and also more accurate than simple OLS. Greater heritability increases the positive bias of GIV-C when there is no true effect of T, and the over-prediction is larger when the pleiotropy is stronger. Nevertheless, GIV-C is clearly more accurate than either simple OLS or MR-based estimators. When the true effect of T on y is zero, GIV-U provides clear evidence of this fact. As before, GIV-C and GIV-U are bounding the true answer when it is specified to be 1.0. When the true answer is zero, GIV-U is very close to the true answer.

Table 4: Endogeneity between y and T due to Pleiotropic Effects, $h^2 = 0.6$ for both y and T

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1$, $\rho = 0.2$	1.1213	1.1999	0.5264	1.1249	1.0798	1.0860	1.0713	0.7222	0.7877	0.6244
	(0.0001)	(0.0002)	(0.0004)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.5$	1.3016	1.5020	0.8074	1.3056	1.1979	1.2134	1.1573	0.8469	0.9255	0.7263
	(0.0001)	(0.0002)	(0.0005)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 1$, $\rho = 0.8$	1.4816	1.8033	1.3247	1.4868	1.3136	1.3470	1.1774	1.0101	1.0985	0.8685
	(0.0001)	(0.0002)	(0.0007)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 0$, $\rho = 0.2$	0.1213	0.1999	0.0562	0.1249	0.0798	0.0860	0.0737	0.0196	0.0331	0.0031
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.5$	0.3016	0.5020	0.1586	0.3056	0.1979	0.2134	0.1733	0.0511	0.0867	0.0028
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.8$	0.4816	0.8033	0.3896	0.4868	0.3136	0.3470	0.2267	0.1036	0.1701	0.0019
	(0.0001)	(0.0002)	(0.0003)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. See the supplementary text for details.

Table 4 has the same layout as Tables 2 and 3, but in Table 4, the heritability for both T and for y is increased to 0.6. Higher heritability slightly increases the positive bias of GIV-C when T is specified to have an actual effect on y, and it also increases the negative bias of GIV-U. Even though GIV-C has positive bias, it is always more accurate than MR and also more accurate than simple OLS. GIV-U continues to under-predict the true answer, and GIV-C and GIV-U together continue to provide bounds on the correct answer, though these bounds become gradually wider as we increase the amount of heritability of T and y

in the simulations. When the true effect of T on y is zero, GIV-U provides clear evidence of this fact.

Table 5: Endogeneity between y and T due to Pleiotropic Effects, $h^2 = 0.8$ for both y and T

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1,\ \rho = 0.2$	1.1613	1.1998	0.4528	1.1629	1.1323	1.1359	1.1283	0.5309	0.6309	0.3819
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1,\ \rho = 0.5$	1.4020	1.5018	0.7269	1.4036	1.3303	1.3389	1.3098	0.7034	0.8323	0.4962
	(0.0001)	(0.0001)	(0.0003)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 1,\ \rho = 0.8$	1.6421	1.8030	1.2567	1.6440	1.5292	1.5461	1.4440	0.9822	1.1310	0.7094
	(0.0001)	(0.0001)	(0.0004)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)
$\delta = 0,\ \rho = 0.2$	0.1613	0.1998	0.0454	0.1629	0.1323	0.1359	0.1293	0.0208	0.0357	0.0033
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)
$\delta = 0,\ \rho = 0.5$	0.4020	0.5018	0.1292	0.4036	0.3303	0.3389	0.3179	0.0565	0.0978	0.0028
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.8$	0.6421	0.8030	0.3323	0.6440	0.5292	0.5461	0.4807	0.1325	0.2188	0.0016
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. See the supplementary text for details.

Table 5 shows simulations where the heritability of T and y is specified to be 0.8. Higher heritability further increases the positive bias of GIV-C though it remains more accurate than simple OLS or MR. Very high heritability also increases the negative bias of GIV-U and further widens the gap between GIV-C and GIV-U, though they continue to bound the true answer. When the true effect of T on y is zero, GIV-C is positively biased though not as much as simple OLS or MR. GIV-U remains very accurate in estimating the true effect of T on y when this effect is actually zero.

Next we consider in Table 6 the case where T has a high heritability of 0.8 while y has a low heritability of 0.2 (this is empirically possible because y has other causes than T and these other causes can be largely non-genetic). The pattern of estimates in Table 6 resembles those of Table 4 where heritability is 0.6 for both T and for y.

Table 7, then shows the case where T has a low heritability of 0.2 while y has a high heritability of 0.8. In this case, GIV-C and GIV-U are giving more accurate answers, and GIV-C is giving a much better answer than simple OLS or MR when the true answer is zero. By comparing Table 6 and Table 7, we see that the size of the upward bias of GIV-C depends on how strong is the relative heritability of T and y as well as on how strong is the pleiotropy. But in all of these cases, GIV-C and GIV-U are bracketing the true effect of

Table 6: Endogeneity between y and T due to Pleiotropic Effects, $h^2 = 0.2$ for y and $h^2 = 0.8$ for T

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1$, $\rho = 0.2$	1.0847	1.1033	0.9286	1.0870	1.0795	1.0821	1.0701	0.8836	0.9535	0.6254
	(0.0004)	(0.0006)	(0.0013)	(0.0007)	(0.0004)	(0.0004)	(0.0005)	(0.0007)	(0.0006)	(0.0036)
$\delta = 1$, $\rho = 0.5$	1.2048	1.2541	1.0996	1.2088	1.1940	1.1988	1.1588	0.9961	1.0686	0.7259
	(0.0004)	(0.0006)	(0.0015)	(0.0008)	(0.0004)	(0.0004)	(0.0005)	(0.0008)	(0.0007)	(0.0043)
$\delta = 1$, $\rho = 0.8$	1.3244	1.4042	1.3181	1.3321	1.3145	1.3188	1.2156	1.1385	1.2037	0.8799
	(0.0004)	(0.0006)	(0.0018)	(0.0013)	(0.0004)	(0.0004)	(0.0012)	(0.0009)	(0.0007)	(0.0057)
$\delta = 0$, $\rho = 0.2$	0.0822	0.1012	0.0539	0.0837	0.0715	0.0749	0.0673	0.0301	0.0438	0.0048
	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.5$	0.2025	0.2521	0.1457	0.2044	0.1777	0.1852	0.1619	0.0801	0.1140	0.0059
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 0$, $\rho = 0.8$	0.3224	0.4025	0.2968	0.3257	0.2923	0.3032	0.2442	0.1656	0.2175	0.0087
	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0004)

Mean of estimated effect for T and its standard error(within parenthesis) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. See the supplementary text for details.

Table 7: Endogeneity between y and T due to Pleiotropic Effects, $h^2 = 0.8$ for y and $h^2 = 0.2$ for T

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1$, $\rho = 0.2$	1.2005	1.9798	0.4651	1.2273	1.0562	1.0700	1.0396	0.8675	0.9016	0.8253
	(0.0005)	(0.0061)	(0.0022)	(0.0017)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
$\delta = 1$, $\rho = 0.5$	1.5004	3.4922	1.0200	1.5280	1.1304	1.1664	1.0776	0.9427	1.0010	0.8689
	(0.0005)	(0.0093)	(0.0029)	(0.0024)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
$\delta = 1$, $\rho = 0.8$	1.8001	5.0029	2.3902	1.8300	1.1725	1.2414	1.0647	1.0432	1.1262	0.9365
	(0.0005)	(0.0153)	(0.0125)	(0.0074)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
$\delta = 0$, $\rho = 0.2$	0.0898	0.4412	0.1173	0.1048	0.0274	0.0353	0.0187	0.0125	0.0220	0.0015
	(0.0001)	(0.0016)	(0.0008)	(0.0008)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.5$	0.2240	1.1174	0.3304	0.2400	0.0648	0.0850	0.0377	0.0310	0.0544	0.0010
	(0.0001)	(0.0023)	(0.0011)	(0.0012)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.8$	0.3579	1.7928	0.8247	0.3782	0.0906	0.1290	0.0335	0.0524	0.0917	0.0000
	(0.0001)	(0.0035)	(0.0036)	(0.0034)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. See the supplementary text for details.

T when the true effect is non-zero, and GIV-U gives accurate answers relative to all other methods when the true effect is zero.

3 Estimating exposure effects in the presence of both pleiotropy and genetic-related endogeneity

Pleiotropy, of course, is not the only potential problem that challenges efforts to estimate the effect of T on y with accuracy, and, indeed, it was not the problem that MR was developed to solve. So now we elaborate the structural model to be as follows:

$$y = \delta T + X\beta_y + \gamma S^*_{y|T} + \nu_y + \epsilon_y \tag{16}$$

$$T = \alpha S^*_T + X\beta_T + \nu_T + \epsilon_T \tag{17}$$

Now the disturbance for both equations has two terms, ν and ϵ. We assume that ϵ_T and ϵ_y are uncorrelated, but that ν_T and ν_y are correlated with each other (with correlation ρ_ν), which produces endogeneity in the structural model. We also assume that ν_T and ν_y are correlated with the genetic markers in S^*_T and $S^*_{Y|T}$ ($\rho_{\nu T}$ and $\rho_{\nu y}$, respectively). There are three principal substantive conditions that could alone or in combination produce this correlation. The first condition is epistasis, meaning that SNPs have nonlinear or interactive effects that are correlated with the linear effects in the PGS. The second condition is when rare alleles have effects on y and on T and when these alleles are correlated with observed alleles. The third condition is "genetic nurturing." Genetic nurturing [27] is the condition where the environment of ego is shaped by genetically related individuals to ego. For example, children live in an environment that is partly created and selected by their parents. If environmental characteristics are related to parents' genes, which of course are correlated with ego's genes, and if the environment affects y while also being correlated with T, then the environment is endogenous to T while also being correlated with ego's genes. The model for height on educational attainment provides a useful example. Taller children could be taller partly for genetic reasons, but also because they grew up in an environment that provided better nutrition. Children who grew up in a better nutritional environment would also be expected to go further in school. Parents who provide a better nutritional environment for their children may have done so in part based on genetic advantages, or on behavioral consequences of genetic advantages (e.g., when a taller parent is rewarded for being tall in school or the workplace and therefore has more money to spend on their children). While recognizing that the substantive reasons for this form of endogeneity can vary, we will refer to it below as genetic nurturing for ease of exposition.

In order to evaluate alternative estimation strategies in the presence of both pleiotropy and genetic nurturing, we elaborated the simulations to include the additional error terms ν_T and ν_y, assuming both to have a variance of 0.1 and assuming correlations of 0.4 between ν_T and ν_y ($\rho_\nu = 0.4$). We assumed varying genetic correlations between S^*_T and ν_T ($\rho_{\nu T}$), and between $S^*_{y|T}$ and ν_y ($\rho_{\nu y}$). We further assumed that the correlation between ν_T and $S^*_{y|T}$ is 0.4 as large as is the correlations set between S^*_T and ν_T, and that the correlation between ν_y and S^*_T is 0.4 as large as is the correlation between $S^*_{y|T}$ and ν_y for that particular set of simulations.

Table 8 shows the results from a set of simulations where the correlation between y and ν_y and also between T and ν_T (i.e., $\rho_{\nu y}$ and $\rho_{\nu T}$) is set at 0.2 and where the heritability for both y and T is set at 0.5. As for the simpler simulations that only included pleiotropy, we find that GIV-C consistently outperforms both OLS and MR. In all these cases, GIV-C is positively biased in its estimate of δ, and these biases are comparable to those that we found for the case of pleiotropy alone when the heritability of y and T was of comparable magnitude. As with the simpler case of pleiotropy without genetic nurturing, we find that GIV-U consistently underestimates the effect of T on y, and that GIV-C and GIV-U bracket the correct answer. The pattern of results for the case of a zero effect are also similar to what we saw in the case of moderate heritability without the additional genetic confounding; GIV-C over-predicts but not as much as for OLS and MR, and GIV-U provides a very accurate answer. This pattern is actually similar to what we find when we increase the extent of genetic confounding (i.e., increase $\rho_{\nu y}$ and $\rho_{\nu T}$), as shown in Table 9, where $\rho_{\nu y} = \rho_{\nu T} = 0.5$, and in Table 10, where $\rho_{\nu y} = \rho_{\nu T} = 0.8$.

Table 8: Genetic-Related Endogeneity, with a Correlation of 0.2 between the Polygenic Score and the Genetically-Related Confounder ($\rho_{\nu y} = \rho_{\nu T} = 0.2$)

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1$, $\rho = 0.2$	1.1664	1.2621	0.5575	1.1649	1.1141	1.1201	1.1005	0.7217	0.7938	0.6117
	(0.0001)	(0.0002)	(0.0004)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.5$	1.3005	1.4781	0.7704	1.3021	1.2043	1.2176	1.1692	0.8158	0.8996	0.6877
	(0.0001)	(0.0002)	(0.0004)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 1$, $\rho = 0.8$	1.4390	1.6917	1.0739	1.4361	1.2993	1.3203	1.2130	0.9371	1.0284	0.7906
	(0.0001)	(0.0002)	(0.0005)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 0$, $\rho = 0.2$	0.1664	0.2621	0.0668	0.1649	0.1141	0.1201	0.1042	0.0250	0.0422	0.0011
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.5$	0.3005	0.4781	0.1400	0.3021	0.2043	0.2176	0.1829	0.0457	0.0806	-0.0003
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.8$	0.4390	0.6917	0.2590	0.4361	0.2993	0.3203	0.2467	0.0817	0.1366	-0.0004
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. The heritability for both y and T is 0.5. The variance of both ν_y and ν_T equals 0.1. Correlation of ν_y and ν_T is 0.4. See the supplementary text for details.

If the reasons for the additional endogeneity do arise from genetic nurturing, and if these genetic nurturing effects are the same for siblings or for dizygotic twins, then the

Table 9: Genetic-Related Endogeneity, with a Correlation of 0.5 between the Polygenic Score and the Genetically-Related Confounder ($\rho_{\nu y} = \rho_{\nu T} = 0.5$)

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1,\ \rho = 0.2$	1.1878	1.2789	0.5538	1.1878	1.1357	1.1411	1.1221	0.6942	0.7741	0.5730
	(0.0001)	(0.0002)	(0.0003)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 1,\ \rho = 0.5$	1.3106	1.4609	0.7257	1.3101	1.2235	1.2334	1.1922	0.7825	0.8726	0.6422
	(0.0001)	(0.0002)	(0.0004)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 1,\ \rho = 0.8$	1.4359	1.6493	0.9807	1.4336	1.3111	1.3285	1.2449	0.8918	0.9910	0.7334
	(0.0001)	(0.0002)	(0.0004)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 0,\ \rho = 0.2$	0.1878	0.2789	0.0675	0.1878	0.1357	0.1411	0.1259	0.0265	0.0449	0.0002
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.5$	0.3106	0.4609	0.1252	0.3101	0.2235	0.2334	0.2037	0.0466	0.0797	-0.0002
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.8$	0.4359	0.6493	0.2186	0.4336	0.3111	0.3285	0.2717	0.0747	0.1274	0.0003
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. The heritability for both y and T is 0.5. The variance of both ν_y and ν_T equals 0.1. Correlation of ν_y and ν_T is 0.4. See the supplementary text for details.

Table 10: Genetic-Related Endogeneity, with a Correlation of 0.8 between the Polygenic Score and the Genetically-Related Confounder ($\rho_{\nu y} = \rho_{\nu T} = 0.8$)

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1,\ \rho = 0.2$	1.2084	1.2951	0.5550	1.2087	1.1559	1.1612	1.1427	0.6734	0.7603	0.5423
	(0.0001)	(0.0002)	(0.0003)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 1,\ \rho = 0.5$	1.3193	1.4520	0.7034	1.3192	1.2383	1.2471	1.2109	0.7540	0.8510	0.6031
	(0.0001)	(0.0002)	(0.0003)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 1,\ \rho = 0.8$										
$\delta = 0,\ \rho = 0.2$	0.2084	0.2951	0.0697	0.2087	0.1559	0.1612	0.1465	0.0283	0.0484	0.0005
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)
$\delta = 1,\ \rho = 0.5$	0.3193	0.4520	0.1181	0.3192	0.2383	0.2471	0.2209	0.0462	0.0793	0.0000
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.8$										

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. The heritability for both y and T is 0.5. The variance of both ν_y and ν_T equals 0.1. Correlation of ν_y and ν_T is 0.4. The empty lines were unobtainable parameter combinations. See the supplementary text for details.

inclusion of family fixed effects are a good strategy that can be used in combination with GIV-C and GIV-U, because in these cases the fixed effects estimator will control for the unobserved but common family effect. Note that a fixed effects model among siblings or dizygotic twins does not solve the problem of endogeneity due to pleiotropy, and GIV-C and GIV-U can be used in combination to address that issue. At the same time, we note that family fixed effects models usually are only possible with smaller samples and they use up many degrees of freedom. Given the simulation results in Tables 8-10, it may be that the greater statistical power available in using GIV-C and GIV-U alone offsets any additional advantage from the fixed effects estimator.

4 Estimating exposure effects in the presence of both pleiotropy and genetic-unrelated endogeneity

Next we use simulated data to estimate the effects of T in the presence of both pleiotropy and genetic-unrelated endogeneity. Table 11 shows simulations where the heritability is 0.5 for both y and T and where there is also pleiotropy but where the error terms in equations 8 and 9 have a 0.4 correlation (ρ_e) with each other and are uncorrelated with the genetic variables $S_{T|X}$ and $S_{y|XT}$. As Table 11 shows, this is the most challenging of all the simulation results obtained so far. Neither OLS or MR provide accurate answers at any level of genetic correlation, and the results for GIV-C are not an improvement. Moreover, GIV-U in this case also has a positive bias, and so GIV-C and GIV-U no longer provide bounds for the true answer. The table has a simple message: when the endogeneity problem stems from non-genetic sources, genetic information will not by itself provide a solution to the estimation strategy. Of course, the validity of this message depends on the extent of the endogeneity problem, as can readily be seen in Table 12. The simulations in Table 12 differ from those in Table 11 only in that the non-genetic endogeneity is much weaker; instead of a 0.4 correlation between the errors in equations 8 and 9, we assume a -0.1 correlation. The consequence of weakening the non-genetic endogeneity is that GIV-C produces very accurate estimates of the effect of T, estimates which are more accurate than those of either OLS or MR. Finally, and consistent with the earlier simulations, the use of $S(y1|T)$ as a proxy control in OLS provides an estimate that is more sensitive to the extent of pleiotropy than is GIV-C, being smaller at low levels of genetic correlation and larger at higher levels. However, it is not as consistent in the sign of its bias and therefore is less useful for the purpose of establishing bounds.

While we do not wish to minimize the challenges posed by non-genetic endogeneity, we also note that this situation provides grounds for optimism. Non-genetic sources of endogeneity can often be measured and included in the model as control variables. Once this is done, the endogeneity problem is reduced in severity. We show this illustratively in Tables 13, 14, and 15. In the simulations reported in these tables, the non-genetic endogeneity correlation (ρ_e) is again 0.4, but we specify it explicitly as the consequence of

Table 11: Genetic-Unrelated Endogeneity, with a Correlation of 0.4 between the Error Terms (ρ_e)

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1$, $\rho = 0.2$	1.3017	1.2004	0.6241	1.3040	1.3436	1.3385	1.3490	1.0720	1.1165	0.9992
	(0.0001)	(0.0003)	(0.0006)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.5$	1.4520	1.5032	0.9145	1.4545	1.4306	1.4346	1.4259	1.1992	1.2481	1.1193
	(0.0001)	(0.0002)	(0.0006)	(0.0002)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.8$	1.6019	1.8051	1.4138	1.6055	1.5225	1.5391	1.4684	1.3539	1.4032	1.2721
	(0.0001)	(0.0002)	(0.0007)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.2$	0.3017	0.2004	0.0618	0.3040	0.3436	0.3385	0.3517	0.2222	0.2342	0.2064
	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.5$	0.4520	0.5032	0.1778	0.4545	0.4306	0.4346	0.4267	0.2713	0.3002	0.2302
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.8$	0.6019	0.8051	0.4271	0.6055	0.5225	0.5391	0.4838	0.3648	0.4093	0.2949
	(0.0001)	(0.0002)	(0.0003)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. The heritability for both y and T is 0.5. See the supplementary text for details.

Table 12: Genetic-Unrelated Endogeneity, with a Correlation of -0.1 between the Error Terms (ρ_e)

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1$, $\rho = 0.2$	1.0509	1.1968	0.5579	1.0546	0.9886	0.9980	0.9706	0.7253	0.7791	0.6431
	(0.0001)	(0.0003)	(0.0005)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.5$	1.2012	1.4988	0.8404	1.2047	1.0741	1.0950	1.0174	0.8188	0.8832	0.7200
	(0.0001)	(0.0003)	(0.0006)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.8$	1.3512	1.8003	1.3490	1.3537	1.1531	1.1962	1.0074	0.9331	1.0066	0.8179
	(0.0001)	(0.0003)	(0.0011)	(0.0005)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\delta = 0$, $\rho = 0.2$	0.0509	0.1968	0.0625	0.0546	-0.0114	-0.0020	-0.0240	-0.0314	-0.0192	-0.0479
	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.5$	0.2012	0.4988	0.1773	0.2047	0.0741	0.0950	0.0386	-0.0068	0.0258	-0.0544
	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.8$	0.3512	0.8003	0.4234	0.3537	0.1531	0.1962	0.0520	0.0249	0.0852	-0.0718
	(0.0001)	(0.0003)	(0.0005)	(0.0005)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. The heritability for both y and T is 0.5. See the supplementary text for details.

two unmeasured variables. We then assume that one of these environmental confounds can be measured, and so we include it explicitly in the regression and re-estimate the models. In Table 13, we assume that the measurable environmental variable accounts for 20% of the variance of the original error term, such that the remaining correlation between the error terms is 0.27. In Table 14, we assume instead that the measurable environmental variable accounts for 50%, implying a remaining correlation of 0.20. In Table 15, we assume that it accounts for 80% of the non-genetic endogeneity, implying a remaining correlation of 0.13. Perhaps not surprisingly, the performance of both GIV-C and GIV-U improve and the level of improvement depends upon the amount of the environmental confounding variables that can be controlled. It's notable that even if 80% of the confounding effects of non-genetic environmental variables were controlled, GIV-C still shows considerable upward bias, though generally not as much as MR. GIV-U, on the other hand, performs reasonably well when most of the environmental confounds are controlled.

We also note that GIV-U does not consistently under-predict the effect of T on y when there are positive-biasing environmental confounds as well as pleiotropy. On the other hand, it does reliably give the most conservative answer of all the estimators we have considered. If the pleiotropy is not extreme and if the amount of uncontrolled environmental endogeneity is not too large, then estimates from GIV-U are in the neighborhood of the true answer.

Table 13: Genetic-Unrelated Endogeneity, Partially Controlling for 20% of the Confounds.

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1$, $\rho = 0.2$	1.2433	1.2051	0.6221	1.2948	1.2878	1.2822	1.2967	0.9814	1.0319	0.8989
	(0.0001)	(0.0002)	(0.0005)	(0.0002)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.5$	1.4008	1.5056	0.9103	1.4456	1.3792	1.3828	1.3720	1.1078	1.1646	1.0154
	(0.0001)	(0.0002)	(0.0005)	(0.0002)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1$, $\rho = 0.8$	1.5582	1.8046	1.4053	1.5961	1.4735	1.4908	1.4046	1.2638	1.3226	1.1667
	(0.0001)	(0.0002)	(0.0007)	(0.0002)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.2$	0.2433	0.2051	0.0552	0.2948	0.2878	0.2822	0.2968	0.1590	0.1717	0.1429
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(<0.0001)	(0.0001)	(0.0001)	(<0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.5$	0.4008	0.5056	0.1704	0.4456	0.3792	0.3828	0.3742	0.2053	0.2361	0.1617
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0$, $\rho = 0.8$	0.5582	0.8046	0.4178	0.5961	0.4735	0.4908	0.4267	0.2904	0.3403	0.2121
	(0.0001)	(0.0002)	(0.0003)	(0.0002)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. The heritability for both y and T is 0.5. Controlling for 20% of the original endogeneity. Remaining correlation between error terms is 0.27. See the supplementary text for details.

Another strategy for addressing the environmental confounds problem is to use fixed effects models where the clustered cases (e.g., siblings) have similar values on the environmental variables that are producing the non-genetic environmental endogeneity. In general, we conclude that non-genetic endogeneity causes potentially large problems for estimat-

Table 14: Genetic-Unrelated Endogeneity, Partially Controlling for 50% of the Confounds.

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1,\ \rho = 0.2$	1.2075	1.2057	0.6227	1.2962	1.2521	1.2466	1.2638	0.9349	0.9872	0.8495
	(0.0001)	(0.0002)	(0.0004)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1,\ \rho = 0.5$	1.3643	1.5064	0.9110	1.4469	1.3431	1.3467	1.3346	1.0555	1.1151	0.9587
	(0.0001)	(0.0002)	(0.0004)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 1,\ \rho = 0.8$	1.5210	1.8054	1.4062	1.5973	1.4355	1.4530	1.3563	1.2042	1.2673	1.1002
	(0.0001)	(0.0002)	(0.0005)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.2$	0.2075	0.2057	0.0562	0.2962	0.2521	0.2466	0.2617	0.1233	0.1360	0.1073
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)
$\delta = 0,\ \rho = 0.5$	0.3643	0.5064	0.1712	0.4469	0.3431	0.3467	0.3378	0.1656	0.1969	0.1215
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.8$	0.5210	0.8054	0.4183	0.5973	0.4355	0.4530	0.3839	0.2407	0.2928	0.1591
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. The heritability for both y and T is 0.5. Controlling for 50% of the original endogeneity. Remaining correlation between error terms is 0.20. See the supplementary text for details.

Table 15: Genetic-Unrelated Endogeneity, Partially Controlling for 80% of the Confounds.

	OLS	MR-based			Conditional Proxy PGS			Unconditional Proxy PGS		
		MR	EMR-1	EMR-2	OLS $S(y\|T)$	OLS $S(y_1\|T)$	GIV-C	OLS $S(y)$	OLS $S(y_1)$	GIV-U
$\delta = 1,\ \rho = 0.2$	1.1736	1.2067	0.6237	1.2981	1.2187	1.2132	1.2332	0.8876	0.9425	0.7978
	(0.0001)	(0.0002)	(0.0003)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(0.0001)
$\delta = 1,\ \rho = 0.5$	1.3312	1.5074	0.9121	1.4487	1.3102	1.3138	1.3003	1.0040	1.0672	0.9012
	(0.0001)	(0.0001)	(0.0003)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(0.0001)
$\delta = 1,\ \rho = 0.8$	1.4887	1.8065	1.4079	1.5989	1.4017	1.4195	1.3121	1.1477	1.2157	1.0353
	(<0.0001)	(0.0001)	(0.0003)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(0.0001)
$\delta = 0,\ \rho = 0.2$	0.1736	0.2067	0.0577	0.2981	0.2187	0.2132	0.2291	0.0888	0.1016	0.0728
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)
$\delta = 0,\ \rho = 0.5$	0.3312	0.5074	0.1725	0.4487	0.3102	0.3138	0.3045	0.1276	0.1596	0.0825
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)
$\delta = 0,\ \rho = 0.8$	0.4887	0.8065	0.4195	0.5989	0.4017	0.4195	0.3450	0.1940	0.2488	0.1082
	(<0.0001)	(0.0001)	(0.0001)	(0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)

Mean of estimated effect for T and its standard error (within parentheses) of twenty simulations for several methods (columns) and different parameters (rows). The effect for T (δ) is equal to either 1 or 0. The correlation between the true polygenic scores (ρ) varies. The heritability for both y and T is 0.5. Controlling for 80% of the original endogeneity. Remaining correlation between error terms is 0.13. See the supplementary text for details.

ing causal effects when pleiotropy is moderate to large in size. Fixed effects models with monozygotic twins will solve pleiotropy problems but it is difficult to obtain monozygotic twin data at sufficient scale to address most problems of interest in the social and behavioral sciences. Well-designed experiments using randomized assignment to treatment would address all the problems considered here, though experiments are frequently infeasible to conduct for well-known reasons. Valid non-genetic environmental IVs would similarly address both the problems of environmental endogeneity and pleiotropy, though these variables are often unavailable. In such cases, the strategy of addressing as much non-genetic endogeneity as possible either with explicit control variables or with fixed-effects models and then calculating both the GIV-C and the GIV-U estimates provides more information about the true effect of T than any of the other strategies considered here.

Table 16: List of Parameters, Variances and Heritability

Name	δ	h_y^2	h_T^2	ρ	$\rho_{\nu y}$	ρ_e	var(y)	True h_y^2	std. δ
Pleiotropic Endogeneity	1	0.20	0.20	0.20	0.00	0.00	2.08	0.23	0.48
	1	0.20	0.20	0.50	0.00	0.00	2.20	0.27	0.45
	1	0.20	0.20	0.80	0.00	0.00	2.32	0.31	0.43
	0	0.20	0.20	0.20	0.00	0.00	1.00	0.20	0.00
	0	0.20	0.20	0.50	0.00	0.00	1.00	0.20	0.00
	0	0.20	0.20	0.80	0.00	0.00	1.00	0.20	0.00
	1	0.40	0.40	0.20	0.00	0.00	2.16	0.44	0.46
	1	0.40	0.40	0.50	0.00	0.00	2.40	0.50	0.42
	1	0.40	0.40	0.80	0.00	0.00	2.64	0.55	0.38
	0	0.40	0.40	0.20	0.00	0.00	1.00	0.40	0.00
	0	0.40	0.40	0.50	0.00	0.00	1.00	0.40	0.00
	0	0.40	0.40	0.80	0.00	0.00	1.00	0.40	0.00
	1	0.60	0.60	0.20	0.00	0.00	2.24	0.64	0.45
	1	0.60	0.60	0.50	0.00	0.00	2.60	0.69	0.38
	1	0.60	0.60	0.80	0.00	0.00	2.96	0.73	0.34
	0	0.60	0.60	0.20	0.00	0.00	1.00	0.60	0.00
	0	0.60	0.60	0.50	0.00	0.00	1.00	0.60	0.00
	0	0.60	0.60	0.80	0.00	0.00	1.00	0.60	0.00
	1	0.80	0.80	0.20	0.00	0.00	2.32	0.83	0.43
	1	0.80	0.80	0.50	0.00	0.00	2.80	0.86	0.36
	1	0.80	0.80	0.80	0.00	0.00	3.28	0.88	0.30
	0	0.80	0.80	0.20	0.00	0.00	1.00	0.80	0.00
	0	0.80	0.80	0.50	0.00	0.00	1.00	0.80	0.00
	0	0.80	0.80	0.80	0.00	0.00	1.00	0.80	0.00
	1	0.20	0.80	0.20	0.00	0.00	5.76	0.20	0.17
	1	0.20	0.80	0.50	0.00	0.00	6.00	0.23	0.17
	1	0.20	0.80	0.80	0.00	0.00	6.24	0.26	0.16
	0	0.20	0.80	0.20	0.00	0.00	1.00	0.20	0.00
	0	0.20	0.80	0.50	0.00	0.00	1.00	0.20	0.00
	0	0.20	0.80	0.80	0.00	0.00	1.00	0.20	0.00
	1	0.80	0.20	0.20	0.00	0.00	7.90	0.91	0.13
	1	0.80	0.20	0.50	0.00	0.00	10.90	0.94	0.09
	1	0.80	0.20	0.80	0.00	0.00	13.90	0.95	0.07
	0	0.80	0.20	0.20	0.00	0.00	1.00	0.80	0.00
	0	0.80	0.20	0.50	0.00	0.00	1.00	0.80	0.00
	0	0.80	0.20	0.80	0.00	0.00	1.00	0.80	0.00

Table 16 – continued

Name	δ	h_y^2	h_T^2	ρ	$\rho_{\nu y/T}$	ρ_e	var(y)	True h_y^2	std. δ
Genetic-Related Endogeneity	1	0.50	0.50	0.20	0.20	0.00	2.53	0.47	0.40
	1	0.50	0.50	0.50	0.20	0.00	2.83	0.53	0.35
	1	0.50	0.50	0.80	0.20	0.00	3.13	0.57	0.32
	0	0.50	0.50	0.20	0.20	0.00	1.09	0.46	0.00
	0	0.50	0.50	0.50	0.20	0.00	1.09	0.46	0.00
	0	0.50	0.50	0.80	0.20	0.00	1.09	0.46	0.00
	1	0.50	0.50	0.20	0.50	0.00	2.91	0.41	0.34
	1	0.50	0.50	0.50	0.50	0.00	3.21	0.47	0.31
	1	0.50	0.50	0.80	0.50	0.00	3.51	0.51	0.29
	0	0.50	0.50	0.20	0.50	0.00	1.22	0.41	0.00
	0	0.50	0.50	0.50	0.50	0.00	1.22	0.41	0.00
	0	0.50	0.50	0.80	0.50	0.00	1.22	0.41	0.00
	1	0.50	0.50	0.20	0.80	0.00	3.28	0.37	0.30
	1	0.50	0.50	0.50	0.80	0.00	3.58	0.42	0.28
	1	0.50	0.50	0.80	0.80	0.00	3.88	0.46	0.26
	0	0.50	0.50	0.20	0.80	0.00	1.36	0.37	0.00
	0	0.50	0.50	0.50	0.80	0.00	1.36	0.37	0.00
	0	0.50	0.50	0.80	0.80	0.00	1.36	0.37	0.00
Genetic-Unrelated Endogeneity	1	0.50	0.50	0.20	0.00	0.40	2.60	0.46	0.38
	1	0.50	0.50	0.50	0.00	0.40	2.90	0.52	0.34
	1	0.50	0.50	0.80	0.00	0.40	3.20	0.56	0.31
	0	0.50	0.50	0.20	0.00	0.40	1.00	0.50	0.00
	0	0.50	0.50	0.50	0.00	0.40	1.00	0.50	0.00
	0	0.50	0.50	0.80	0.00	0.40	1.00	0.50	0.00
	1	0.50	0.50	0.20	0.00	-0.10	2.10	0.57	0.48
	1	0.50	0.50	0.50	0.00	-0.10	2.40	0.63	0.42
	1	0.50	0.50	0.80	0.00	-0.10	2.70	0.67	0.37
	0	0.50	0.50	0.20	0.00	-0.10	1.00	0.50	0.00
	0	0.50	0.50	0.50	0.00	-0.10	1.00	0.50	0.00
	0	0.50	0.50	0.80	0.00	-0.10	1.00	0.50	0.00

List of parameters and the corresponding variance of y. The variance of T is always equal to 1. The effect sizes of the genetic markers are kept constant in each table, so there is no compensation for an increase in genetic correlation or for the correlation with the confounds. Hence the true heritability of y changes. δ is the effect of T, ρ is the genetic correlation, ρ_e is the correlation between the error terms in y and T (to create environmental confounds) and $\rho_{\nu y}$ is the correlation between the genetic confounds for y and polygenic score for y. In all simulations $\rho_{\nu y} = \rho_{\nu T}$. The last two columns show the actual heritability for y and the standardized effect size for T.

5 Empirical application

We used data from the Health and Retirement Survey (HRS) for our empirical example [28]. The HRS is a longitudinal survey on health, retirement and aging which is presentative for the US population aged 50 years or older. The survey consists of eleven waves from 1992 to 2012. We used phenotypic data that has been cleaned and harmonized by the RAND cooperation.[8]

Since 2006, data collection has expanded to include biomarkers and a subset of the participants has been genotyped.[9] Autosomal SNPs were imputed using the worldwide reference panel from phase I of the 1000 Genomes project (v3, released March 2012) [29]. If the uncertainty about the genotype of an individual was greater than 10 percent, the SNP was removed. Furthermore, SNPs were removed from the entire sample if the imputation quality was below 70 percent, if the minor allele frequency was smaller than 1 percent, or if the SNP was missing in over 5 percent of the sample. Our analyses were restricted to unrelated participants of European descent according to the standard HRS protocol. Specifically, HRS filtered out parent-offspring pairs, siblings and half-siblings. Selection on European descent was done based on self reported race and principal component analysis [30]. The PGS for educational attainment is negatively correlated with birth year ($r =$ -0.06; $p < 0.0001$) and educational attainment has been shown to affect longevity [31, 32]. Thus, age-related sample selection is likely to be correlated with educational attainment and its PGS, which could potentially bias our results. Since the HRS is a sample of an older population spanning across many birth years, we further restricted our analysis sample to a a relatively younger group of people born between 1935 and 1945. This subsample is still large enough to for our analyses ($N = 2,839$), yet less likely to be affected by age-related sample selection. .

We constructed polygenic scores starting with a set of 2,224,079 SNPs that were either directly genotyped in HRS or present in the HapMap3 reference panel [33], providing us with a high-resolution coverage of common genetic variants. To control for linkage disequilibrium (LD) between SNPs, we constructed all polygenic scores using LDpred [6] with the default LD window (total number of SNPs divided by 3000) and assuming that all of the SNPs are causal.

The first unconditional polygenic score for educational attainment was constructed by using GWAS results provided by the Social Science Genetic Association Consortium (SSGAC) [26], excluding HRS, UK Biobank and the *23andMe* cohort from the meta-analysis. The remaining SSGAC sample consists of several cohorts from around the world ($n = 207,605$, see Supplementary Table 18). We included all SNPs that overlapped with our initial set in LDpred. After LDpred filtered out ambiguous SNPs and SNPs with minor

[8] RAND HRS Data, Version O. Produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration. Santa Monica, CA (August 2016). See `http://www.rand.org/labor/aging/dataprod/hrs-data.html` for additional information.

[9] See https://hrs.isr.umich.edu/data-products/genetic-data

allele frequency smaller than 0.01; 1,849,602 autosomal SNPs remained.

The second unconditional polygenic score for educational attainment was constructed by using results from a GWAS in the UK Biobank, also provided by the SSGAC [34] ($n = 442,183$; 1,870,853 SNPs).

The first unconditional polygenic score for height was constructed using the publicly available GWAS summary results from the GIANT consortium ($n = 253,288$) [35],[10] which are based on ≈ 2.5 million autosomal SNPs that were imputed using the HapMap 2 CEU reference panel [36] (See Supplementary Table 19). Merging this set with the directly genotyped and HapMap 3 SNPs resulted in 1,264,571 SNPs that were included in the score by LDpred.

We conducted three GWASs in the UK Biobank (UKB) to obtain the other required polygenic scores. The UKB is a publicly available population-based prospective study of individuals aged 40-69 years during recruitment in 2006-2010 [37]. We restricted the analysis to unrelated Brits of European descent [38] that were available in the full release of the genetic data ($n = 441,298$). Autosomal SNPs were imputed using the UK10K reference panel. Details on genotyping, pre-imputation quality control, and imputation have been documented extensively elsewhere [38].

To obtain a second unconditional polygenic score for height, the GWAS analysis included as control variables dummies for genotyping batches and sex. We also included a third order polynomial of age and it's interaction terms with sex. Furthermore, the first 20 principal components of the genetic data were also included to control for subtle population structure. The obtained GWAS results underwent quality control following an extended version of the EasyQC protocol [39] described in detail elsewhere [40]. Two loci had SNPs with p-values that were numerically equal to 0, these could not be entered into LDpred. From each of the two loci one SNP was included into the score after LDpred was done. This yielded a score consisting of 1,861,847 autosomal SNPs.

For the conditional polygenic score for educational attainment, we included as control variables height, genotyping batches, sex, age, age and height squared and cubed, the interaction terms between the terms for age and height, as well as their interaction terms with sex. Furthermore, the first 20 principal components of the genetic data were included as controls for population stratification and the GWAS results underwent quality control, yielding 1,861,878 autosomal SNPs.

For the conditional polygenic score for height, a identical GWAS analysis was conducted where we controlled for educational attainment instead (including squared, cubic and interaction terms). This yielded a score based on 1,861,847 autosomal SNPs (including the same two SNPs that were manually added, as described above).

There is an overlap in the cohorts used by the GIANT consortium in the GWAS on height and by the SSGAC GWAS on educational attainment [26]. To ensure independence

[10] http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files\#GWAS_Anthropometric_2014_Height

of measurement errors in the PGS, whenever the GIANT height PGS was used, we excluded the other as an instrument and used a PGS constructed from the UK Biobank GWAS results instead.

Using these data, we demonstrate the value of the GIV regression approach in several important empirical applications. First, we estimated the chip heritability of educational attainment (EA) in the HRS data from a PGS for EA. We use the residual of EA after regressing it on control variables. The results are shown in Table 1 of the main text. All reported coefficients are standardized. Since the squared standardized coefficient in OLS equals R^2, our OLS result in column 1 of Table 1 implies that the PGS for EA currently captures 6.8% of the variance in EA.

Using the GIV regression results reported in columns 2 and 3 of Table 1 and the error correction described above 1.1, we obtain chip heritability estimates of 13.4% (95% CI +/- 3.9%) and 13.8% (+/- 4.0%), respectively.

Second, we estimated the (causal) effect of body height on EA. Earlier studies have reported a positive relationship between these variables [41, 42, 43]. Third, we present results from a negative control that estimates the (causal) effect of EA on body height (which should be zero). We estimated these effects using OLS, MR, GIV-C, and GIV-U regression. In each regression, we included birth year, birth year squared, educational attainment of both parents and (in pooled models) sex as control variables. We included PGS of EA or height depending on the method. All variables have been standardized. The results are shown and discussed in main text (see Tables 3 and 4).

6 Practical recommendations

We discussed two main sources of bias in this paper – direct pleiotropy and unobserved environmental confounds that may or may not covary with genetic effects. These sources of bias are relevant in almost all research questions in the social sciences and epidemiology when experimental data is not available.

The existing literature addresses these challenges with various strategies. All of them have their advantages and disadvantages. For example, panel data that contain repeated measures for each individual over time can be used for individual fixed-effects models that control for all unobserved heterogeneity among people, including genetic and environmental factors. Unfortunately, individual fixed-effects models do not allow investigating variables that do not vary over time for a particular person, such as the relationship between educational attainment and body height.

The gold standard to address potential bias arising from genetic and family-specific environmental confounds is a comparison among MZ twin pairs. These pairs are (almost) genetically identical and share the same family environment. However, very large samples of MZ twin pairs are necessary for this approach because within MZ twin pair variation tends to be very small or non-existent. Also, this approach does by itself not control for unobserved environmental confounds that are individual-specific.

Probably the most popular approach to identify causal effects in non-experimental data are instrumental variable techniques. Yet, convincing environmental instruments are rare and they limit the scope of research questions to scenarios to which the instruments apply. Furthermore, as discussed earlier, genetic instruments are invalid when they have direct pleiotropic effects on the exposure and the outcome or if they are correlated with other unobserved confounds.

This leaves a broad class of important applied research questions for which GIV regression offers a new approach to obtain more precise estimates than ordinary multiple regression techniques or approaches that use invalid instruments. Table 17 provides an overview of different types of applied research questions and our recommended estimation strategy in cross-sectional population samples that lack an experimental design, or valid non-genetic instruments, or relevant natural experiments. We differentiate these research questions based on the expected degree of pleiotropic confounds and whether environmental sources of endogeneity may also exist or not. Unfortunately, environmental endogeneity is hard to rule out in almost all non-experimental research scenarios.

Mendelian Randomization is in principle a great idea for addressing environmental endogeneity, but its application is limited to scenarios where direct pleiotropy between the exposure and the outcome is of no concern. An example may be the influence of number of cigarettes smoked per day on the number of biological offspring – smoking intensity seems to be regulated by a relatively limited number of genes with strong effects and clear biological functions that are unlikely to have direct pleiotropic effects on reproductive success [44]. Yet, even in this situation, genes related to smoking may still violate the exclusion restriction via LD with other genes or via their correlation with unobserved environmental confounds, such as parental socioeconomic status. In short, it is difficult to argue convincingly that the assumptions of MR are actually satisfied. The assumptions of MR are less likely to hold the more genetically complex the investigated traits are, the higher their genetic correlation is, and the more likely it is that the genes associated with these traits work via unobserved environmental channels.

We argue that GIV regression is a reasonable estimation strategy whenever pleiotropic confounds are a possible concern. If genetic and environmental confounds are both likely to exist, we recommend the combination of GIV regression with control variables that correct for non-genetic endogeneity as far as possible, ideally in samples that also allow controls for family-fixed effects (e.g. siblings or DZ twins). Examples of research questions with both sources of endogeneity are plentiful, e.g. the relationships between body height and educational attainment (low pleiotropy), diet and body mass (probably with a medium degree of direct pleiotropy), and the returns to schooling (probably with a high degree of direct pleiotropy on educational attainment and personal income, and quite likely mediated by factors such as cognitive ability and personality). GIV regression in combination with environmental controls is a reasonable estimation strategy in all of these cases.

An important practical question is data availability for GIV regression. In addition to a

Table 17: Guidance for applications

Direct pleiotropy	Environment endogeneity	Recommended method	Example
No / very low	Yes	MR	Smoking intensity on number of children
	No	OLS	Randomized controlled trials
Low	Yes	GIV + FFE + Controls*	Body height on educational attainment
	No	GIV	x
Medium	Yes	GIV + FFE + Controls*	Diet on body mass
	No	GIV	x
High	Yes	GIV + FFE + Controls*	Returns to schooling
	No	GIV	x

MR – Mendelian Randomization, OLS – Ordinary Least Squares, GIV – Genetic Instrumental Variable regression, FFE – Family fixed-effects, ideally estimated in pairs of dizygotic twins. *How well this strategy works depends on the strength of the residual environmental endogeneity after adjusting for controls and FFE; more residual environmental endogeneity will lead to more bias in the estimates of treatment T and outcome y.
x No good example is known to us.

genotyped prediction sample, the researcher will need GWAS summary statistics from non-overlapping samples to construct the conditional and unconditional scores. Unconditional scores for many traits can often be constructed using publicly available GWAS results from consortia such as GIANT [11], SSGAC [12], PGC [13], or CHARGE [14]. Most of these consortia did not include data from the UKB in their earlier publications. Thus, the publicly available UKB data can often be used to obtain a second score from an independent sample. Unconditional GWAS results for virtually all traits in the UKB are publicly available from the Broad Institute [15]. The UKB, or any other large, publicly available biobank, can also

[11] https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
[12] https://www.thessgac.org
[13] https://www.med.unc.edu/pgc/results-and-downloads
[14] http://www.chargeconsortium.com/main/results
[15] http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank

be considered as a source for obtaining conditional GWAS results. If the researcher does not have access to these data or lacks the resources for large-scale GWAS analyses, it is always possible to team up with one of the many research groups around the world that have the necessary data and resources. Running a conditional GWAS in a large sample like the UKB can often be done in a matter of hours by experienced research teams. Importantly, GIV regression only requires one conditional score which can be instrumented by an unconditional score from a non-overlapping sample. Thus, data access or computational resources should not be a serious practical limitation for GIV regression.

Furthermore, the UKB is large enough to be split into three sub-samples for GIV regression. One particularly appealing approach would be to use the subsample of siblings in the UKB as a prediction sample that allows the researcher to control for family fixed-effects. The remaining unrelated individuals can be split into two, still very large, subsamples to conduct conditional and unconditional GWAS analyses. Because all participants in the UKB have been recruited at about the same time and in the same country, the genetic correlations for a given trait are likely to be perfect for randomly chosen subsets of the data.

An important practical issue is that the prediction sample should not be included in any of the GWAS samples used to construct the scores to avoid overfitting. Reassuringly, this is not a problem either because most GWAS consortia provide meta-analysis results excluding specific samples upon request. If this is not possible, an alternative strategy is to conduct the GWAS on y in the prediction sample and to subtract the effect of each SNP in this cohort from the publicly available results using the meta-analysis formula that the consortium used to aggregate effects. For example, if the meta-analysis used sample size weights to obtain the z-scores of each SNP, the corrected z-scores excluding the prediction sample could be obtained by simply subtracting the z-score in the prediction sample using the appropriate weight [45]. Furthermore, many samples have only recently been genotyped and are therefore not included yet in published GWAS studies. These samples could be readily employed for GIV regression using the approaches described above.

Overall, we believe that GIV regression has substantial practical utility for many researchers and across a wide range of important applied research questions. The usefulness of GIV regression will increase further in the future as a result of the growing availability of accurate, cheap genetic data and GWAS results on many traits from ever growing sample sizes.

Table 18: Cohort List for Educational Attainment Score from SSGAC

Study	Full name	Sampling	Country	Sample size
ACPRC	Manchester Studies of Cognitive Ageing	Population-based	England	1713
AGES	Age, Gene / Environment Susceptibility-Reykjavik Study	Population-based	Iceland	3212
ALSPAC	Avon Longitudinal Study of Parents and Children	Population-based birth cohort	England	2877
ASPS	Austrian Stroke Prevention Study	Population-based	Austria	777
BASE-II	Berlin Aging Study II	Population-based	Germany	1619
CoLaus	Cohorte Lausannoise	Population-based	Switzerland	3269
COPSAC2000	Copenhagen Studies on Asthma in Childhood 2000	Case-control birth cohort	Germany	318
CROATIA-Korčula	Croatia Korčula	Population-based (Isolate)	Croatia	842
deCODE	deCODE genetics	Population-based	Iceland	46758
DHS	Dortmund Health Study	Population-based	Germany	953
DIL	Wellcome Trust Diabetes and Inflammation Laboratory	Population-based	England	2578
EGCUT1	Estonian Genome Center, University of Tartu	Population-based	Estonia	5597
EGCUT2	Estonian Genome Center, University of Tartu	Population-based	Estonia	1328
EGCUT3	Estonian Genome Center, University of Tartu	Population-based	Estonia	2047
ERF	Erasmus Rucphen Family Study	Family-based	Nether-lands	2433
FamHS	Family Heart Study	Family-based	USA	3483
FINRISK	The National FINRISK Study	Case-control (Cardiovascular health)	Finland	1685
FTC	Finnish Twin Cohort	Family-based	Finland	2418
GOYA	Genetics of Overweight Young Adults	Case-control (Obesity)	Denmark	1459

Table 18 – continued

Study	Full name	Sampling	Country	Sample size
GRAPHIC	Genetic Regulation of Arterial Pressure in Humans	Population-based	England	727
GS	Generation Scotland	Population-based	Scotland	8776
H2000 Cases	Health 2000	Case-control (Metabolic syndrome)	Finland	797
H2000 Controls	Same as above	Case-control (Metabolic syndrome)	Finland	819
HBCS	Helsinki Birth Cohort Study	Population-based birth cohort	Finland	1617
HCS	Hunter Community Study	Population-based	Australia	1946
HNRS (CorexB)	Heinz Nixdorf Recall Study	Population-based	Germany	1401
HNRS (Oexpr)	Same as above	Same as above	Germany	1347
HNRS (Omni1)	Same as above	Same as above	Germany	778
Hypergenes	Hypergenes	Case-control	Italy/ UK/ Belgium	815
INGI-CARL	Italian Network of Genetic Isolates - Carlantino	Population-based (Isolate)	Italy	947
INGI-FVG	Italian Network of Genetic Isolates - Friuli Venezia Giulia	Population-based (Isolate)	Italy	943
KORA S3	Kooperative Gesundheitsforschung in der Region Augsburg	Population-based	Germany	2655
KORA S4	Same as above	Population-based	Germany	2721
LBC1921	Lothian Birth Cohort 1921	Population-based birth cohort	Scotland	515
LBC1936	Lothian Birth Cohort 1936	Population-based birth cohort	Scotland	1003
LifeLines	The LifeLines Cohort Study	Population-based	Nether-lands	12539

Table 18 – continued

Study	Full name	Sampling	Country	Sample size
MCTFR	Minnesota Center for Twin and Family Research	Family-based, but only founders used.	USA	3819
MGS	Molecular Genetics of Schizophrenia	Population-based	USA	2313
MoBa	Mother and Child Cohort of NIPH	Population-based (Nested case-control)	Norway	622
NBS	Nijmegen Biomedical Study	Population-based	Nether-lands	1808
NESDA	Netherlands Study of Depression and Anxiety	Case-control (Mental health)	Nether-lands	1820
NFBC66	Northern Finland Birth Cohort 1966	Population-based	Finland	5297
NTR	Netherlands Twin Register	Family-based	Nether-lands	5246
OGP	Ogliastra Genetic Park	Population-based	Italy	370
OGP-Talana	Ogliastra Genetic Park-Talana	Population-based (Isolate)	Italy	544
ORCADES	Orkney Complex Disease Study	Population-based (Isolate)	Scotland	1828
PREVEND	Prevention of Renal and Vascular End-stage Disease	Population-based	Nether-lands	3578
QIMR	Queensland Institute of Medical Research	Family-based	Australia	8006
RS-I	Rotterdam Study Baseline	Population-based	Nether-lands	6108
RS-II	Rotterdam Study Extension of Baseline	Population-based	Nether-lands	1667
RS-III	Rotterdam Study Young	Population-based	Nether-lands	3040
Rush-MAP	Rush University Medical Center - Memory and Aging Project	Community-based	USA	887
Rush-ROS	Rush University Medical Center - Religious Orders Study	Community-based	USA	808

Table 18 – continued

Study	Full name	Sampling	Country	Sample size
SardiNIA	SardiNIA Study of Aging	Family-based	Italy	5616
SHIP	Study of Health in Pomerania	Population-based	Germany	3556
SHIP-TREND	Study of Health in Pomerania	Population-based	Germany	901
STR - Salty	Swedish Twin Registry	Family-based	Sweden	4832
STR - Twingene	Swedish Twin Registry	Family-based	Sweden	9553
THISEAS	The Hellenic Study of Interactions between SNPs & Eating in Atherosclerosis Susceptibility	Case-control	Greece	829
TwinsUK	St Thomas UK Adult Twin Registry	Population-based	England	4012
WTCCC58C	1958 British Birth Cohort	Population-based	England	2804
YFS	The Cardiovascular Risk in Young Finns Study	Population-based	Finland	2029

This table contains the list of cohorts used in the GWAS of Educational Attainment of [26], excluding the Health and Retirement Study and 23andMe cohorts. A more detailed list and description can be found in the supplementary materials of [26]

Table 19: Cohort List for Height Score from GIANT

Study	Full name	Sampling	Country	Sample size
ACTG	The AIDS Clinical Trials Group	Population-based	Interna- tional	1055
ADVANCE	Atherosclerotic Disease, VAscular FunctioN, and GenetiC Epidemiology	Population-based case-control	USA	584
AE	Athero-Express Biobank Study	patient-cohort	The Nether- lands	686
AGES	Age, Gene/Environment SusceptibilityReykjavik Study	Population-based	Iceland	3219
Amish HAPI Heart Study	Amish Heredity and Phenotype Intervention Heart Study	Founder population	USA	907
ARIC	Atherosclerosis Risk in Communities Study	Population-based	USA	8110
ASCOT	AngloScandinavian Cardiac Outcome Trial	"Randomised control clinical trial"	UK, Ireland and Nordic Regions	3802
B58C-T1DGC	British 1958 birth cohort (Type 1 Diabetes Genetic Consortium controls)	Populationbased birth cohort	UK	2591
B58C-WTCCC	British 1958 birth cohort (Wellcome Trust Case Control Consortium controls)	Populationbased birth cohort	UK	1479
BHS	Busselton Health Study	Population-based	Australia	1328
BLSA	Baltimore Longitudinal Study on Aging	Population-based	USA	844
B-PROOF	Baltimore Longitudinal Study on Aging	"Randomised control clinical trial"	Nether- lands	2669

Table 19 – continued

Study	Full name	Sampling	Country	Sample size
BRIGHT	British Genetic of Hypertension (BRIGHT) study	Hypertension cases	UK	1806
CAD-WTCCC	WTCCC Coronary Arteryt Disease cases	Case series	UK	1879
CAPS1 cases	Cancer Prostate in Sweden 1	Case-control	Sweden	489
CAPS1 controls	Cancer Prostate in Sweden 1	Case-control	Sweden	491
CAPS2 cases	Cancer Prostate in Sweden 2	Case-control	Sweden	1483
CAPS2 controls	Cancer Prostate in Sweden 2	Case-control	Sweden	519
CHS	Cardiovascular Health Study	Population-based	USA	3228
CoLaus	Cohorte Lausannoise	Population-based	Switserland	5409
Corogene	Genetic Predisposition of Coronary Heart Disease in Patients Verified with Coronary Angiogram	Population-based	Finland	3758
deCODE	deCODE genetics sample set	Population-based	Iceland	26799
DESIR	Data from an Epidemiological Study on the Insulin Resistance syndrome	Population-based	France	716
DGI cases	Diabetes Genetics Initiative	Case-control	Scandinavia	1317
DGI controls	Diabetes Genetics Initiative	Case-control	Scandinavia	1090
DNBC	Danish National Birth Cohort - Preterm Delivery Study	Case-control	Denmark	1802
EGCUT	Estonian Genome Center, University of Tartu	Population-based	Estonia	1417
EGCUT-370	Estonian Genome Center, University of Tartu	Population-based	Estonia	866
EGCUT-OMNI	Estonian Genome Center, University of Tartu	Population-based	Estonia	1356

Table 19 – continued

Study	Full name	Sampling	Country	Sample size
EPIC-Obesity Study	European Prospective Investigation into Cancer and Nutrition - Obesity Study	Population-based	UK	3552
ERF	Erasmus Rucphen Family Study	Family-based	Nether-lands	2726
FamHS	Family Heart Study	Population-based	USA	1463
Fenland	Fenland Study	Population-based	UK	1402
FINGES-TURE cases	Finnish Genetic Study of Arrhythmic Events	Disease cohort (MI cases only)	Finland	943
FRAM	Framingham Heart Study	Population-based, multi-generational	USA	8089
FTC	Finnish Twin Cohort	Monozygotic twins	Finland	125
FUSION cases	Finland-United States Investigation of NIDDM Genetics	Case-control	Finland	1082
FUSION controls	Finland-United States Investigation of NIDDM Genetics	Case-control	Finland	1167
GENMETS cases	Health 2000 / GENMETS substudy of Metabolic syndrome	Case-control	Finland	824
GENMETS controls	Health 2000 / GENMETS substudy of Metabolic syndrome	Case-control	Finland	823
GerMiFSI (cases only)	German Myocard Infarct Family Study I	Case-control	Germany	600
GerMiFSII (cases only)	German Myocard Infarct Family Study II	Case-control	Germany	1124

Table 19 – continued

Study	Full name	Sampling	Country	Sample size
GOOD	Gothenburg Osteoporosis and Obesity Determinants Study	Population-based	Sweden	938
HBCS	Helsinki Birth Cohort Study	Birth cohort study	Finland	1726
Health ABC	Health, Aging, and Body Composition Study	longitudinal cohort study	USA	1655
HERITAGE Family Study	Health, Risk Factors, Training and Genetics (HERITAGE) Family Study	Family Study, baseline data from an exercise training intervention	USA	500
HYPER-GENES Cases	HYPERGENES	Case-control	Italy/ UK/ Belgium	1841
HYPER-GENES Controls	HYPERGENES	Case-control	Italy/ UK/ Belgium	1900
InCHIANTI	Invecchiare in Chianti	Population-based	Italy	1138
IPM Mount Sinai BioMe	The Charles Bronfman Institute for Personalized Medicine BioMe Biobank Program	Hospital-based	USA	2867
KORA S3	Cooperative Health Research in the Region of Augsburg, KOoperative Gesundheitsforschu ng in der Region Augsburg	Population-based	Germany	1643
KORA S4	Cooperative Health Research in the Region of Augsburg, KOoperative Gesundheitsforschu ng in der Region Augsburg	Population-based	Germany	1811

Table 19 – continued

Study	Full name	Sampling	Country	Sample size
LifeLines	LifeLines Cohort study	Population-based	Netherlands	8118
LLS	Leiden Longevity Study	Family based	Netherlands	1903
LOLIPOP_EW610	London Life Sciences Prospective Population Study	Population-based	UK	927
LOLIPOP_EWA	London Life Sciences Prospective Population Study	Population-based with some enrichment	UK	513
LOLIPOP_EWP	London Life Sciences Prospective Population Study	Population-based with some enrichment	UK	651
MGS	Molecular Genetics of Schizophrenia/NIMH Repository Control Sample	Population-based (survey research method)	USA	2597
MICROS	MICROS (EUROSPAN)	Population-based	Italy	1079
MIGEN	Myocardial Infarction Genetics Consortium	Case-control	USA / Finland / Italy / Spain / Sweden	2652
NBS-WTCCC	WTCCC National Blood Service donors	Population-based	UK	1441
NELSON	Dutch and Belgian Lung Cancer Screening Trial		Netherlands and Belgium	2668
NFBC1966	Northern Finland Birth Cohort 1966	Population-based	Finland	4499
NHS	The Nurses' Health Study	Nested case-control	USA	3217
NSPHS	Northern Sweden Population Health Study (EUROSPAN)	Population-based	Sweden	652

Table 19 – continued

Study	Full name	Sampling	Country	Sample size
NTRNESDA	Netherlands Twin Register & the Netherlands Study of Depression and Anxiety	Case-control	Nether-lands	3522
ORCADES	Orkney Complex Disease Study (part of EUROSPAN)	Population-based	Scotland	695
PLCO	The Prostate, Lung Colorectal and Ovarian Cancer Screening Trial	Case-control	USA	2244
PLCO2 controls	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	Population-based case-control	USA	1193
PREVEND	Prevention of REnal and Vascular ENdstage Disease (PREVEND) Study	Population-based	Nether-lands	3624
PROCARDIS	Precocious Coronary Artery Disease	Population-based	UK	7000
PROSPER/ PHASE	The PROspective study of Pravastatin in the Elderly at Risk for vascular disease	Randomized controlled trial	Nether-lands, Scotland and Ireland	5244
QFS	Quebec Family Study	Family-based???	Canada	860
QIMR	Twin study at Queensland Institute of Medical Research	Population-based	Australia	3627
RISC	Relationship between Insulin Sensitivity and Cardiovascular disease Study	Population-based	Europe	1031
RS-I	Rotterdam Study I	Population-based	Nether-lands	5744
RS-II	Rotterdam Study II	Population-based	Nether-lands	2124
RS-III	Rotterdam Study III	Population-based	Nether-lands	2009

Table 19 – continued

Study	Full name	Sampling	Country	Sample size
RUNMC	Nijmegen Bladder Cancer Study (NBCS) & Nijmegen Biomedical Study (NBS), Radboud University Nijmegen Medical Centre	Population-based	Nether-lands	2873
SardiNIA	SARDINIA	Population-based	Italy	4298
SASBAC cases	Swedish And Singapore Breast Association Consortium	Case-control	Sweden	794
SASBAC controls	Swedish And Singapore Breast Association Consortium	Case-control	Sweden	758
SEARCH / UKOPS	Studies of Epidemiology and Risk factors in Cancer Heredity / UK Ovarian Cancer Population Study	Population-based	UK	1592
SHIP	Study of Health in Pomerania	Population-based	Germany	4092
SHIP-TREND	Study of Health in Pomerania - TREND	Population-based	Germany	986
Sorbs	Sorbs are selfcontained population from Eastern Germany, European Descent	Population-based	Germany	907
T2D-WTCCC	WTCCC Type 2 Diabetes cases	case series	UK	1903
TRAILS	Tracking Adolescents' Individual Lives Survey	Population-based (measured at 18yrs of age)	Nether-lands	1139
TWINGENE	TWINGENE	Population-based	Sweden	9380
TwinsUK	TwinsUK	Twins pairs	UK	1479
VIS	VIS (EUROSPAN) and KORCULA	Population-based	Croatia	784

Table 19 – continued

Study	Full name	Sampling	Country	Sample size
WGHS	Women's Genome Health Study	Population-based	USA	23099
YFS	The Cardiovascular Risk in Young Finns Study	Population-based cohort	Finland	1995

This table contains the list of cohorts used in the GWAS of height from [35]. A more detailed list and description can be found in the supplementary materials of [35] and [46]

References

[1] Chabris CF, Lee JJ, Cesarini D, Benjamin DJ, Laibson DI (2015) The fourth law of behavior genetics. *Curr. Dir. Psychol. Sci.* 24(4):304–312. 1.1

[2] Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 9(3):e1003348. 1.1, 1.1

[3] McCarthy MI et al. (2008) Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9(5):356–369. 1.1

[4] Price AL et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38(8):904–909. 1.1

[5] Abdellaoui A, Al. E (2013) Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* 21(11):1277–1285. 1.1

[6] Vilhjálmsson BJ et al. (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97(4):576–592. 1.1, 1.1, 5

[7] Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3(10):e3395. 1.1

[8] Witte JS, Visscher PM, Wray NR (2014) The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* 15(11):765–776. 1.1

[9] Rietveld CA et al. (2013) GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 340(6139):1467–1471. 1.1, 1.1

[10] Branigan AR, McCallum KJ, Freese J (2013) Variation in the Heritability of Educational Attainment: An International Meta-Analysis. *Soc. Forces* 92(1):109–140. 1.1

[11] Bielby WT, Hauser RM, Featherman DL (1977) Response errors of nonblack males in models of the stratification process. *Journal of the American Statistical Association* 72(360a):723–735. 1.1

[12] Bollen KA (2002) Latent variables in psychology and the social sciences. *Annual Review of Psychology* 53(1):605–634. 1.1

[13] Angrist J, Pischke JS (2009) *Mostly harmless econometrics: an empiricist's companion.* (Princeton University Press, NJ, USA). 1.1, 1.1

[14] Burgess S, Small DS, Thompson SG (2015) A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research* p. 0962280215597579. 1.1, 2

[15] Davey Smith G, Hemani G (2014) Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* 23(R1):R89–98. 1.1, 2

[16] Auton A et al. (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74. 1.1

[17] Bansal V et al. (2017) Gwas results for educational attainment aid in identifying genetic heterogeneity of schizophrenia. *bioRxiv.* 1.1

[18] Wray NR et al. (2014) Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry* 55(10):1068–1087. 1.1

[19] de Vlaming R et al. (2017) Meta-gwas accuracy and power (metagap) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. *PLOS Genetics* 13(1):1–23. 1

[20] Davidson R, MacKinnon JG (2004) *Econometric Theory and Methods.* (Oxford University Press), pp. 177 – 212. 1.1

[21] Tucker-Drob EM (2017) Measurement Error Correction of Genome-Wide Polygenic Scores in Prediction Samples. 3

[22] Wickens MR (1972) A note on the use of proxy variables. *Econometrica: Journal of the Econometric Society* pp. 759–761. 2

[23] Aigner DJ (1974) MSE dominance of least squares with errors-of-observation. *Journal of Econometrics* 2(4):365–372. 2

[24] Burgess S, Butterworth A, Malarstig A, Thompson SG (2012) Use of Mendelian randomisation to assess potential benefit of clinical intervention. 2

[25] Bulik-Sullivan B et al. (2015) An atlas of genetic correlations across human diseases and traits. *Nature genetics* 47(11):1236–1241. 2

[26] Okbay A et al. (2016) Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*. 2, 5, 18

[27] Kong A et al. (2017) The nature of nurture: effects of parental genotypes. *bioRxiv* p. 219261. 3

[28] Sonnega A et al. (2014) Cohort profile: the Health and Retirement Study (HRS). *Int. J. Epidemiol.* 43(2):576–585. 5

[29] The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65. 5

[30] Weir D (2012) Quality control report for genotypic data (`http://hrsonline.isr.umich.edu/sitedocs/genetics/HRS_QC_REPORT_MAR2012.pdf`). Accessed: 2017-04-06. 5

[31] van Kippersluis H, O'Donnell O, van Doorslaer E (2011) Long-run returns to education. *Journal of Human Resources* 46(4):695–721. 5

[32] Cutler DM, Lleras-Muney A (2008) Education and health: Evaluating theories and evidence in *Mak. Am. Heal. Soc. Econ. Policy as Heal. Policy*, eds. House J, Schoeni R, Kaplan G, Pollack H. (Russell Sage Foundation, New York), p. 37. 5

[33] Consortium TIH (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58. 5

[34] Lee JJ et al. (2018) Gene discovery and polygenic prediction from a 1.1-million-person gwas of educational attainment. *in press*. 5

[35] Wood AR et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46(11):1173–1186. 5, 19

[36] Consortium TIH (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861. 5

[37] Sudlow C et al. (2015) UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12(3):e1001779. 5

[38] Marchini J et al. (2015) Genotype imputation and genetic association studies of UK Biobank. 5

[39] Winkler TW et al. (2014) Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* 9(5):1192–1212. 5

[40] Okbay A et al. (2016) Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* 5

[41] Silventoinen K, Kaprio J, Lahelma E (2000) Genetic and environmental contributions to the association between body height and educational attainment: a study of adult Finnish twins. *Behavior genetics* 30(6):477–485. 5

[42] Case A, Paxson C, Islam M (2009) Making sense of the labor market height premium: Evidence from the British Household Panel Survey. *Economics letters* 102(3):174–176. 5

[43] AnneÂ Case, ChristinaÂ Paxson (2008) Stature and status: Height, ability, and labor market outcomes. *Journal of Political Economy* 116(3):499–532. 5

[44] The Tobacco and Genetics Consortium (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* 42(5):441–7. 6

[45] Willer CJ, Li Y, Abecasis GR (2010) METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26(17):2190–2191. 6

[46] Allen HL et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467(7317):832–838. 19