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

Thomas A. DiPrete
Columbia University

Casper Burik and Philipp Koellinger
VU University Amsterdam

May 4, 2017

This research was facilitated by the Social Science Genetic Association Consortium (SSGAC) and by the research group on genetic and social causes of life chances at the Zentrum für interdisplinäre Forschung (ZiF) Bielefeld. Data analyses make use of the UK Biobank resource under application number 11425. We acknowledge data access from the Genetic Investigation of ANthropometric Traits Consortium (GIANT). We used data from the Health and Retirement Study (HRS), which is supported by the National Institute on Aging (NIA U01AG009740, RC2 AG036495, RC4 AG039029). HRS genotype data can be accessed via the database of Genotypes and Phenotypes (dbGaP, accession number phs000428.v1.p1). Researchers who wish to link genetic data with other HRS measures that are not in dbGaP, such as educational attainment, must apply for access from HRS. We are very grateful to Richard Karlsson Linnér for help with the GWAS analyses in the UK Biobank and to Aysu Okbay for providing us with subsets of the GWAS meta-analysis on educational attainment. We thank Patrick Turley, Daniel J. Benjamin, Jonathan P. Beauchamp, Niels Rietveld, Eric Slob, Hans van Kippersluis, Benjamin Domingue, and Lisbeth Trille Loft for productive discussions and comments on earlier versions of the manuscript. The study was supported by funding from an ERC Consolidator Grant (647648 EdGe, Philipp D. Koellinger).

Abstract

We introduce Genetic Instrumental Variables (GIV) regression – a method to estimate causal effects in non-experimental data with many possible applications in the social sciences and epidemiology. In non-experimental data, genetic correlation between the outcome and the exposure of interest is a source of bias. Instrumental variable (IV) regression is a potential solution, but valid instruments are scarce. Existing literature proposes to use genes related to the exposure as instruments (i.e. Mendelian Randomization – MR), but this approach is problematic due to possible pleiotropic effects of genes that can violate the assumptions of IV regression. In contrast, GIV regression provides accurate estimates for the causal effect of the exposure and gene-environment interactions involving the exposure under less restrictive assumptions than for MR. As a valuable byproduct, GIV regression also provides accurate estimates of the chip heritability of the outcome variable. GIV regression uses polygenic scores (PGS) for the exposure and the outcome of interest, both of which can be constructed from genome-wide association study (GWAS) results. By splitting the GWAS sample for the outcome into non-overlapping subsamples, we obtain multiple indicators of the outcome PGS that can be used as instruments for each other. In two empirical applications, we demonstrate that our approach produces reasonable estimates of the chip heritability of educational attainment (EA) and, unlike the results using MR, GIV regression estimates find that the positive relationship between body height and EA is primarily due to genetic confounds that have pleiotropic effects on both traits.

Introduction

A major challenge in the social sciences and in epidemiology is the identification of causal effects in non-experimental data. In these disciplines, ethical and legal considerations along with practical constraints often preclude the use of experiments to randomize the assignment of observations between treatment and control groups or to carry out such experiments in samples that represent the relevant population [1]. Instead, many important questions are studied in field data which make it difficult to discern between causal effects and (spurious) correlations that are induced by unobserved factors [2]. Obviously, confusing correlation with causation is not only a conceptual error, it can also lead to ineffective or even harmful recommendations, treatments, and policies, as well as a significant waste of resources (e.g., as in [3]).

One important source of bias in field data is genetic effects: Twin studies [4] as well as methods based on molecular genetic data [5, 6] can be used to estimate the proportion of variance in a trait that is due to the linear genetic effects (so-called narrow-sense heritability). Using these and related methods, an overwhelming body of literature demonstrates that almost all important human traits, behaviors, and health outcomes are influenced both by genetic predisposition as well as environmental factors ([7, 8, 9]). Most of these traits are "genetically complex", which means that the observed heritability is due to the accumulation of effects from a very large number of genes that each have a small, often statistically insignificant, influence [10]. Furthermore, genes often influence several seemingly unrelated traits (i.e. they have "pleiotropic effects") [11] and genetic correlations between many traits have been convincingly demonstrated [12], giving rise to unobserved variable bias in field studies that do not control for the genetic predisposition of individuals for the exposure and the outcome of interest.

One popular strategy to isolate causal effects in non-experimental data is to use instrumental variables (IVs) which "purge" the exposure of its correlation with the error term in the regression [13]. IVs need to satisfy two important assumptions. First, they need to be correlated with the exposure of interest conditional on the other control variables in the regression (i.e. IVs need to be "relevant"). Second, they need to be independent of the error term of the regression conditional on the other control variables and produce their correlation with the outcome solely through their effect on the exposure. In practice, finding valid IVs that satisfy both requirements is difficult. In particular, the second requirement (the so-called exclusion restriction) is challenging.

Epidemiologists have proposed to use genetic information to construct IVs and termed this approach Mendelian Randomization (MR) [14, 15, 16, 17]. The idea is in principle appealing because genotypes are randomized in the production of gametes by the process of meiosis. Thus, conditional on the genotype of the parents, the genotype of the offspring are the result of a random draw. So if it would be known which genes affect the exposure, it may be possible to use them as IVs to identify the causal influence of the exposure on some outcome of interest. Yet, there are four challenges to this idea. First, we need to know which genes affect the exposure and isolate true genetic effects from environmental confounds that are correlated with ancestry. Second, if the exposure is a genetically complex trait, any gene by itself will only capture a very small part of the variance in the trait, which leads to the well-known problem of weak instruments [18, 19]. Third, genotypes are only randomly assigned *conditional* on the genotype of the parents. Unless it is possible to control for the genotype of the parents, the genotype of the offspring is *not* random and correlates with everything that the genotypes of the parents correlate with (e.g. parental environment, personality, and habits) [20]. Fourth, the function of most genes is not completely understood. Therefore, it is difficult to rule out direct pleiotropic effects of genes on the exposure and the outcome, which would violate the exclusion restriction [16].

Recent advances in complex trait genetics make it possible to address the first two challenges of MR. Array-based genotyping technologies have made the collection of genetic data fast and cheap. As a result, very large datasets are now available to study the genetic architecture of many human traits and a plethora of robust, replicable genetic associations has recently been reported in large-scale genome-wide association studies (GWAS) [21]. These results begin to shed a light on the genetic architecture that is driving the heritability of traits such as body height [22], BMI [23], schizophrenia [24], Alzheimer's disease [25], depression [26], or educational attainment (EA) [27]. High quality GWASs use several strategies to control for genetic structure in the population and indeed, empirical evidence suggests that the vast majority of the reported genetic associations for many traits is not confounded by ancestry [28, 29, 30, 31]. Furthermore, so-called polygenic scores (PGS) have become the favored tool for summarizing the genetic predispositions for genetically complex traits [32, 33, ?, 27]. PGS are linear indices that aggregate the effects of all currently measured genetic variants (typically single nucleotide polymorphisms, a.k.a. SNPs), and recent studies demonstrate the ability of PGS to predict genetically complex outcomes such as height, BMI, schizophrenia, and EA [34, 22, 23, 24, ?]. For example, a polygenic score for EA currently captures 4-6% of the variance in the trait and replicates extremely well across different hold-out samples [27]. Although PGS still capture substantially less of the variation in traits than suggested by their heritability [35] (an issue we return to below), PGS capture a much larger share of the variance of genetically complex traits than individual genetic markers. The third challenge could in principle be addressed if the genotypes of the parents and the offspring are observed (e.g. in a large sample of trios) or by using large samples of dizygotic twins where the genetic differences between siblings are random draws from the parent's genotypes. However, the fourth challenge (i.e. pleiotropy) remains a serious obstacle despite recent efforts to relax the exogeneity assumptions in MR ([36, 37]).

Here, we present a novel method that we call Genetic Instrumental Variables (GIV) regression that can be implemented using widely available statistical software. In contrast to MR, GIV regression does not require strong assumptions about the causal mechanism of genes because it effectively controls for possible pleiotropic effects of genes. In particular, GIV regression is based on the insight that adding the true PGS for the outcome to a regression model would effectively eliminate bias arising from a genetic correlation between the outcome and an exposure of interest. Furthermore, we argue that the attenuated predictive accuracy of PGS is conceptually similar to the well-known problem of measurement error in regression analysis. Instrumental variable (IV) techniques can correct attenuation bias in regression coefficient estimates that results from measurement error [38]. We argue that it is possible to obtain a valid IV for a PGS by randomly splitting the GWAS sample that was used for its construction. Typically, a GWAS is used to estimate the effects of individual SNPs in a discovery sample. Then, the estimated effects are utilized as weights for the genetic data in an independent prediction sample. By splitting the GWAS sample into independent subsamples, one can obtain several PGS (i.e. multiple indicators) in the prediction sample. Each will have even lower predictive accuracy than the original score due to the smaller GWAS subsamples used in their construction, but these multiple indicators can be used as IVs for each other, and the instruments will satisfy the assumptions of IV regression to the extent that the measurement errors (the difference between the true and calculated PGS) are uncorrelated.

We show that it is possible under plausible assumptions to obtain consistent estimates of the narrow-sense heritability of a trait by using IV regression that utilizes two PGS that were constructed this way. Then we extend the idea to the problem of estimating causal effects in non-experimental data. We argue that using multiple indicators of the PGS of the outcome together with a PGS for the exposure produce IVs that come reasonably close to satisfying the assumptions of IV regression. Finally, we demonstrate how our approach can be straightforwardly extended to obtain causal estimates of gene-environment interactions (GxE) on outcomes.

We begin by laying out the assumptions of our approach and prove that GIV regression yields consistent estimates for the effect of the PGS on the outcome variable, when the other covariates in the model are exogenous and when the true PGS is uncorrelated with the error term net of the included covariates. We then turn to the more complex case of when a regressor of interest (T) is potentially correlated with unobserved variables in the error term because of pleiotropy, and we show that the bias under these assumptions with GIV regression is generally smaller than with OLS, MR, or what we will term an enhanced version of MR (EMR). We then use simulations to test how our approach behaves in finite sample under plausible assumptions about genetic correlations and then show how sensitive our method is to violations of the assumptions in comparison to MR and EMR.

Next, we demonstrate the practical usefulness of our approach in empirical applications using the publicly available Health and Retirement Study [39]. First, we demonstrate that a consistent estimate of the so-called chip heritability [35] of EA can be obtained with our method. Then, we estimate the effects of body height on EA. As a "negative control," we check whether our method finds a causal effect of EA on body height (it should not). [1]

[1] Note that a clean experimental design which randomizes people into groups based on body height or EA is not possible. Thus, any attempt to study the causal relationship between the two variables must rely on observational data and naturally occurring experiments like the genetic endowment of individuals, which we exploit here.

Theory

Assumptions

The methods we describe builds on the standard identifying assumptions of IV regression [13]. In the context of our approach, this implies four specific conditions:

1. Complete genetic information: The available genetic data include all variants that influence the variable(s) of interest.

2. Genetic effects are linear: All genetic variants influence the variable(s) of interest via additive linear effects. Thus, there are no genetic interactions (i.e. epistasis) or dominant alleles.

3. Genome-wide association studies successfully control for population structure: In other words, the available regression coefficients for the genetic variants are not systematically biased by omitted variables that describe the genetic ancestry of the population. Failure to control for population structure can lead to spurious genetic associations [20].

4. It is possible to divide GWAS samples into non-overlapping sub-samples drawn from the same population as the sample used for analysis.

Estimating narrow-sense SNP heritability from polygenic scores

Under these assumptions, consistent estimates of the chip heritability of a trait[2] can be obtained from polygenic scores (for full details, see Supporting Information section 2). If y is the outcome variable, X is a vector of exogenous control variables, and S_y^* is a summary measure of genetic tendency for y, then one can write

$$\begin{aligned} y &= \alpha + X\beta + \gamma S_y^* + \epsilon \\ &= \alpha + X\beta + \gamma(G\zeta_y) + \epsilon \end{aligned} \tag{1}$$

where G is an $n \times m$ matrix of genetic markers, and ζ is the $m \times 1$ vector of SNP effect sizes, where the number of SNPs is typically in the millions. If the true effects of each SNP on the outcome were known, the true genetic tendency S_y^* would be expressed by the PGS for y, and the marginal R^2 of S_y^* 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 ([40]; [33]; [35]). This is akin to the well-known attenuation bias resulting from measurement error [41]. We refer to the estimate of the PGS from available GWAS data as S_y, where

$$S_y = S_y^* + v_1 = G\zeta_y + Gu_y \tag{2}$$

and substitute S_y for S_y^* 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_y and converges to the SNP-based 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 [35].

It has long been understood that multiple indicators can, under certain conditions, provide a strategy to correct regression estimates for attenuation from measurement error ([42]; [43]). 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 ([38]; [44]): (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 [38]. 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 regression.

[2] i.e. the proportion of variance in a trait that is due to linear affects of currently measurable SNPs

The most straightforward solution to the problem of attenuation bias is to obtain multiple indicators of the PGS by splitting the GWAS discovery sample for y into two mutually exclusive subsamples. This produces noisier estimates of S^*_y, with lower predictive accuracy, but the multiple indicators can be used as IVs for each other (SI appendix). Standard 2SLS regression using S_{y1} as an instrument for S_{y2} will then recovered a consistent estimate of γ in equation 1.

Reducing bias arising from genetic correlation between exposure and outcome

The logic from above can be extended to situations where the question of interest is not the chip heritability of y per se, but rather the effect of some non-randomized exposure on y (e.g. a behavioral or environmental variable, or a non-randomized treatment due to policy or medical interventions). We can rewrite equation 1 by adding a treatment variable of interest T, such that

$$y = \alpha + \delta T + X\beta + (\gamma S^*_y + \epsilon) \tag{3}$$

where, for example, y is EA and T is body height. 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 (S^*_y) is also correlated with the exposure represented by T in equation (3). If S^*_y is not observed and controlled for in equation 3, $\hat{\delta}$ will be a biased estimate of the effect of T on y.[3]

In standard Mendelian randomization (MR), a measure of genetic tendency (S_T) for a behavior of interest (T in equation 3) 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, is exogenous ([46];[44]). One such example would be the use of a PGS for height as an instrument for height in a regression of EA on height. The problem with this approach is that the PGS for height will fail to satisfy the exclusion restriction if (some) of the genes affecting height also have a direct effect on EA (e.g. via healthy cell growth and metabolism) or if they are correlated with unmeasured environmental factors that affect EA.[4] Note that this problem arising from pleiotropic effects of genes is not solved even if infinitely large GWAS samples would be available.

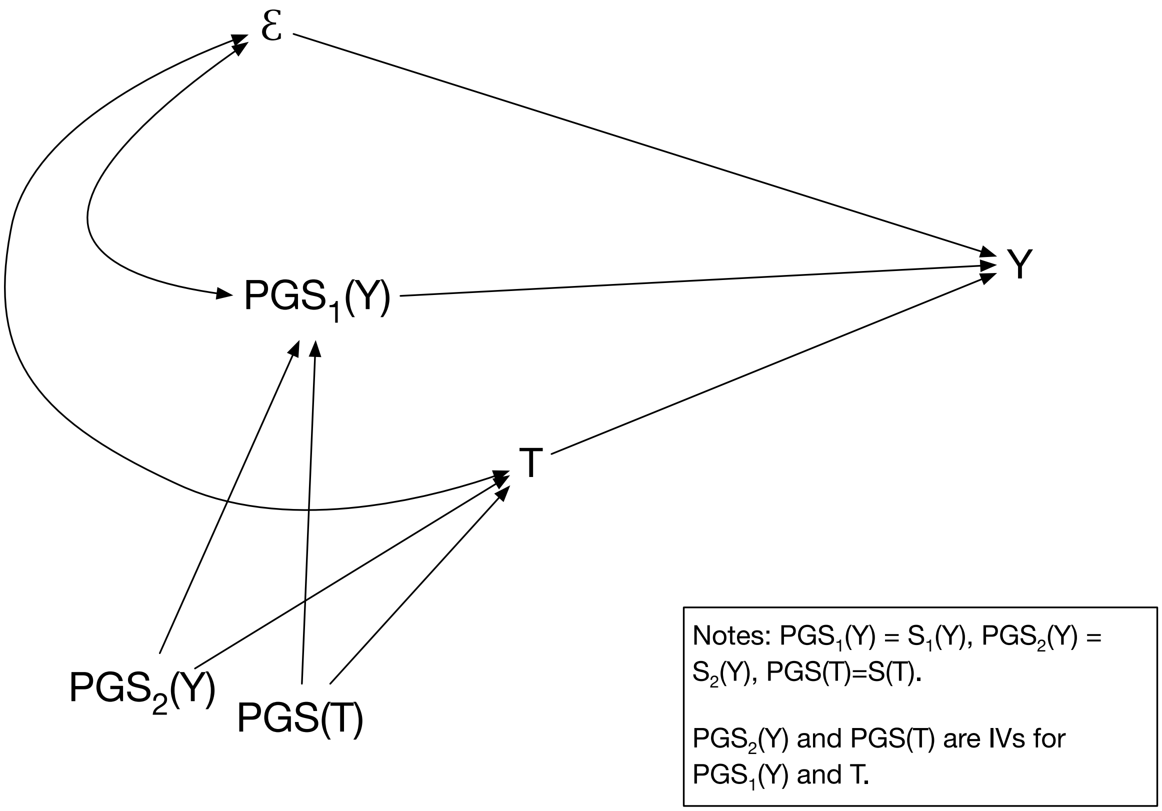
The multiple indicator strategy described above provides multiple approaches for addressing the bias in MR. If the genetic propensity for y could be directly controlled in the regression, MR would provide less biased estimates of the effect of T. We refer to the combined use of S_{y1} as a control and S_T as an IV as "enhanced Mendelian Randomization" (EMR). However, controlling for S_{y1} as a proxy for S^*_y is not adequate, both because it leaves a component of S^*_y in the error term which causes the exclusion restriction assumption of MR to fail, and because the bias in the estimated coefficient of S_{y1} also produces bias in the estimated coefficient of T. The bias arising from the use of a proxy for S^*_y is a form of omitted variable bias (SI appendix).

Violation of the exclusion restriction due to genetic correlation is potentially solved (or at least is less severe) when a third indicator of the PGS for y, i.e., S_{y3} is used to instrument simultaneously both S_{y1} and T in equation 1. However, the practical problem with using two indicators S^*_y as the sole instruments is that their mutual correlation will be relatively high (depending on their reliability) and they are weak instruments for T. As a practical strategy, the best solution is arguably to use S_T along with S_{y2} (or S_{y2} and S_{y3}) as instruments for S_{y1} and T. S_T will still violate the exclusion restriction to the extent that it is correlated with Gu_{y1}. However, the extent of the violation will be reduced by the presence of S_{y1} in the regression. Arguably a strategy that both reduces the correlation between S_T and ϵ (through the inclusion of S_{y1} in the model) and eliminates or greatly reduces omitted variable bias through the inclusion of an instrument for S_{y1} in the first stage equation will outperform MR in the estimation of a consistent effect of T that is purged of genetic correlation. Figure 1 illustrates the GIV regression strategy we propose.

[3] This is the standard case of omitted variable bias, see [45].

[4] Classic MR typically does not use PGS as instruments. Instead, the idea is to use single genetic variants that are known to affect the exposure via well-understood biological mechanisms that make it unlikely to violate the exclusion restriction. In practice, limited knowledge about the biological function of most genes make it difficult to argue that direct pleiotropic effects of the gene on the exposure and the outcome of interest exist.

Figure 1: Genetic Instrumental Variables (GIV) regression



Notes: PGS$_1$(Y) = S$_1$(Y), PGS$_2$(Y) = S$_2$(Y), PGS(T)=S(T).

PGS$_2$(Y) and PGS(T) are IVs for PGS$_1$(Y) and T.

As noted above, if not all relevant genetic effects are contained in the PGS (e.g. interaction effects, structural variants, or rare alleles may be missing given currently available GWAS data), the PGS instruments above will not perfectly satisfy the exclusion restriction to the extent that S_y^* is correlated with the omitted genetic variables. However, the above approach would generally be expected to reduce bias due to genetic correlation, given that a large fraction of heritability can be attributed to linear effects of common SNPs that are well tagged by currently available genotyping arrays [35, 47, 34, 48]). See the SI appendix for details.

Gene-environment interactions

We next generalize equation (3) to the case of gene-environment interactions, where the effect of T varies with the PGS. In principle, these interactions could be extremely complicated and so for practical reasons, swe focus her on obtaining plausible estimates of the linear interaction between S_y^* and T. We rewrite equation (3) as

$$\begin{aligned} y &= \alpha + \delta_1 T + X\beta + (\gamma S_y^* + \delta_2 T S_y^* + \epsilon) \\ &= \alpha + \delta_1 T + X\beta + (\gamma\,(S_{y1} - v_1) \\ &\qquad\qquad\qquad\qquad + \delta_2 T\,(S_{y1} - v_1) + \epsilon) \\ &= \alpha + \delta_1 T + X\beta + \gamma S_{y1} + \delta_2 T S_{y1} \\ &\qquad\qquad\qquad\qquad + (\epsilon - v_1 - \delta_2 T v_1) \end{aligned} \tag{4}$$

Now there are three endogenous variables, T, S_{y1}, and TS_{y1}. Also the disturbance term has now been elaborated to include a term that is a function of T, and so an additional PGS for y is needed as an additional instrumental variable. This additional PGS for y will allow the use of IV regression to estimate δ_2. In the simulations described in section SI 3.3 (see SI Figure 14), GIV regression performs better than OLS, MR, or EMR in estimating the parameters of equation 4. Of course, the term TS_y^* may not fully capture all gene-environment interactions involving T or other environmental variables. Correlations between the IVs and variables in the error term of equation 4 will violate the assumptions of IV regression. We address issues of violated assumptions below.

Simulations

We explored the robustness of GIV regression in finite sample sizes using a range of simulation scenarios (SI appendix). The simulations generate data from a set of known models, which we then analyzed to produce coefficient estimates of the effect of the PGS for y on y and the effect of T on y. We produce these estimates using OLS, MR, EMR, and GIV regression, and compare these results with the true answer across a range of parameter values. The simulations specify that the true PGS scores for y and T are correlated and that the observed PGS scores for y and T are constructed with error. We make the conservative assumption that the entire genetic correlation between the traits is due to Type 1 pleiotropy, i.e. all genes that are associated with both phenotypes have direct effects on both.[5] In practice, this is unlikely to be the case, but it is equally unlikely that one can put a credible upper bound on (or completely rule out) Type 1 pleiotropy. In one set of scenarios, we make the assumption that the entire endogeneity problem arises from the genetic correlation between y and T, a problem which would be solved if we could measure the PGS for these two phenotypes without error. In a second set of simulations, we make the additional assumption that endogeneity arises from other (e.g., environmental) sources that cause the disturbance term in the structural equation for y to be correlated with the disturbance term in the structural model for T even if the true PGS for y and for T were in the respective structural equations. In a third set of simulations, we assume that the genetic factors that affect T are correlated with the environmental factors in the disturbance term for y, as would be the case if parental genes, which affect the PGS for T, also either cause or select for environmental factors that affect y net of T and the PGS for y. We conducted simulations which alternatively specify that the effects of T and the PGS for y are additive and that the effects interact (i.e., where the effect of T on y depend on the PGS for y). The simulations alternately assume that the underlying distributions for the errors in the structural equations for y and T are multivariate normal and deviate from normal via the introduction of skew and kurtosis. They also alternately assume that the errors in the equation for the observed PGS for T and the multiple observed PGS for y are independent or correlated. Finally, we simulate the scenario where even the true PGS for y and T fail to capture all the genetic effects on y and T because they omit rare genetic variants, and where the rare variants for y are correlated with the rare variants for T.

The details of these simulation results are described in the SI appendix. The results provide considerable support for the claim that GIV regression greatly improves our ability to estimate the effects of variables that may have a causal effect on an outcome variable but where genetic correlation and other forms of endogeneity are present. When the only problem is measurement error in the PGS for y and T, GIV regression produces results that bracket the true answer. GIV regression also provides accurate results when the errors of the two structural models are correlated for reasons beyond measurement error in the PGS. Skewness and Kurtosis in the distributions of the errors do not much affect the quality of GIV regression estimates.

When the measurement errors of the PGS are correlated (i.e. when overlapping GWAS samples were used to construct the PGS for y and the genetic IVs), the exclusion restriction is violated and GIV regression estimates are biased. However, we find that GIV regression still outperforms MR or EMR for small to moderately correlated measurement errors ($\rho < 0.5$). It is also encouraging to find that missing genetic variants from the PGS for y and T do not lead to noteworthy bias in the GIV regression estimate for the effect of T on y. Finally, we find that in situations where endogeneity is induced by the effects of or cor-

[5] as opposed to cascade effects where, for example, a component of the PGS for y affects y indirectly through its effect on height that then causes higher EA

relation between parental genes and the environment of the parent's children, estimates from all methods are, as expected, biased. However, we still find that GIV regression outperforms all other methods in terms of the size of the bias of the estimated effect of T on the outcome. In the next section, we discuss these scenarios and our results in greater detail.

Violated assumptions

We now elaborate on the assumptions that GIV regression is based on, and discuss what our simulations tell us about how GIV regression performs under potential violations in comparison with OLS, MR, and EMR.

1. Complete genetic information: Current GWAS are based on two technologies to obtain genetic data. First, so-called genotyping arrays are used to extract information from DNA samples for a selected sub-set of genetic markers. Array technologies allow high throughput and are substantially cheaper than sequencing the entire human genome, which mostly consists of genetic information that does not vary among humans. Instead, array technologies focus on genetic markers that are known to vary within or across specific human populations. Second, one makes use of the fact that genetic markers which are physically close to each other on a chromosome tend to be correlated. This allows genotyping arrays to focus on one or a few SNPs per region that represent the genetic variations (so-called haplotypes) which can be found among humans. Next, information from fully sequenced reference samples is used to impute the missing SNPs [49]. This approach yields highly accurate information for common genetic polymorphisms [50]. However, genotyping and imputation accuracy attenuate strongly for rare polymorphisms as well as for so-called structural genetic variants (e.g. deletions, insertions, inversions, copy-number variants) that are not directly included in the genotyping array. Newer genotyping arrays tend to capture more and better selected polymorphisms than older arrays. Furthermore, increasing sample sizes of completely sequenced reference populations allow imputation of missing genetic variants with ever increasing accuracy [50]. Nevertheless, this implies that the assumption of complete genetic information is violated in practice, although this is likely to be a temporary issue. Another implication of this assumption is that it will be important in practice to ensure that all PGS used in GIV regression are constructed from the same or at least from largely overlapping sets of SNPs.[6]

While it is not possible to know the impact of genetic variants that are not yet included in GWAS data, recent research [47] finds that the *1000 Genomes* imputed data imply very little bias for our method arising from correlation between missing genetic information and the SNPs used to estimate the PGS for height, because the *1000 Genomes* imputed data contains almost all the narrow-sense heritability of these traits. Specifically, we used Yang et al's results to infer the effects of rare variants on y (and also on T) in the SI appendix, and we then computed the bias via simulations using a range of correlations between common and rare genetic variants. The simulation shows that our results are robust across a range of plausible values for these correlations (Supplementary Figure 12).

2. Genetic effects are linear: Possible violations of this assumption could arise if non-linear genetic effects such as systematic gene-gene interactions (a.k.a. epistasis) ended up in the error terms of both scores or if y is affected by genetic dominance or by unmeasured genetic markers (e.g. very rare alleles or structural variants not included in the GWAS or the prediction sample). In other words, suppose that the true structural equation is

$$\begin{aligned} y &= \alpha + \delta T + X\beta + \gamma S_y^* + f(G) + \epsilon \qquad (5) \\ &= \alpha + \delta T + X\beta + \gamma \left(S_{y_1}^* + v_1\right) + f(G) + \epsilon \\ &= \alpha + \delta T + X\beta + \gamma S_{y1} + (f(G) - v_1 + \epsilon) \end{aligned}$$

[6] If it were actually possible to observe the PGS for y based on the true effect sizes of all genetic variants that affect y (call this S_y^{**} to distinguish it from S_y^*, which is the PGS obtained from the true coefficients and all the genetic markets in G – see equation 2), an IV model with S_T as an instrument would no longer be identified. This problem could in principle be solved by isolating the genetic information that only affects y through T from S_y^{**} , though there is no statistical strategy for determining this. If one somehow knew S_y^{**}, one could regress S_y^{**} on S_T and use the residual as the control and S_T as the IV for T. This model would be identified, but it might violate the exclusion restriction if S_T affected y through other channels than T, meaning that S_T would, in this specification, be correlated with now-omitted (genetic) information in the error term of the structural model.

where $f(G)$ includes interaction terms between the various genetic markers in G, the effects of unmeasured genetic markers and other nonlinear effects. The presence of $f(G)$ in equation (5) may cause the exclusion rule to be violated; S_{y2} may be correlated with the disturbance term because S_y^* may be correlated with the non-zero interaction effects in $f(G)$. This problem is not solved even if the measurement error in the PGS was essentially eliminated through the use of extremely large GWAS samples and using ordinary least squares to estimate equation (5); the problem stems from the failure to control for (or find an instrument for) $f(G)$. Imagine that S_y^* perfectly captures the linear effects of the measured SNPs. If $f(G)$ is uncorrelated with S_y^* , then the estimated effect of S_y^* will be consistent, but the estimate of the proportion of variance in y that is statistically explained by genetic factors will be underestimated and the standard error of the effect of S_y^* will be higher than if $f(G)$ were observed. If $f(G)$ is positively correlated with S_y^*, then the true effect of S_y^* will be overestimated and the total variance explained by genetic factors will be underestimated. If $f(G)$ is negatively correlated with S_y^*, then the true effect of S_y^* will be underestimated and the total variance explained by genetic factors will also be underestimated. Epistasis certainly exists to some extent. However, the observed twin correlations for the majority of traits (69%) are consistent with a simple and parsimonious model where twin resemblance is solely due to additive genetic variation and where epistasis is therefore not a major problem [8].

3. Genome-wide association studies successfully control for population structure: Violations of this assumption lead to biased PGS or to PGS that predict y for non-genetic reasons (as when a model without population controls makes it seem as if Italians like pasta or the Chinese use chopsticks for genetic reasons) [20]. This can lead to the violation of the exclusion restriction if the population structure variables that are correlated with S_y^* are not controlled for in the PGS or the structural model, and if these population variables affect the outcome of interest. Multiple indicators of S_y^* would not resolve this omitted variable bias because each of these indicators would also be correlated with the omitted variables.

4. It is possible to divide GWAS samples into non-overlapping sub-samples drawn from the same population as the sample used for analysis. In principle, this assumption seems unproblematic: the availability of large-scale, population-based, genotyped datasets such as the UK Biobank makes it straightforward to randomly split the sample into parts and to exclude genetically related or identical observations. One practical issue is that one may want to use results from published GWAS studies to construct polygenic scores. In this case, it should be verified that the genetic architecture of the trait is identical in the GWAS results and the analysis sample (e.g. using bivariate LD score regression [12]). Furthermore, most GWAS studies are conducted as a meta-analysis of summary statistics from various samples. Meta-analysis circumvents legal, practical, and logistic challenges that would have to be overcome to pool data from several providers on one central location is as statistically powerful as analyzing the raw data directly [51]. However, the meta-analysis approach makes it difficult to check if the same or closely related individuals have been included in several samples. It is currently unknown if and to which extent such hidden overlap between GWAS samples is a real issue. We explore in SI section 3.2.2 the consequences of a correlation between the measurement errors of the polygenic scores. As can be seen from Supplementary Figures 10 and 11, when the measurement errors for S_{y1} and S_{y2} are not independent, all methods produce biased estimates. When the correlation between the measurement errors of the PGS for y is small, GIV regression outperforms the other methods. When the correlation becomes moderate to strong, none of the methods produce accurate estimates. When the measurement errors in S_{y1} and S_{y2} are correlated with the measurement error in T, there is a region of small to moderate correlation strength in which EMR performs better than GIV regression or MR. When the correlation is strong, none of the methods produce accurate estimates.

5. The PGS for the outcome is uncorrelated with omitted inherited environmental factors that affect the outcome. An example of potential bias stemming from omitted inherited environmental factors would arise from the correlation between the PGS for height and the PGS for parent's height, which is correlated with parent's height, under conditions when parents get an environmental effect from height (e.g., higher pay for being taller) that affects the quality of the childhood health environment that could be correlated both with child's height and with child's EA. Violation of the exclusion restriction would be avoided by controlling for parental height, or for the parental resources that are consequences of parental height. More generally, however, there might be other causal pathways between parental genetic factors that affect the child's environment in ways that affect her EA (e.g., via the BMI of parents). One alternative strategy for blocking these pathways would be to construct and control for a parental PGS for the

child's EA. However, large family samples including biological parents and their offspring would be required for this. Such samples are still very rare and often not available in the public domain. Furthermore, in the absence of sufficiently large GWAS samples to estimate parental PGS with high accuracy, controlling for the observed parental PGSs would not be perfect, for the same reasons described throughout this paper in the context of a person's own PGS (though it would then in principle be possible to pursue a multiple indicator strategy for parental PGS as we do for the child's PGS). Whether it would be preferable to control for the parental PGS (and to use an additional indicator of the PGS to obtain consistent estimates of the effect of the parental PGS on the outcome) or for parental phenotypical characteristics that might affect a person's life course environment, or for the environmental characteristics themselves would depend upon whether the causal pathways are well-enough understood and whether sufficient information is available about a person's environment, her parent's phenotypical characteristics, or her parent's PGS for the person's outcome of interest in the analysis.

To assess the potential impact of bias from omitted inherited environmental factors that are correlated with the IVs, we carried out a final simulation where we assumed that a genetic marker of the parents (P_T) affects y net of S_y^* and T. We assumed a range of values for the effect of P_T allowing its effect to range from zero to the same size as T in our structural model (equation 3). Supplementary Figure 13 shows that GIV regression generally outperforms OLS, MR, and EMR, but that all methods produce substantial bias if the reduced form effect of P_T, which affects y entirely indirectly through omitted environmental factors, rivals the effect of T on y. In practice, the problem is unlikely to be this large; even if the indirect effects of parental genes through their effect on the child's environment are sizable, much of the bias can be removed from the estimation via controls for these consequential environmental variables or for the parental phenotypical characteristics that produce these environmental effects (e.g., parental education or income or height), or for the parental genotype (via a PGS for the genetic effects of parents on y) or through the use of data on twins that allows an effective control of parental genotype via the estimation of within-family regression models.

Empirical applications

We illustrate the practical use of GIV regression in a variety of important empirical applications using data from the Health and Retirement Survey (HRS) for 8,638 unrelated individuals of North-West European descent who were born between 1935 and 1945 (SI appendix).

The narrow-sense SNP heritability of educational attainment

First, we estimate the chip heritability of EA using GIV regression (see SI appendix).

The results are displayed in Table 1. The standard OLS estimate of the PGS only explains 4% of the variance in EA, which is similar to the results reported by the Social Science Genetic Association Consortium [27]. This is substantially lower than the 17.3% estimate of chip heritability (with a 95% confidence interval of +/- 4%) reported by [52] in the same data using genomic-relatedness-matrix restricted maximum likelihood (GREML).[7] Columns 2 and 3 show GIV regression results using one or the other of the two subset PGS scores as the covariate and as the instrumental variable. The 2SLS regression in column 2 gives an estimate of .364, while the 2SLS regression in column 3 gives an estimate of .454. These estimates imply 13% and 21% respectively as the estimates of SNP-based heritability, and so confidence intervals around these point estimates are consistent with the GREML estimate from [52]. In other words, GIV regression recovers the correct chip heritability from polygenic scores.[8]

The relationship between body height and educational attainment

Previous studies using both OLS and sibling or twin fixed effects methods have found that taller people generally have higher levels of EA [53, 54, 55]. They are also more likely to perform well in various other

[7] GREML yields unbiased estimates of SNP-based heritability that are not affected by attenuation, see [?].

[8] GIV regression is less efficient than GREML in obtaining unbiased estimates of the chip heritability. Nevertheless, we present these results as a proof of concept.

Table 1: Effects of the PGS on Educational Attainment in the HRS subsample

	(1) OLS	(2) IV1	(3) IV2
PGS EA Total	0.192***		
	(0.0166)		
PGS EA SSGAC		0.364***	
		(0.0421)	
PGS EA UKB			0.454***
			(0.0436)
Birth year	3.616	8.074	-1.107
	(22.20)	(22.69)	(23.73)
Birth year, squared	-3.583	-8.039	1.135
	(22.20)	(22.69)	(23.73)
Gender	-0.0829***	-0.0791***	-0.0821***
	(0.0164)	(0.0168)	(0.0176)
Mother's EA	0.252***	0.231***	0.253***
	(0.0202)	(0.0209)	(0.0216)
Father's EA	0.199***	0.183***	0.176***
	(0.0202)	(0.0210)	(0.0219)
N	2839	2839	2839

Note: Regression results of education in years (standardized) on the polygenic scores (standardized) and control variables. Standard errors are in parentheses. Coefficients for the 10 principal components are not shown. **significant at 0.1% level.

life domains, including earnings, higher marriage rates for men (though with higher probabilities of divorce), and higher fertility [56, 57, 58, 59, 60, ?]. The question is what drives these results. Can they be attributed to genetic effects that jointly influence these outcomes? Are there social mechanisms that systematically favor taller or penalize shorter individuals? Or are there non-genetic factors (e.g., the uterine and post-birth environments especially related to nutrition or disease) that affect both height and these life course outcomes? The literature on the relationship between height and EA has found evidence that the association arises largely through the relationship between height and cognitive ability, which may suggest that the height-EA association is driven largely by genetic association between height and cognitive ability. We use GIV regression with individual-level data from the HRS to clarify the influence of height on EA, and we compare these results with those obtained from OLS and from MR. In addition, we conduct a "negative control" experiment that estimates the causal effect of EA on body height (which should be zero). A complete description of the materials and methods is available in the SI Appendix.

GWAS summary statistics for height were obtained from the Genetic Investigation of ANthropometric Traits (GIANT) consortium [22] and by running a GWAS on height using the interim release of genetic data in the UK Biobank [61], which was not part of the GIANT sample. We refer to these as *Height_ GIANT* and *Height_ UKB*, respectively. GWAS summary statistics for EA were obtained from the Social Science Genetic Association Consortium (SSGAC). The most recent study of the SSGAC on EA used a meta-analysis of 64 cohorts for genetic discovery and the interim release of the UKB for replication [27]. We refer to these samples as *EA_ SSGAC* and *EA_ UKB*, respectively. There is an overlap in the cohorts between *Height_ GIANT* and *EA_ SSGAC*. To ensure independence of measurement errors in the PGS, whenever one of the two was used as regressor, we excluded the other as instrument and used a PGS from UK Biobank data instead.

The OLS results in Table 2 show that height (in meters) appears to have a strong effect on years of EA, with two additional centimeters in height generating one additional month of EA. MR appears to confirm the causal interpretation of the OLS result; indeed, the point estimate from MR is even slightly larger than from OLS. As discussed above, MR suffers from probable violations of the exclusion restriction. These violations could stem from the possibility that the some genes have direct effects on both height and EA (i.e. Type 1 pleiotropy).[9] They could also stem from the possibility that the PGS for height by itself is correlated with the genetic tendency for parents to have higher EA and income, and therefore a lower nutritional or disease risk for their children, who therefore are more likely to reach their full cognitive potential and have higher EA. Controlling for the PGS is an imperfect strategy for eliminating this source of endogeneity, because the bias in the estimated effect of the PGS score also biases the estimated effect of height (the omitted variable bias discussed earlier).

In contrast, estimates from GIV regression in Table 2 show both a considerably larger effect of the education PGS score on EA, and a small and statistically insignificant effect of height on EA. These results imply that the positive correlation between height and EA is not a causal relationship. Rather, the observed phenotypic correlation is primarily due to the genetic correlation between the two traits. Furthermore, our "negative control" using GIV regression finds no causal effect of EA on height, as expected. One might contrast our results also to those of [53], who found a correlation between the height and EA of Finnish monozygotic (MZ) twins. Silventoinen et al's study effectively controls for all genetic correlation between height and EA. However, their result would still suffer from endogeneity bias to the extent that the difference in MZ twin heights is related to intra-uterine or post-birth environmental differences that cause one twin to be taller and have higher cognitive or non-cognitive abilities than the other twin. GIV regression is arguably superior to twins fixed effects to the extent that these environmental variables are uncorrelated with the PGS for height once the PGS for EA is effectively controlled, making the combination of the PGS for height and the PGS for EA to be valid instruments.

Table 2: Regression of educational attainment on height in the Health and Retirement Study (HRS)

	(1)	(2)	(3)	(4)	(5)
	OLS	MR	EMR	GIV1	GIV2
Height	0.120***	0.130*	0.0523	0.000594	0.0697
	(0.0252)	(0.0599)	(0.0590)	(0.0609)	(0.0706)
PGS_EA_Total	0.189***		0.191***		
	(0.0165)		(0.0166)		
PGS_EA_SSGAC				0.364***	
				(0.0425)	
PGS_EA_UKB					0.447***
					(0.0438)
Birth year	1.825	0.0763	2.834	8.066	-2.086
	(22.11)	(22.63)	(22.16)	(22.72)	(23.63)
Birth year, squared	-1.790	-0.0439	-2.800	-8.030	2.115
	(22.11)	(22.63)	(22.16)	(22.72)	(23.63)
Gender	0.00707	0.0109	-0.0435	-0.0787	-0.0297
	(0.0250)	(0.0481)	(0.0473)	(0.0487)	(0.0559)
Mother's EA	0.247***	0.256***	0.250***	0.231***	0.250***
	(0.0201)	(0.0207)	(0.0203)	(0.0210)	(0.0217)
Father's EA	0.198***	0.220***	0.199***	0.183***	0.176***
	(0.0201)	(0.0205)	(0.0201)	(0.0210)	(0.0218)
N	2839	2839	2839	2839	2839

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Standard errors in parentheses. All variables have been standardized. EA is measured in years of schooling needed to obtain the highest achieved educational degree according to ISCED classifications. The first 10 principal components in the genetic data were included as control variables. *PGS_EA_UKB*: PGS for EA using UKB data. *PGS_EA_Total*: PGS for EA using GWAS meta analysis of UKB + SSGAC[27], excluding data from *23andMe* and HRS; *PGS_EA_SSGAC*: PGS for EA using meta analysis from [27], excluding data from *23andMe*, UKB, and HRS. MR and EMR use *PGS_Height_UKB* as instrument for height. GIV1 uses *PGS_Height_UKB and PGS_EA_UKB* as instruments for height and *PGS_EA_SSGAC.* GIV2 uses *PGS_Height_GIANT and PGS_EA_SSGAC* as instruments for height and *PGS_EA_UKB.*

[9] Results from [?] and [27] suggest a genetic correlation between height and EA of about 0.15.

Conclusion

Accurate estimation of causal relationships with observational data is one of the biggest and most important challenges in epidemiology and the social sciences - two fields of inquiry where many questions of interest cannot be adequately addressed with properly designed experiments due to practical or ethical constraints. Here, we have proposed a method that allows genetic data to be used for this purpose. Thinking of genetic data as a sort of naturally occurring experiment is appealing because the genotypes that arise from two mates are randomized by the process of meiosis. Thus, given that virtually all human traits are heritable to some extent, an individual's genotype could in principle be used to identify causal effects across a wide range of important scientific questions. Thanks to cheap and accurate genotyping technologies and growing insights into the genetic architecture of many traits via large-scale GWAS, this general idea becomes practically more and more feasible. In principle, it is this idea which underlies so-called Mendelian Randomization (MR)–a method suggested by epidemiologists that uses genetic data as instrumental variables.

The crucial identifying assumption of MR is that the genes which are used as instruments do not also affect the outcome through other causal pathways via so-called pleiotropic effects. In light of the widespread and often substantial genetic correlations between many traits, this assumption seems problematic. We have proposed a new strategy that we call genetic instrumental variable (GIV) regression, that eliminates or at least substantially reduces the bias of MR due to pleiotropy under a set of arguably more realistic assumptions. We have explored conditions where the assumptions underlying GIV regression will fail and conclude that GIV regression outperforms OLS and MR in a broad range of realistic scenarios.

The simulations described in the paper certainly do not cover all conceivable data generating processes, but they are nonetheless of considerable utility, we would argue, in assessing the performance of GIV regression. Analyses with real data demonstrate that GIV regression recovers estimates of the effect size of the outcome PGS that are consistent with alternative approaches to estimate the extent of narrow-sense heritability. Our analyses also provide reason to be cautious when using OLS or MR to estimate causal effects between variables that are known to be genetically correlated. Existing knowledge about the effects of epistasis, rare or dominant alleles, structural variants, or population structure provide good grounds to be cautiously optimistic that GIV regression provides an important tool for assessing causal effects when unmeasured genetic correlation is likely to be a serious issue. In particular, constant improvements in genotyping technology, increasing GWAS samples, and even better statistical methods to control for population structure in GWAS will make it less and less likely in the future that the assumptions underlying our approach will be seriously violated. Additional knowledge in this rapidly developing field will provide further guidance for assessing the extent of remaining bias in GIV regression estimates. The combination of new estimation tools and continued rapid advancements in genetics should provide a significant improvement in our understanding of the effects of behavioral and environmental variables on important socioeconomic and medical outcomes.

References

[1] McNeill PM (1993) *The Ethics and Politics of Human Experimentation.* (Cambirdge University Press).

[2] Stigler SM (2005) Correlation and causation: A comment. *Perspectives in Biology and Medicine* 48(1 Supplement):88–S94.

[3] Lawlor DA, Smith GD, Ebrahim S (2004) Commentary: The hormone replacement–coronary heart disease conundrum: is this the death of observational epidemiology? *International Journal of Epidemiology* 33(3):464–467.

[4] Plomin R (1999) Genetics and general cognitive ability. *Nature* 402(Supplement):25–29.

[5] Yang J et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42(7):565–569.

[6] Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88(1):76–82.

[7] Turkheimer E, Haley A, Waldron M, D'Onofrio B, Gottesman II (2003) Socioeconomic status modifies heritability of IQ in young children. *Psychological science* 14(6):623–628.

[8] Polderman TJC et al. (2015) Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* 47:702–709.

[9] Conley D (2016) Socio-Genomic Research Using Genome-Wide Molecular Data. *Annual Review of Sociology* 42(1):275–299.

[10] 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.

[11] Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW (2013) Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* 14(7):483–495.

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

[13] Wooldridge JM (2002) *Econometric Analysis of Cross Section and Panel Data.* (Massachusetts Institute of Technology, Cambridge, MA), pp. 83–113.

[14] Smith GD, Ebrahim S (2003) 'mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 32(1):1–22.

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

[16] Pickrell J (2015) Fulfilling the promise of Mendelian randomization. *bioRxiv* p. 018150.

[17] Davey Smith G (2015) Mendelian randomization: a premature burial? *bioRxiv* p. 021386.

[18] Hahn J, Hausman J (2003) Weak instruments: Diagnosis and cures in empirical econometrics. *The American Economic Review* 93(2):118–125.

[19] Murray MP (2006) Avoiding invalid instruments and coping with weak instruments. *The journal of economic perspectives* 20(4):111–132.

[20] Hamer D, Sirota L (2000) Beware the chopsticks gene. *Mol. Psychiatry* 5(1):11–13.

[21] Welter D et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42(Database issue):D1001–1006.

[22] 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.

[23] Locke AE et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518(7538):197–206.

[24] Ripke S et al. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511(7510):421–427.

[25] Lambert JC et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* 45(12):1452–1458.

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

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

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

[29] Rietveld CA et al. (2014) Replicability and robustness of GWAS for behavioral traits. *Psychol. Sci.* 25(11):1975–1986.

[30] Bulik-Sullivan BK et al. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47(3):291–295.

[31] Loh PR et al. (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* 47(3):284–290.

[32] Purcell SM et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460(7256):748–752.

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

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

[35] 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.

[36] Bowden J, Davey Smith G, Burgess S (2015) Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44(2):512–25.

[37] van Kippersluis H, Rietveld CA (2017) Pleiotropy-robust mendelian randomization. *International Journal of Epidemiology.*

[38] Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: An empiricist's companion.* (Princeton University Press).

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

[40] 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.

[41] Wooldridge JM (2013) *Introductory Econometrics: A Modern Approach.* (Cengage Learning), 5th edition edition, pp. 317–323.

[42] 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.

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

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

[45] Wooldridge JM (2013) *Introductory Econometrics: A Modern Approach.* (Cengage Learning), 5th edition edition, pp. 83–92.

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

[47] Yang J et al. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47(10):1114–20.

[48] Plomin R et al. (2013) Common DNA markers can account for more than half of the genetic influence on cognitive abilities. *Psychological Science* 24(4):562–568.

[49] Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39(7):906–913.

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

[51] Lin DY, Zeng D (2010) Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. *Genetic epidemiology* 34(1):60–66.

[52] de Vlaming R et al. (2016) Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies, (Cold Spring Harbor Labs Journals), Technical report.

[53] 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.

[54] 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.

[55] Case A, Paxson C (2006) Stature and status: Height, ability, and labor market outcomes, (National Bureau of Economic Research), Technical report.

[56] Heineck G (2005) Up in the skies? The relationship between body height and earnings in Germany. *Labour* 19(3):469–489.

[57] Heineck G (2009) Too tall to be smart? The relationship between height and cognitive abilities. *Economics Letters* 105(1):78–80.

[58] Schick A, Steckel RH (2015) Height, Human Capital, and Earnings: The Contributions of Cognitive and Noncognitive Ability. *Journal of Human Capital* 9(1):94–115.

[59] Weitzman A, Conley D (2014) From Assortative to Ashortative Coupling: Men's Height, Height Heterogamy, and Relationship Dynamics in the United States, (National Bureau of Economic Research), Technical report.

[60] Kanazawa S (2005) Big and tall parents have more sons: further generalizations of the Trivers–Willard hypothesis. *Journal of Theoretical Biology* 235(4):583–590.

[61] Marchini J et al. (2015) Genotype Imputation and Genetic Association Studies of Uk Biobank: Interim Data Release, Technical report.

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

Supplementary Information

Thomas A. DiPrete, Casper Burik, & Philipp Koellinger

May 3, 2017

Contents

1 Introduction

The Supplementary Information for this article consists of four sections. In section 2, we provide technical details of Genetic Instrumental Variables (GIV) regression. In section 3, we describe a set of simulations to illustrate and explore how GIV regression performs in finite samples when the model assumptions are satisfied or violated in various ways. Section 4 describes the data and methods used for our empirical examples. In section 5, we provide supplementary information about the empirical examples described in the article.

2 Genetic Instrumental Variables (GIV) regression

2.1 Estimating narrow-sense SNP heritability from polygenic scores.

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 exogenous control variables, and S_y^* is a summary measure of genetic tendency for y, then one can write

$$y = \alpha + X\beta + \gamma S_y^* + \epsilon \tag{1}$$

where, for example, y is educational attainment. 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^* 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 are the predicted values for y in the equation

$$y = G\zeta_y + \varepsilon_y \tag{2}$$

where G is an $n \times m$ matrix of genetic markers, and ζ is the $m \times 1$ vector of LD-adjusted 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^* would be expressed by the PGS for y, and the marginal R^2 of S_y^* 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 ([7]; [2]; [8]). We refer to the estimate of the PGS from available GWAS data as S_y, and substitute S_y for S_y^* 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 [9] and [10], 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_g^2 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% [11].

It has long been understood that multiple indicators can, under certain conditions, provide a strategy to correct regression estimates for attenuation from measurement error ([12]; [13]). 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 ([14]; [15]): (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 [14]. 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) [16].

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^*, with lower predictive accuracy. However, it also produces an IV for S_y that has desirable properties. Formally, we let $\hat{\zeta_{y1}}$ be the estimated coefficient vector for ζ_y in equation 2 from the first training sample, and $\hat{\zeta_{y2}}$ be the coefficient vector estimated from the second training sample. It follows then that

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

for the j-th genetic marker, where u_{y1} and u_{y2} are asymptotically normally distributed errors with $E(u_{y1j}) = E(u_{y2j}) = 0$ and $V(u_{y1j}) = V(u_{y2j}) = \sigma_\varepsilon^2 n^{-1}/var(x_j)$, and where x_j is the observed number of reference alleles for location j. Because the two discovery samples are non-overlapping, u_{y1} and u_{y2} 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,

$$\begin{aligned}S_{y1} &= S_y^* + v_1 = G\zeta_y + Gu_{y1} = S_y^* + Gu_{y1} \\ S_{y2} &= S_y^* + v_2 = G\zeta_y + Gu_{y2} = S_y^* + Gu_{y2}\end{aligned} \tag{4}$$

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

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

As can be seen from equation 5, the PGS S_{y1} is correlated with the error term via its correlation with Gu_{y1} 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} meets the two requirements to be a valid instrument for S_{y1}, namely that it is correlated with S_{y1} (through their mutual dependence on S_y^*) and uncorrelated with the disturbance term, because neither S_y^* $(= G\zeta_y\ )$ nor Gu_{y2} are correlated with Gu_{y1}. Therefore, the covariance of Gu_{y1} and Gu_{y2} is

$$\begin{aligned}
&Cov(Gu_{y1}, Gu_{y2}) = E([Gu_{y1}][Gu_{y2}]) \\
&= E\left\{\sum_{i=1}^{m} g_i^2 u_{y1i}u_{y2i} + \sum_{i=1}^{m}\sum_{j\neq i}^{m} g_i g_j u_{y1i}u_{y2j}\right\} \\
&= \sum_{i=1}^{m} E(g_i^2)E(u_{y1i}u_{y2i}) + \sum_{i=1}^{m}\sum_{j\neq i}^{m} E(g_i g_j)E(u_{y1i}u_{y2j}) \\
&= 0 \qquad (6)
\end{aligned}$$

with equation (6) true because of the fact that u_{y1i} and u_{y2j} – being random measurement error – are independent of the genetic markers and uncorrelated with each other, under the assumption that equation (2) is the correct specification of the relationship between G and y.[1] It follows, therefore, that the IV S_{y2} is uncorrelated with the error term in equation 5, i.e.,

$$\begin{aligned}
plim\,\frac{1}{n}\sum_i (S_{y2})_i(\epsilon_i - (Gu_{y1})_i) &= plim\,\frac{1}{n}\sum_i (S_y^* + Gu_{y2})_i(\epsilon_i - (Gu_{y1})_i) \\
&= plim\,\frac{1}{n}\sum_i \left[(S_y^*)_i\epsilon_i + (S_y^*)_i(Gu_{y1})_i + (Gu_{y2})_i\epsilon_i + (Gu_{y1})_i(Gu_{y2})_i\right] \\
&= 0
\end{aligned}$$

At the same time, S_{y2} is correlated with S_{y1} through their common dependence on S_y^*. Under the assumptions that X is uncorrelated with ϵ net of S_y^*, and that S_y^* is uncorrelated with ϵ net of X, then S_{y2} is a valid IV for the estimation of γ in equation 5.

The above derivation assumes that the true coefficients of the genetic markers in G do not vary in the population. More generally, we might assume that the

[1] This conclusion assumes that the two PGS are estimated from the same population. In principle, the PGS for a trait could vary across sub-populations. Using a PGS from one subpopulation as an instrument for a PGS from another subpopulation could cause a violation of the exclusion restriction. This potential problem is solved if the two scores are estimated from randomly chosen subsamples of a single GWAS sample after randomly excluding related individuals so that the final sample consists only of unrelated individuals. This can be done using the UK Biobank.

population consists of a finite number of (possibly latent) groups, $k = 1, ., , , K$ with the kth group having the polygenic score S^*_{yk}. 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 [14]. 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

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

where $P = \{p_1, p_2, ..., p_K\}$ is the group composition that defines population P and Δ_{yk} 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

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

where S^*_{yik} is the true PGS for trait y for individual i in group k, and where $\bar{S}_{y1iP}$ 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 $\bar{S}_{y2P}$ is the IV, then $\bar{S}_{y2P}$ is correlated with Δ_{yk} 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.[2]

2.2 Reducing bias due to genetic correlation between exposure and outcome

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

$$y = \alpha + \delta T + X\beta + (\gamma S_y^* + \epsilon) \tag{7}$$

where, for example, y could be educational attainment and T could be body height. 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 (7). If S_y^* is not observed, it is part of the disturbance term. Equation (7) is written without any interaction terms involving S_y^* and T, implying that while the effect of T may vary with S_y^*, δ is a (variance-weighted) average effect of T across values of S_y^* [14]. If the same (uncontrolled for) genetic tendencies that affect the outcome variable also affect or are otherwise correlated with T (e.g., if T is influenced by parental genes that are correlated with S_y^*), then $\hat{\delta}$ will be a biased estimate of the effects of T.

In standard MR, a measure of genetic tendency (S_T) for a behavior of interest (T in equation 7) 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, is exogenous ([18];[15]). Implicitly, the true PGS for y is in the error term, as is shown in equation 7. One such example 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

$$y = \alpha + \delta\hat{T} + X\beta + \{\epsilon + \gamma S_y^* + \delta(T - \hat{T})\} \tag{8}$$

[2] 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 [17].

The problem with this approach is that the PGS for height will typically fail to satisfy the exclusion restriction because of so-called Type 1 pleiotropy [16]: the genetic variants that predispose individuals to be tall may also directly increase the predisposition for higher educational attainment [19, 20] (e.g. via healthy cell growth and metabolism). This problem is not solved even if we could use the true PGS S^*_T as the IV.

The multiple indicator strategy described above provides potentially attractive approaches for addressing the bias in MR. If the genetic propensity for y could be directly controlled in the regression, MR would provide less biased estimates of the effect of T. We refer to the combined use of S_{y1} as a control and S_T as an IV as "enhanced Mendelian Randomization" (EMR). However, controlling for S_{y1} as a proxy for S^*_y is not adequate, both because it leaves a component of S^*_y in the error term which causes the exclusion restriction assumption of MR to fail, and because the bias in the estimated coefficient of S_{y1} also produces bias in the estimated coefficient of T. The bias arising from the use of a proxy for S^*_y as a control variable in OLS regression is a form of omitted variable bias. To see this, assume that S_T is a valid instrument for T if S^*_y were measured and controlled. In this case, the second stage of 2SLS would involve the substitution of $\hat{T}$ for T, and the regression would give a consistent estimate of δ if S^*_y were observed, i.e.

$$\begin{aligned} y &= \alpha + \delta \mathrm{T} + \gamma S^*_y + \epsilon \\ &= \alpha + \delta \hat{T} + \gamma S_{y1} + \{\epsilon - \gamma v_1 + \delta(T - \hat{T})\} \end{aligned} \tag{9}$$

where

$$v_1 = G u_{y1}$$

and where (for simplicity) we drop other covariates from the model.[3] If v_1 is omitted from the regression, then the bias in δ and γ is equal to the product of γ and the regression coefficients of the regression of v_1 on $\hat{T}$ and S_{y1}.
More generally, if a set of variables z is omitted from a regression of y on a vector X , then

$$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + Z\lambda + \epsilon)$$

and

$$\begin{aligned} E(\hat{\beta}|X) = \beta + (X'X)^{-1}X'Z\lambda,\ i.e., \\ Bias = (X'X)^{-1}X'Z\lambda \end{aligned} \tag{10}$$

[3] We could, for example, imagine replacing each of the variables in equation 8 with the residual of this variable from an OLS regression of that variable on the variables in X.

where X and Z are the matrices of included and omitted variables. The expression $(X'X)^{-1}X'Z$ gives the matrix of coefficients from regressions of each of the omitted variables on the included variables, and λ is the vector of coefficients of the variables in z in the regression of y on x and z. If z consists of a single omitted variable, then $E(\hat{\beta}|X) = \beta + (X'X)^{-1}X'z\lambda$, and $(X'X)^{-1}X'z$ is the vector of estimated regression coefficients of z on the included variables x.

Violation of the exclusion restriction in EMR due to genetic correlation is potentially solved (or at least is less severe) when an additional indicator of the PGS for y, i.e., S_{y3}, is used to instrument simultaneously both S_{y1} and T in equation 1. The practical problem with using two indicators of S^*_y as the sole instruments is that their mutual correlation will be relatively high (depending on their reliability) and they are weak instruments for T. Instead, as a practical – and, we will argue, effective– strategy, the best solution is arguably to use S_T along with S_{y2} (or S_{y2} and S_{y3}) as instruments for S_{y1} and T. S_T will still violate the exclusion restriction to the extent that it is correlated with v_1. However, the extent of the violation will be reduced by the presence of S_{y1} in the regression.

Arguably a strategy that both reduces the correlation between S_T and ϵ (through the inclusion of S_{y1} in the model) and eliminates or greatly reduces omitted variable bias through the inclusion of an instrument for S_{y1} in the first stage equation will outperform MR in the estimation of a consistent effect of T that is purged of genetic correlation. As noted above, if not all relevant genetic effects are contained in the PGS (e.g. interaction effects, structural variants, or rare alleles may be missing given currently available GWAS data), the PGS instruments above will not perfectly satisfy the exclusion restriction to the extent that S^*_y is correlated with the omitted genetic variables. However, the above approach would generally be expected to reduce bias due to genetic correlation, given that a large fraction of heritability can be attributed to linear effects of common SNPs that are well tagged by currently available genotyping arrays [8, 21, 9, 22]).

This argument can be made more formally. MR assumes a regression of y on covariates and T, with S^*_y in the error term of equation 7. If, as above, we omit the covariates X (or, more precisely, residualize y, T, and S_{y1} from their dependence on X), then the bias in MR equals

$$Bias(\hat{\delta}^{(MR)}) = \eta_1^{(MR)}\lambda_1^{(MR)} + \eta_2^{(MR)}\lambda_2^{(MR)}$$

where coefficient $\eta_1^{(MR)}$ is the coefficient of $\hat{T}$ in the regression of S^*_y on $\hat{T}$, $\lambda_1^{(MR)}$ is the effect of S^*_y on y, net of $X$and T, and the second term is ignorable because $\hat{T}$ is orthogonal to its residual.

If we estimate δ using EMR, the omitted variables are now v_1 (instead of S_y^*) and $(T - \hat{T})$, and the bias in the estimator for δ is

$$Bias(\hat{\delta}^{EMR}) = \eta_{11}^{(EMR)}\lambda_1^{(EMR)} + \eta_{12}^{(EMR)}\lambda_2^{(EMR)}$$

where the first term is the product of $\eta_{11}^{(EMR)}$ (the coefficient of $\hat{T}$ in the regression of v_1 on $\hat{T}$ and S_{y1}) and coefficient $\lambda_1^{(MR)}$, which is the effect of v_1 on y, net of X, $\hat{T}$, and S_{y1}. The second term is the product of the coefficient of $\hat{T}$ in the regression of $T - \hat{T}$ on $\hat{T}$ and S_{y1} and the coefficient of $T - \hat{T}$ on y, net of X, $\hat{T}$, and S_{y1}. As with MR, the second term is ignorable.

Lastly, we consider GIV regression. The second stage of GIV regression is

$$\begin{aligned} y &= \alpha + \delta \mathrm{T} + \gamma S_y^* + \epsilon \\ &=\alpha + \delta\hat{T} + \gamma S_{y1} + \{\delta(T - \hat{T}) + \epsilon - \gamma v_1\} \\ &=\alpha + \delta\hat{T} + \gamma\hat{S}_{y1} + \{\epsilon + \gamma(S_y^* - \hat{S}_{y1}) + \delta(T - \hat{T})\} \end{aligned} \tag{11}$$

In GIV regression, the bias of δ is given by

$$Bias(\hat{\delta}^{GIV}) = \eta_{11}^{(GIV)}\lambda_1^{(GIV)} + \eta_{12}^{(GIV)}\lambda_2^{(GIV)} \tag{12}$$

The first term is the product of $\eta_{11}^{(GIV)}$, which is the coefficient of $\hat{T}$ in the regression of $(S_y^* - \hat{S}_{y1})$ on $\hat{T}$ and $\hat{S}_{y1}$ multiplied by λ_1, the effect of $(S_y^* - \hat{S}_{y1})$ on y. The second term is the coefficient of $\hat{T}$ in the regression of $(T - \hat{T})$ on $\hat{T}$ and $\hat{S}_{y1}$ multiplied by $\lambda_{2,}$, the effect of $(T - \hat{T})$ on y. As with MR and EMR, the second term is ignorable.

The major difference between the bias in MR and the bias in GIV regression stems from the relative sizes of $\eta_{11}^{(MR)}$ and $\eta_{11}^{(GIV)}$. In general, we expect that the correlation between the true PGS for y and that part of T which is predicted by the observed PGS for T (which drives $\eta_{11}^{(MR)}$) will be stronger than is the correlation between the residual PGS for y (i.e., the difference between the true PGS for y and the predicted PGS from S_{y2}) and that part of T which is predicted by the observed PGS for T and S_{y2} (which drives $\eta_{11}^{(GIV)}$). Therefore, in general, we expect a lower bias in the estimate of δ using GIV regression than using MR. We find support for our expectation in the simulations and empirical analyses discussed below.

2.3 Gene-environment interactions

We next generalize equation (7) to the case of gene-environment interactions, where the effect of T varies with the PGS. In principle, these interactions could be extremely complicated and so for practical reasons, we focus her on obtaining plausible estimates of the linear interaction between S_y^* and T. We rewrite equation (7) as

$$\begin{aligned}
y &= \alpha + \delta_1 T + X\beta + \left(\gamma S_y^* + \delta_2 T S_y^* + f(G) + \epsilon\right) \\
&= \alpha + \delta_1 T + X\beta + \left(\gamma \left(S_{y1} - G u_{y1}\right)\right. \\
&\qquad\qquad\qquad \left. + \delta_2 T \left(S_{y1} - G u_{y1}\right) + f(G) + \epsilon\right) \\
&= \alpha + \delta_1 T + X\beta + \gamma S_{y1} + \delta_2 T S_{y1} \\
&\qquad\qquad\qquad + \left(\epsilon + f(G) - G u_{y1} - \delta_2 T G u_{y1}\right).
\end{aligned} \tag{13}$$

Now there are three endogenous variables, T, S_{y1}, and TS_{y1}. Also the disturbance term has now been elaborated to include a term that is a function of T, and so an additional instrument, S_{y4}, is needed. As before, S_{y4} will be a valid instrument for the same reasons as in equation (6) to the extent that problems deriving from correlation between the instrument and $f(G)$ are relatively small.

3 Simulations

We describe the basic simulation model and the data generating process in 3.1. Section 3.2 studies various violations of the model assumptions.

3.1 Standard model

Our interest lies in studying the effect of a treatment T on an outcome y. Both T and y are partly heritable and the genetic propensities of individuals for both variables are summarized by polygenic scores, S_T^* and S_y^*. These polygenic scores are not observed directly. Instead, they are empirically approximated from genome-wide association study (GWAS) results for T and y with finite sample sizes. The estimated regression coefficients from the GWAS are used as weights to construct the scores in an independent sample with genetic data. Since the GWASs were conducted in finite sample sizes, the estimated betas will have standard errors greater than zero, which implies that the constructed PGS will be noisy proxies of the true PGS [2]. We denote the actually available (noisy) PGS for T and y for individuals $i = 1, ..., N$ as S_{Ti} and S_{yi}, respectively.

The data generating process is as follows:

$$y = \gamma S_y^* + \delta T + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{14}$$

$$T = \beta S_T^* + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2) \tag{15}$$

$$S_{yi} = S_y^* + \varepsilon_{yi}, \qquad \varepsilon_{yi} \sim \mathcal{N}(0, \sigma_{yi}^2) \tag{16}$$

$$S_{Ti} = S_T^* + \varepsilon_{Ti}, \qquad \varepsilon_{Ti} \sim \mathcal{N}(0, \sigma_{Ti}^2) \tag{17}$$

S_y^* and S_T^* are drawn from a multivariate normal distribution with non-zero covariance and have a correlation of ρ_G. We assume that all measurement errors are independent of each other. Both y and T are standardized and have a mean of 0 and a variance of 1. Furthermore, the true polygenic scores (S_T^* and S_y^*) are also assumed to be standardized. This implies the following variances for T and y:

$$\text{var}(T) = \beta^2 + \sigma_\eta^2 = 1 \tag{18}$$

$$\text{var}(y) = \gamma^2 + \delta^2\beta^2 + 2\gamma\delta\beta\rho_G + \sigma_\epsilon^2 + \delta^2\sigma_\eta^2 + 2\delta\sigma_{\epsilon,\eta} = 1 \tag{19}$$

where $\sigma_{\epsilon,\eta}$ is the covariance between ϵ and η. One can calculate from this variance decomposition what β and γ should be in terms of their heritability (h_T^2 and h_y^2) and their genetic correlation:

$$\beta = \sqrt{h_T^2} \tag{20}$$

$$\gamma = -\delta\beta\rho_G + \sqrt{\delta^2\beta^2\rho_G^2 - \delta^2\beta^2 + h_y^2}. \tag{21}$$

Similarly, σ_ϵ^2 and σ_η^2 can be expresses as

$$\sigma_\eta^2 = 1 - h_T^2 \tag{22}$$

$$\sigma_\epsilon^2 = (-\delta\rho_e\sigma_\eta + \sqrt{\delta^2\rho_e^2\sigma_\eta^2 - \delta^2\sigma_\eta^2 - 1 - h_y^2})^2 \tag{23}$$

where ρ_e is the correlation between ϵ and η.

In practice, an important concern is endogeneity and bias due to unobserved environmental factors that jointly influence T and y (i.e., $\rho_e \neq 0$). Our simulations cover two broad scenarios. In one scenario, we assume that endogeneity in the naive OLS estimates arises solely due to genetic correlations between T and y (i.e., $\rho_G \neq 0$). In this case, the endogeneity problem would be solved if the true PGS S_y^* would be known. We simulate this scenario of entirely genetically caused endogeneity by drawing ϵ and η from independent distributions. In the second scenario, there is "additional endogeneity" due to unobserved non-genetic effects that matter for both

T and y (i.e., $\rho_e \neq 0$). We simulate this more realistic scenario by drawing ϵ and η from a multivariate normal distribution with $\rho_e = 0.4$.

The variances of the error terms σ^2_ϵ and σ^2_η can be derived using the model of Dudbridge [2]. In the original Dudbridge model, the polygenic scores are not standardized and have a variance equal to the heritability of the trait. Specifically, it is assumed that

$$y = G\zeta + u, \tag{24}$$

where y is the phenotype, ζ is a vector of effect sizes and G is a matrix with standardized genetic markers. The effect sizes are assumed to be randomly distributed across the genome. Therefore, the true polygenic score and the estimated polygenic score are defined as

$$S^*_y = G\zeta \tag{25}$$
$$S_y = G\hat{\zeta}. \tag{26}$$

The variance of this score is equal to

$$h^2_y = \text{var}(S^*_y) = \sum_{i=1}^{m} var(\zeta_i G_i) \approx m\text{var}(\zeta_i) \tag{27}$$

where m is the number of independent genetic markers. The variance of the estimation error can now be written as

$$\text{var}(S_y - S^*_y) = \text{var}(G(\hat{\zeta} - \zeta)) \approx m\text{var}(\hat{\zeta} - \zeta) = m\frac{1 - \text{var}(\zeta_i)}{n} = m\frac{1 - h^2_y/m}{n}. \tag{28}$$

Here n is the sample size used to estimate $\hat{\zeta}$. Since we standardized our polygenic scores in our simulations, we divide this variance by the heritability such that our true polygenic score has a variance of one. Therefore, the variance of the error in the polygenic score for observation i and trait k is

$$\sigma^2_{ki} = \frac{m_{ki}}{h^2_k}\frac{1 - h^2_k/m_{ki}}{n_{ki}}. \tag{29}$$

For our simulations, we use parameters in a range close to the empirical values we estimated in the Health and Retirement Study (see section 4). For all scores we assume m_{ki} equal to 300,000. The GWAS discovery sample sizes are assumed to vary between 200,000 and 2,000,000 per trait. Several recently published GWAS already had sample sizes exceeding 200,000 [23, 24, 25, 26] and 2,000,000 will be a realistic sample size for many traits in the near future. For models that include multiple polygenic scores per trait, we assume that the GWAS discovery sample was divided into equal parts per score. The size of our estimation sample is set to 8,600 (again similar to the sample size of the HRS, see section 4). We assume chip heritabilities of $h_y^2 = 0.2$ (similar to results reported for educational attainment [9, 17]) and $h_T^2 = 0.55$ (similar to results reported for body height [21]). For δ, we assume 0.15, which is the phenotypic correlation between height and educational attainment in the HRS. The genetic correlation between height and educational attainment has been estimated to be 0.15 by [26]. Hence, we use this value for ρ_G. Note that this implies we assume that the entire genetic correlation between the traits is due to Type 1 pleiotropy, i.e. we assume that all genes that are associated with both phenotypes have direct effects on both rather than some of the genes having cascade effects (e.g. a direct effect on height that triggers higher educational attainment, which shows up in the genetic correlation estimate). Surely, this is a conservative assumption to the disadvantage of classic MR. However, since there is no way to exclude the possibility that Type 1 pleiotropy is underlying the observed genetic correlation between the two traits, we prefer this conservative assumption.

All simulations were written in MatLab and the code is posted on Github (https://github.com/cburik/GIVsim). Each simulation is conducted as follows:

i. Calculate the simulation parameters from the input parameters.

ii. Draw the true PGSs from a multivariate normal distribution.

iii. Draw the error terms and measurement errors from their respective distributions;

iv. Compute the "measured" PGS, T, and y from equations 14-17;

v. Estimate $\hat{\beta}$ and $\hat{\gamma}$ in the simulated data and save the estimation results.

vi. Repeat steps ii-v to create a distribution of estimated effect sizes and their confidence intervals for each method.

We estimate the coefficients:

1. using OLS.

2. in $y = \delta\hat{T} + \epsilon$ using 2SLS and S_T as the IV (i.e., Mendelian Randomization - MR).

3. using 2SLS with S_T as the instrument for T, treating S_{y1} as exogenous (i.e., Enhanced Mendelian Randomization - EMR).

4. using 2SLS with S_{y2} as the instrument for T, treating S_{y1} as exogenous (i.e. EMR with an alternative instrument).

5. using 2SLS with S_{y2} and S_{y3} as instruments (i.e., Genetic Instrumental Variables regression - GIV).

6. using 2SLS with S_{y2} and S_T as instruments (i.e., Genetic Instrumental Variables regression - GIV).

7. using 2SLS with S_{y2}, S_{y3}, and S_T as instruments (i.e., Genetic Instrumental Variables regression - GIV).

The results of the simulations are shown in Supplementary Figures 1-14. The results of the standard model (with valid assumptions) are depicted in Supplementary Figures 1 to 6.

Supplementary Figures 1-3 shows results from the relatively simple scenario of genetic endogeneity only ($\rho_g = 0.15$, but $\rho_e = 0$). In this case, we would not need an instrument for T if the true PGS for y (S_y^*) would be known or if measurement error in S_y^* would be dealt with. This is also apparent in this figure – when the GWAS sample becomes large enough, the OLS estimates converge to the true coefficients. We see the same tendency for EMR (method 3) in Supplementary Figure 1, but it performs slightly worse then OLS. MR estimates (Supplementary Figures 1 and 2) are biased for all sample sizes due to the omitted effect of S_y^* and the correlation of the instrument S_{y1} with that omitted variable. Supplementary Figure 1 also reports estimates for GIV regression using the 6th method, using S_{y2} and S_T as instruments. This version of GIV regression outperforms the other methods and provides consistent estimates even for GWAS samples of 200,000 observations that are split into two samples of 100,000 to create two indicators of the score. The variance of the estimates is larger compared to the other methods, but decreases with total GWAS sample size.

Supplementary Figure 2 shows the same results, but now with GIV regression estimates using the 5th method that uses S_{y2} and S_{y3} as instruments and EMR estimates using the 4th method that uses S_{y2} as instrument. In this version, the variance of the GIV regression estimates is so large that the confidence bounds are outside of the figure. The high correlation between S_y^1, S_y^2 and S_y^3 implies large standard errors of the estimated coefficients due to multicollinearity. In other words, the instruments do not contain enough additional information to get precise estimates for S_y^* and T. EMR using method 4 also does not perform as well as EMR using method 3.

Supplementary Figure 3 compares the OLS estimates with all three GIV regression methods (5, 6 and 7). Again, it is clear that method 5 yields very imprecise results. Methods 6 and 7 perform equally well. However, method 7 requires the construction of an additional polygenic score. This additional effort does not seem to be justified compared to the results from method 6. Thus, we recommend method 6 for most practical applications.

Supplementary Figures 4, 5 and 6 show simulation results of the same methods, but now for the more complex scenario with additional endogeneity ($\rho_g = 0.15$, $\rho_e > 0$). As can be seen from Supplementary Figure 4, the OLS estimates are biased for the effects of both S_y^* and T even if the PGS was based on large GWAS samples. As the GWAS sample size increases, the estimate for S_y^* converges towards the true value, while the effect of T remains systematically overestimated due to unobserved environmental effects. The estimates of MR are also biased in this scenario due to the omitted effect of S_y^*. EMR estimates are likewise biased. However, EMR estimates of T are much closer to the true value than OLS and MR estimates. Furthermore, the EMR estimates converge towards the true coefficients as the GWAS sample sizes increase towards infinity. Furthermore, Supplementary Figure 4 also shows that the GIV regression estimates using method 6 are very close to being consistent for all GWAS sample sizes. Supplementary Figures 5 and 6 show that methods 4 and 5 again perform very poorly. Furthermore, we find no performance increase of method 7 compared to our preferred method 6.

3.2 Simulations of violated assumptions

We now turn to a set of simulations that systematically violate the identifying assumptions of GIV regression. The goal of this exercise is to explore how sensitive GIV regression reacts to these violated assumptions in finite sample sizes. Our simulations focus on the best performing GIV regression and EMR variants from section 3.1 (i.e., methods 6 and 3). We add OLS and standard MR estimates as benchmarks.

3.2.1 Skewness and kurtosis

Since GIV regression is a instrumental variables method, theoretical proofs of consistency rely on the central limit theorem. Yet, the more relevant question in practice is how sensitive the method reacts to skewness and kurtosis in finite sample sizes. To explore this question, we simulate samples with different degrees of skewness and kurtosis by drawing variables from a normal distribution using Fleishman's power method[27, 28] , keeping the GWAS sample size fixed at $n = 1,000,000$. The other parameters remained the same as in the standard model with additional endogeneity ($\rho_g = 0.15$, $\rho_e = 0.4$).

We investigate three different models. First, we add skewness to y via ϵ in equation 14. Second, we add skewness to T via η in equation 15. Note that if T is skewed, y will also become skewed. Third, we simulate a model with kurtosis in both y and T via ϵ and η.

For the first and second scenario, we used a kurtosis of 3 (corresponding to a normal distribution), unless that was not possible. Not all combinations of kurtosis and skewness are attainable, in the case of high skewness more kurtosis is needed. In those cases, we used the minimum kurtosis needed to obtain a certain skewness. The minimum kurtosis is found via the formula from [28]:

$$k = 1.7735511 + 1.6410373s^2$$

where k is the minimal amount of kurtosis needed and s is skewness.

Supplementary Figures 7 and 8 present the results for scenarios 1 and 2, with skewness in T and y, respectively. In both cases, it appears that only the OLS estimates are affected by skewness, while GIV regression estimates remain consistent. Supplementary Figure 9 shows the results for scenario 3 with kurtosis. Again, GIV regression estimates remain consistent, while kurtosis also does not diminish the performance of the other methods relative to the base-line scenario with normally distributed variables.

3.2.2 Dependent measurement errors

One of the key assumptions of GIV regression is independence of the measurement errors of the polygenic scores. Specifically, the measurement errors of the scores used as instruments ($\varepsilon_{y2}, \varepsilon_T$) should be independent from the score used as regressor (ε_{y1}). The independence of the measurement errors is violated if there is an overlap in the GWAS samples used to construct the different scores and the correlation between$\varepsilon_{y1}$and ε_{y2} will depend on the extent to which there is sample overlap. Furthermore, if the GWAS samples used to construct S_{y1} and S_T overlap, then the strength of the correlation between the measurement errors also depends on the extent to which the outcome variables (y and T) are correlated.

We simulate a violation of the independent measurement errors assumption by drawing the measurement errors from a multivariate normal distribution using a non-zero correlation between them. We fixed the GWAS sample size at $n = 1,000,000$ in these simulations and varied the correlation between the measurement errors. Again, there are different versions of this model. In the first scenario, only the measurement errors in S_{y1} and S_{y2} (ε_{y1} and ε_{y2}) are correlated with each other. In the second scenario, the measurement error of S_T (ε_T) is also correlated with the others.

The results of the first scenario are shown in Supplementary Figure 10. GIV regression outperforms the alternative methods for small to moderately correlated measurement errors ($\rho < 0.5$). As the correlation of the measurement errors increases, GIV regression starts to underestimate the genetic effects on y due to the attenuation bias. At the same time, GIV begins to overestimate the effect of the T, but only to a small extent. The GIV regression estimates of the treatment remain much closer to the true effect than the OLS estimates, even for severe violations of the independent error assumptions. However, for very strong violations (i.e. for very strong sample overlap between the samples used to construct S_{y1} and S_{y2}), GIV regression performs slightly worse than MR and EMR.

Supplementary Figure 11 displays the results of the second scenario, now with two invalid instruments (S_{y2} and S_T). Again, more strongly correlated measurements errors induce a stronger bias in the GIV regression estimates. However, in contrast to the first scenario, GIV regression now underestimates both the genetic effect and the effect of the treatment. None of the displayed estimation methods get anywhere close to the true parameters in the case of strongly correlated measurement errors.

3.2.3 Missing genetic variants

We simulate a situation where not all genetic variants are captured by the polygenic scores. A situation that will be common in practice, since GWAS results never include all genetic variants and rare variants may be left out. In this situation, we augment the model of equations 14-17 by splitting the scores in two parts: one part for common variants and the second part for rare variants. This situation is described with the following equations:

$$y = \gamma_1 S_y^* + \gamma_2 S_{y,rare}^* + \delta T + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{30}$$

$$T = \beta_1 S_T^* + \beta_2 S_{T,rare}^* + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2) \tag{31}$$

$$S_{yi} = S_y^* + e_{yi}, \qquad e_{yi} \sim \mathcal{N}(0, \sigma_{yi}^2) \tag{32}$$

$$S_{Ti} = S_T^* + e_{Ti}, \qquad e_{Ti} \sim \mathcal{N}(0, \sigma_{Ti}^2) \tag{33}$$

We determined β_1, β_2, γ_1 and γ_2 using the observed heritability and the estimated missing heritability. [29] estimate the chip heritability of height to be 0.56 and state that the total narrow-sense heritability is likely to be between 0.6 and 0.7 (the current estimate is still attenuated due to imperfect tagging of rare and structural genetic variants). For this simulation exercise we assume the observed chip heritability of T to be 0.55 and the total narrow-sense heritability to be 0.65. For y, we extrapolate this ratio and assume the observed chip heritability and the total narrow-sense heritability to be 0.2 and 0.24, respectively.

$$h^2_{T,tot} = \beta_1^2 + \beta_2^2 + 2\rho_r\beta_1\beta_2 \tag{34}$$

$$h^2_{T,obs} = \mathrm{cov}(T, S_T^*)^2 = (\beta_1 + \rho_r\beta_2)^2 \tag{35}$$

$$\Rightarrow \beta_2 = \sqrt{\frac{h^2_{T,tot} - h^2_{T,obs}}{1 - \rho_r^2}} \tag{36}$$

$$\beta_1 = \sqrt{h^2_{T,obs}} - \rho_r\beta_2 \tag{37}$$

similarly:

$$h^2_{y,tot} = \gamma_1^2 + 2\gamma_1\gamma_2\rho_r + 2\gamma_1\delta\beta_1\rho_G + 2\gamma_1\delta\beta_2\rho_G\rho_r + \gamma_2^2 + 2\gamma_2\delta\beta_1\rho_G\rho_r + 2\gamma_2\delta\beta_2\rho_G \tag{38}$$

$$+ \delta^2\beta_1^2 + 2\delta\beta_1\beta_2\rho_r + \delta^2\beta_2 \tag{39}$$

$$h^2_{y,obs} = \gamma_1^2 + \delta^2\beta_1^2 + 2\gamma_1\delta\beta_1\rho_r + 2\gamma_2\delta\beta_1\rho_G\rho_r + 2\gamma_1\delta\beta_2\rho_G\rho_r + 2\delta^2\beta_1\beta_2\rho_r \tag{40}$$

$$+ \rho_r^2(\gamma_2^2 + \delta^2\beta_2^2 + 2\gamma_2\delta\beta_2\rho_G) \tag{41}$$

$$\Rightarrow \gamma_2 = (-2\delta\beta_1\rho_G + \sqrt{(2\delta\beta_1\rho_G)^2 - 4(\delta^2\beta_2^2 - \frac{h^2_{y,tot} - h^2_{y,obs}}{1 - \rho_r^2})}\ )/2 \tag{42}$$

$$\gamma_1 = (-(2\delta\beta_1\rho_G + 2\delta\beta_2\rho_G\rho_r) + \sqrt{(2\delta\beta_1\rho_G + 2\delta\beta_2\rho_G\rho_r)^2 - 4C}\ )/2 \tag{43}$$

with:

$$C = \delta^2\beta_1^2 + 2\gamma_2\delta\beta_1\rho_G\rho_r + 2\delta^2\beta_1\beta_2\rho_r + \rho_r^2(\gamma_2^2 + \delta^2\beta_2^2 + 2\gamma_2\delta\beta_2\rho_G) \tag{44}$$

We assume that the rare variants are correlated with the common variants, but they are unobserved. In practice, we do not know the size of this correlation. To get a sense, we deconstructed the scores used in the empirical part of our paper into different parts. Specifically, we constructed PGS for EA and height using three different subsets of GWAS data from the 1000 Genomes project: (G1) a PGS score based only on SNPs included in HapMap 3 (G2) a PGS score based on common SNPs not included in HapMap 3, but included in 1000 Genomes (MAF>5%), and (G3) a PGS score based on rare SNPs not included in HapMap 3, but included in 1000 Genomes (MAF $\leq$ 5%). Due to LD, the three scores should be correlated

	EA_G1	EA_G2	EA_G3
EA_G1	1		
EA_G2	0.8257	1	
EA_G3	0.0549	0.0539	1

with each other, but the question is the extent to which this correlation affects the correlation of PGS for two different traits. The following correlation matrices show the correlation of G1 with the common SNPs not included in Hapmap 3, and the correlation between G2 and the rare SNPs included in 1000 Genomes (MAF $\leq 5\%$).

	Height_G1	Height_G2	Height_G3
Height_G1	1		
Height_G2	0.8685	1	
Height_G3	0.0542	0.0447	1

During the simulations, we again fix the GWAS sample at 1,000,000. We vary the correlation between the common and the rare genetic variants (ρ_r). Because γ_1 depends on ρ_r, γ_1 will have different values depending on the input parameters.

The simulation results are shown in Supplementary Figure 12. We see that the effect of T is still consistently estimated across all simulated scenarios. Thus, we conclude that missing (rare or structural) genetic variants do not lead to a noteworthy bias in the GIV regression estimate of the treatment T on outcome y.

3.2.4 Parental effects

The last situation we simulate is concerned with unobserved parental effects. In particular, the genetic correlation between a biological parent and his or her offspring is on average 0.5. Since the genotypes of the parents partly influence the environment of the offspring (e.g. via socio-economic status and parental habits), it is possible that environmental factors that are "inherited" via the parents will violate the exclusion restriction of IV regression because unobserved genotypes of the parents would partly correlate with the polygenic scores of the offspring and residual environmental factors captured by the error term. In principal, it is possible to control for parental genotypes directly (via genetic data from the parents) or indirectly (via the inclusion of family fixed-effects, e.g. in a large sample of dizygotic twins). However, only very few samples currently exist that either contain a large enough sample of genotyped trios (i.e. mother, father, child) or genotyped dizygotic twins and also the phenotypes of interest.

Our simulations below explore the consequences arising from unobserved "inherited environments" in a standard GIV regression model. Specifically, we model a situation where the unobserved weighted average polygenic score of the parents for the treatment variable (P_T) has a direct or indirect effect on the outcome, y. We augment the standard model as follows:

$$y = \gamma S_y^* + \delta T + \theta P_T + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{45}$$

$$T = \beta S_T^* + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2) \tag{46}$$

$$S_{yi} = S_y^* + e_{yi}, \qquad e_{yi} \sim \mathcal{N}(0, \sigma_{yi}^2) \tag{47}$$

$$S_{Ti} = S_T^* + e_{Ti}, \qquad e_{Ti} \sim \mathcal{N}(0, \sigma_{Ti}^2) \tag{48}$$

We assume P_T has a correlation of 0.5 with S_T^* and a correlation of $\rho_g = 0.5$ with S_y^*. Furthermore, we assume that the parental effect is directly related to S_T^* and only related to S_y^* through pleiotropy between T and y. A scenario where the parental genetic effect is more directly related to S_y^* can also be imagined. In both cases, the assumptions for GIV regression will be violated. However, a stronger bias can be expected when the omitted parental environment has a stronger correlation with the instrument for the treatment variable (S_T). Hence, these simulations can be seen as a worst case scenario. Furthermore, we vary θ from 0 to 0.15 (i.e., the effect of P_T on y is not stronger than the effect of T on y). In this situation β is the same as in the standard model. However, the simulations need to adjust γ to account for overestimation of the heritability due to the unobserved effect of P_T:

$$h_y^2 = \gamma^2 + 2\gamma\delta\beta\rho_G + \gamma\theta\rho_G + \delta^2\beta^2 + \delta\beta\theta + \theta^2 \tag{49}$$

$$\Rightarrow \gamma = (-2\delta\beta\rho_G - \theta\rho_G + \sqrt{(2\delta\beta\rho_G + \theta\rho_G)^2 + 4(\delta^2\beta^2 + \delta\beta\theta + \theta^2 - h_y^2)}\,)/2 \tag{50}$$

As before, the simulations fixed the GWAS sample at $n = 1,000,000$. We varied the strength of the parental effects (θ). Because γ depends on θ, γ has different values depending on the input parameters.

The results of these simulations are shown in supplementary figure 13.While we can see from the top panel that the genetic effects are estimated consistently with GIV regression, it is clear from the bottom panel that the parental effects bias the estimated effect of T. Our treatment effects are overestimated because of the omitted variable bias caused by P_T. Note that GIV regression still outperforms all the other models even in this case.

3.3 Simulation of gene-environment interactions

Next, we simulate the gene-environment interaction model described in SI section 2.3. In principle, these gene-environment interactions could be extremely complicated, but for practical reasons we focus on a simple linear interaction between T and S_y^*. The equations of the data generating process have been augmented to equations 51-54.

$$y = \gamma S_y^* + \delta_1 T + \delta_2 T S_y^* + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{51}$$

$$T = \beta S_T^* + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2) \tag{52}$$

$$S_{yi} = S_y^* + e_{yi}, \qquad e_{yi} \sim \mathcal{N}(0, \sigma_{yi}^2) \tag{53}$$

$$S_{Ti} = S_T^* + e_{Ti}, \qquad e_{Ti} \sim \mathcal{N}(0, \sigma_{Ti}^2) \tag{54}$$

We assume here that δ_2 is of the same size of δ_1 (0.15). Hence, the interaction term is relatively important. The other parameters have the same value as they do in the standard model (including the additional endogeneity), only we have to take the interaction term into account when calculating γ.

$$h_y^2 = \gamma^2 + \delta_1^2 \beta^2 + 2\gamma\delta_1\beta\rho_G + \delta_2(1 + \rho_G^2) \tag{55}$$

$$\Rightarrow \gamma = (-2\delta_1\beta\rho_G + \sqrt{(2\delta_1\beta\rho_G + \theta\rho_G)^2 + 4(\delta_1^2\beta^2 + \delta_2(1 + \rho_G^2) - h_y^2)}\,)/2 \tag{56}$$

As in the standard model, the GWAS sample varied from 200,000 to 2,000,000. The parameters are estimated with two methods, OLS and GIV regression. For GIV regression, S_{y1}, T, and TS_{y1} are used as endogenous regressors and S_{y2}, S_T, and $S_{y2}S_T$ are used as instruments.

The results of these simulations are shown in SI figure 14. From all three panels it is clear that the GIV regression estimates all three coefficients of equation 51 consistently, while the OLS estimates are clearly biased also for larger GWAS samples. The variance of the GIV regression estimates is relatively large compared to OLS. This should be taken into consideration if the effect size of the interaction term is expected to be smaller. We have not compared GIV regression to MR in this scenario, as there is not a standard way to do MR with polygenic scores and gene-environment interactions.

4 Data

We use data from the Health and Retirement Survey (HRS)[30]. 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.[4]

Since 2006, data collection has expanded to include biomarkers and a subset of the participants has been genotyped.[5] Autosomal SNPs were imputed using the world-wide reference panel from phase I of the 1000 Genomes project (v3, released March 2012)[31]. 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 are 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 [32]. The polygenic score for educational attainment is negatively correlated with birth year ($r =$ -0.06; $p < 0.0001$) and educational attainment influences longevity [33, 34]. Since the HRS is a sample of an older population, we further restrict our sample to a smaller range of birth years (1935 to 1945) to reduce sample selection bias that is correlated with the PGS. This resulted in a sample of 2,839 individuals.

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 [35], 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 30 percent of the SNPs are causal.

The polygenic scores for educational attainment were constructed by using GWAS results provided by the Social Science Genetic Association Consortium [20], excluding HRS and the *23andMe* cohort from the meta-analysis, but including the UK Biobank (see Supplementary Table 1). Three polygenic scores are constructed. First, a score using a meta-analysis of all available data. This score uses a sample of 318,954 individuals (1,873,557 SNPs). The other two scores are created by splitting the sample in two. One score is created by only using data of the UK Biobank ($n = 111,349$; 1,873,557 SNPs) and a score using the remaining sub-sample of several cohorts from around the world ($n = 207,605$; 1,849,602 SNPs).

The first polygenic score for height was constructed using the publicly available

[4] 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.

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

GWAS summary results from the GIANT consortium ($n = 253,288$) [25][6] which are based on ≈ 2.5 million autosomal SNPs that were imputed using the HapMap 2 CEU reference panel [36] (See Supplementary Table 2). 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.

For the second polygenic score for height, we conducted a GWAS on body height in the UK Biobank (UKB). 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 interim release of the genetic data ($n = 112,151$). Autosomal SNPs were imputed using the UK10K reference panel. Details on genotyping, pre-imputation quality control, and imputation have been documented extensively elsewhere [38]. GWAS analysis included as control variables dummies for genotyping batches, years of age, sex, and all interaction terms between age dummies and sex. Furthermore, the first 15 principal components of the genetic data were also included to control for subtle population structure. GWAS results underwent quality control following an extended version of the EasyQC protocol [39] described in detail elsewhere [23]. This yielded 1,861,232 autosomal SNPs that were included in the LDpred scores.

5 Empirical applications

We demonstrate the value of GIV regression approach in several important empirical applications. First, we estimate the chip heritability of educational attainment (EA) from a PGS for EA. Second, we estimate the (causal) effect of body height on EA. Earlier studies have reported a positive relationship between these variables [40, 41, 42]. Third, we present results from a negative control that estimates the (causal) effect of EA on body height (which should be zero). We estimate these effects using OLS, MR, EMR, and GIV regression. We include birth year, birth year squared, educational attainment of both parents and (in pooled models) gender as control variables. We included PGS of EA or height depending on the method. All variables have been standardized.

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

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

results instead.

The OLS results in Table 2 in the main text appear to show that height has a strong effect on EA. MR seems to confirm the causal interpretation of the OLS result; indeed, the point estimate from MR is even slightly larger than from OLS. However, and as discussed above, MR suffers from probable violations of the exclusion restriction. These violations could stem from the possibility that the same genetic factors that increase height are also directly increasing EA. They could also stem from the possibility that the PGS for height by itself is correlated with the genetic tendency for parents to have higher EA and income, and therefore a lower nutritional or disease risk for their children, who therefore are more likely to reach their full cognitive potential and have higher EA. Controlling for the PGS (i.e., EMR) is an imperfect strategy for eliminating this source of endogeneity, because the bias in the estimated effect of the PGS score also biases the estimated effect of height (the omitted variable bias discussed earlier).

In contrast, estimates from both GIV regressions in Table 2 in the main text show both a considerably larger effect of the education PGS score on EA, and a small and statistically insignificant effect of height on EA. These results imply that the positive correlation between height and EA is not a causal relationship. Instead, the phenotypic correlation seems to be entirely explained by the genetic correlation between the two traits.

Supplementary Table 3 shows the estimates of EA on height (which should be zero) using the four estimation strategies. OLS and MR both find (erroneously) a statistically significant positive effect of EA on height. The GIV regression estimate for the effect of EA on height is indistinguishable from zero in both specifications. In this application, EMR also finds a small and statistically insignificant effect of EA on height, though it underestimates the genetic contribution to height.

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. 2.1

[2] Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 9(3):e1003348. 2.1, 2.1, 3.1, 3.1

[3] McCarthy MI et al. (2008) Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9(5):356–369. 2.1

[4] Price AL et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38(8):904–909. 2.1

[5] Abdellaoui A, Al. E (2013) Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* 21(11):1277–1285. 2.1

[6] Vilhjálmsson BJ et al. (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97(4):576–592. 2.1, 4

[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. 2.1

[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. 2.1, 2.2

[9] Rietveld CA et al. (2013) GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science (80-.).* 340(6139):1467–1471. 2.1, 2.1, 2.2, 3.1

[10] Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 9(3):e1003348. 2.1

[11] Branigan AR, McCallum KJ, Freese J (2013) Variation in the heritability of educational attainment: An international meta-analysis. *Social forces* 92(1):109–140. 2.1

[12] 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. 2.1

[13] Bollen KA (2002) Latent variables in psychology and the social sciences. *Annual Review of Psychology* 53(1):605–634. 2.1

[14] Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: An empiricist's companion.* (Princeton University Press). 2.1, 2.1, 2.2

[15] Burgess S, Small DS, Thompson SG (2015) A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research* p. 0962280215597579. 2.1, 2.2

[16] Davey Smith G, Hemani G (2014) Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* 23(R1):R89–98. 2.1, 2.2

[17] de Vlaming R et al. (2016) Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies, Technical report. 2, 3.1

[18] Burgess S, Butterworth A, Malarstig A, Thompson SG (2012) Use of Mendelian randomisation to assess potential benefit of clinical intervention. 2.2

[19] Bulik-Sullivan B et al. (2015) An atlas of genetic correlations across human diseases and traits. *Nature genetics.* 2.2

[20] Okbay A et al. (2016) Genome-wide association study identifies 74 loci associated with educational attainment. *Nature.* 2.2, 4

[21] Yang J et al. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47(10):1114–20. 2.2, 3.1

[22] Plomin R et al. (2013) Common DNA markers can account for more than half of the genetic influence on cognitive abilities. *Psychol. Sci.* 24(4):562–568. 2.2

[23] Okbay A et al. (2016) Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* 3.1, 4

[24] Locke AE et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518(7538):197–206. 3.1

[25] 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. 3.1, 4

[26] Okbay A et al. (2016) Genome-wide association study identifies 74 loci associated with educational attainment. *Nature.* 3.1, 5

[27] Fleishman AI (1978) A method for simulating non-normal distributions. *Psychometrika* 43(4):521–532. 3.2.1

[28] Luo H (2011) Generation of non-normaldata : A study offleishman's powermethod, (Uppsala University, Department of Statistics), Technical Report 2011:1. 3.2.1

[29] Yang J et al. (2015) Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* 19(7):807–812. 3.2.3

[30] Sonnega A et al. (2014) Cohort profile: the Health and Retirement Study (HRS). *Int. J. Epidemiol.* 43(2):576–585. 4

[31] The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65. 4

[32] 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. 4

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

[34] 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. 4

[35] Consortium TIH (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58. 4

[36] Consortium TIH (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861. 4

[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. 4

[38] Marchini J et al. (2015) Genotype imputation and genetic association studies of UK Biobank. 4

[39] Winkler TW et al. (2014) Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* 9(5):1192–1212. 4

[40] 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

[41] 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

[42] Case A, Paxson C (2006) Stature and status: Height, ability, and labor market outcomes, (National Bureau of Economic Research), Technical report. 5

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

Supplementary Tables

Thomas DiPrete, Casper Burik and Philipp Koellinger

May 4, 2017

SI table 1: Cohort list for Educational Attainment

Study	Full name	Sampling	Country	Sample size
ACPRC	Manchester Studies of Cognitive Ageing	Population-based	England	1713
AGES	Age, Gene/ Environment SusceptibilityReykjavik 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-Korula	Croatia Korula	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	Netherlands	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

SI table 1 – 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	Netherlands	12539
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	Netherlands	1808
NESDA	Netherlands Study of Depression and Anxiety	Case-control (Mental health)	Netherlands	1820
NFBC66	Northern Finland Birth Cohort 1966	Population-based	Finland	5297
NTR	Netherlands Twin Register	Family-based	Netherlands	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

SI table 1 – continued

Study	Full name	Sampling	Country	Sample size
PREVEND	Prevention of Renal and Vascular End-stage Disease	Population-based	Netherlands	3578
QIMR	Queensland Institute of Medical Research	Family-based	Australia	8006
RS-I	Rotterdam Study Baseline	Population-based	Netherlands	6108
RS-II	Rotterdam Study Extension of Baseline	Population-based	Netherlands	1667
RS-III	Rotterdam Study Young	Population-based	Netherlands	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
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 [1], excluding the Health and Retirement Study and 23andMe cohorts. A more detailed list and description can be found in the supplementary materials of [1]

SI table 2: Cohort list for Height

Study	Full name	Sampling	Country	Sample size
ACTG	The AIDS Clinical Trials Group	Population-based	International	1055
ADVANCE	Atherosclerotic Disease, VAscular FunctioN, and GenetiC Epidemiology	Population-based case-control	USA	584
AE	Athero-Express Biobank Study	patient-cohort	The Netherlands	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”	Netherlands	2669
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

SI table 2 – continued

Study	Full name	Sampling	Country	Sample size
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
EPIC-Obesity Study	European Prospective Investigation into Cancer and Nutrition - Obesity Study	Population-based	UK	3552
ERF	Erasmus Rucphen Family Study	Family-based	Netherlands	2726
FamHS	Family Heart Study	Population-based	USA	1463
Fenland	Fenland Study	Population-based	UK	1402
FINGESTURE 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
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
HYPERGENES Cases	HYPERGENES	Case-control	Italy/ UK/ Belgium	1841
HYPERGENES 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

SI table 2 – continued

Study	Full name	Sampling	Country	Sample size
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
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
NTRNESDA	Netherlands Twin Register & the Netherlands Study of Depression and Anxiety	Case-control	Netherlands	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	Netherlands	3624
PROCARDIS	Precocious Coronary Artery Disease	Population-based	UK	7000

SI table 2 – continued

Study	Full name	Sampling	Country	Sample size
PROSPER/ PHASE	The PROspective study of Pravastatin in the Elderly at Risk for vascular disease	Randomized controlled trial	Netherlands, 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	Netherlands	5744
RS-II	Rotterdam Study II	Population-based	Netherlands	2124
RS-III	Rotterdam Study III	Population-based	Netherlands	2009
RUNMC	Nijmegen Bladder Cancer Study (NBCS) & Nijmegen Biomedical Study (NBS), Radboud University Nijmegen Medical Centre	Population-based	Netherlands	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)	Netherlands	1139
TWINGENE	TWINGENE	Population-based	Sweden	9380
TwinsUK	TwinsUK	Twins pairs	UK	1479
VIS	VIS (EUROSPAN) and KORCULA	Population-based	Croatia	784
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 Educational Attainment by Wood et al. (2014). A more detailed list and description can be found in the supplementary materials of [2] and [3]

SI table 3: Regression of height on educational attainment in the Health and Retirement Study (HRS)

	(1)	(2)	(3)	(4)	(5)
	OLS	MR	EMR	GIV1	GIV2
EA	0.0563***	0.156*	-0.00686	0.0396	-0.0646
	(0.0121)	(0.0645)	(0.0577)	(0.0882)	(0.0679)
PGS_Height_GIANT	0.165***		0.165***	0.634***	
	(0.0139)		(0.0140)	(0.0293)	
PGS_Height_UKB	0.206***		0.208***		0.512***
	(0.0125)		(0.0126)		(0.0252)
Birth Year	2.524	14.43	2.558	-2.961	-3.793
	(14.54)	(16.58)	(14.62)	(17.24)	(16.23)
Birth Year Squared	-2.534	-14.44	-2.566	2.955	3.786
	(14.54)	(16.58)	(14.62)	(17.24)	(16.23)
Gender	-0.759***	-0.739***	-0.765***	-0.773***	-0.771***
	(0.0108)	(0.0135)	(0.0120)	(0.0150)	(0.0135)
Mother's EA	0.0202	0.00109	0.0367	0.0146	0.0525*
	(0.0136)	(0.0226)	(0.0201)	(0.0278)	(0.0230)
Father's EA	-0.000745	-0.0222	0.0132	0.0231	0.0126
	(0.0134)	(0.0207)	(0.0184)	(0.0251)	(0.0209)
N	2839	2839	2839	2839	2839

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Standard errors in parentheses. All variables have been standardized. EA is measured in years of schooling needed to obtain the highest achieved educational degree according to ISCED classifications. The first 10 principal components in the genetic data were included as control variables. PGS_Height_GIANT: PGS for Height using data from the GIANT consortium [2]. PGS_Height_UKB: PGS for Height using UKB data. PGS_EA_UKB: PGS for EA using UKB data. PGS_EA_Total: PGS for EA using GWAS meta analysis of UKB + SSGAC [1], excluding data from 23andMe and HRS; PGS_EA_SSGAC: PGS for EA using meta analysis from Okbay et al [1], excluding data from23andMe, UKB, and HRS. MR and EMR use PGS_EA_Total as an instrument for EA. GIV1 uses PGS_Height_UKB and PGS_EA_UKB as instruments for EA and PGS_Height_GIANT. GIV2 uses PGS_Height_GIANT and PGS_EA_SSGAC as instruments for EA and PGS_Height_UKB.

References

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

[2] 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.

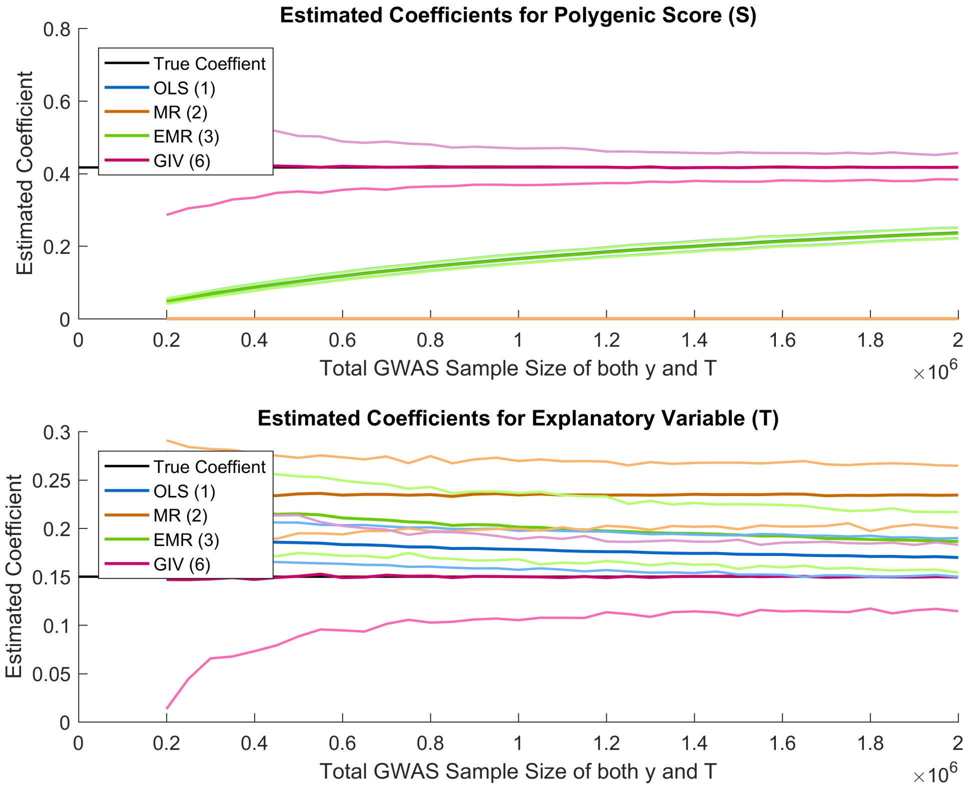
[3] Allen HL, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467(7317):832–838.

Supplementary Figures

Thomas DiPrete, Casper Burik, Philipp Koellinger

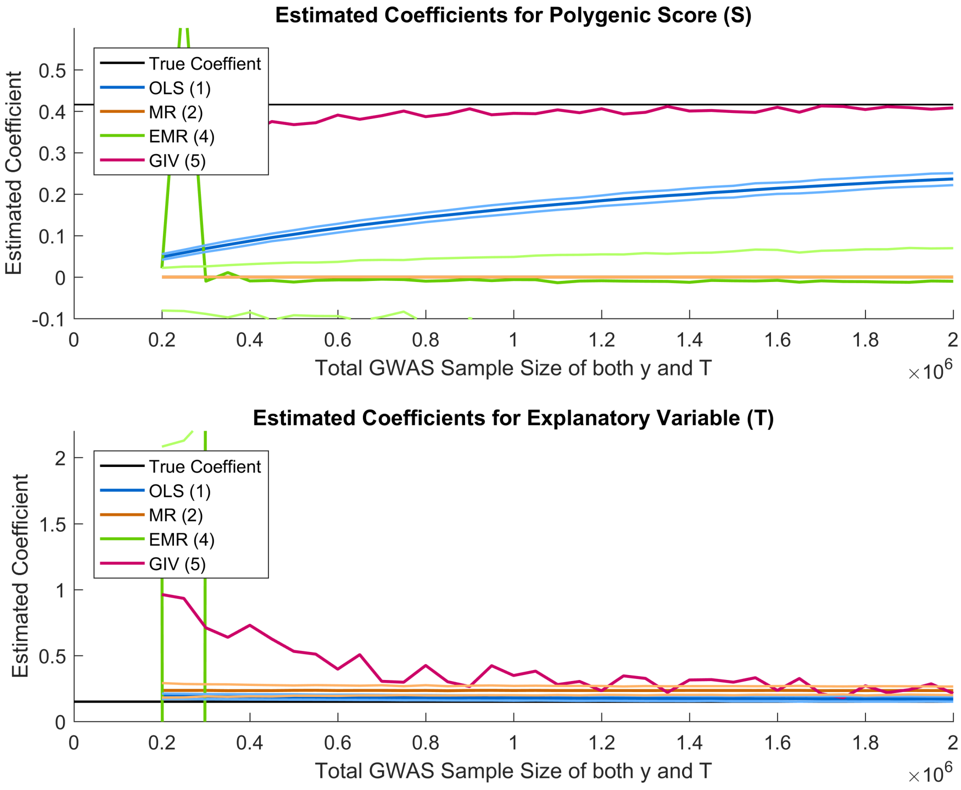
April 3, 2017

SI figure 1: Model with measurement errors and no additional endogeneity and methods 1,2,3,6 (see main text)



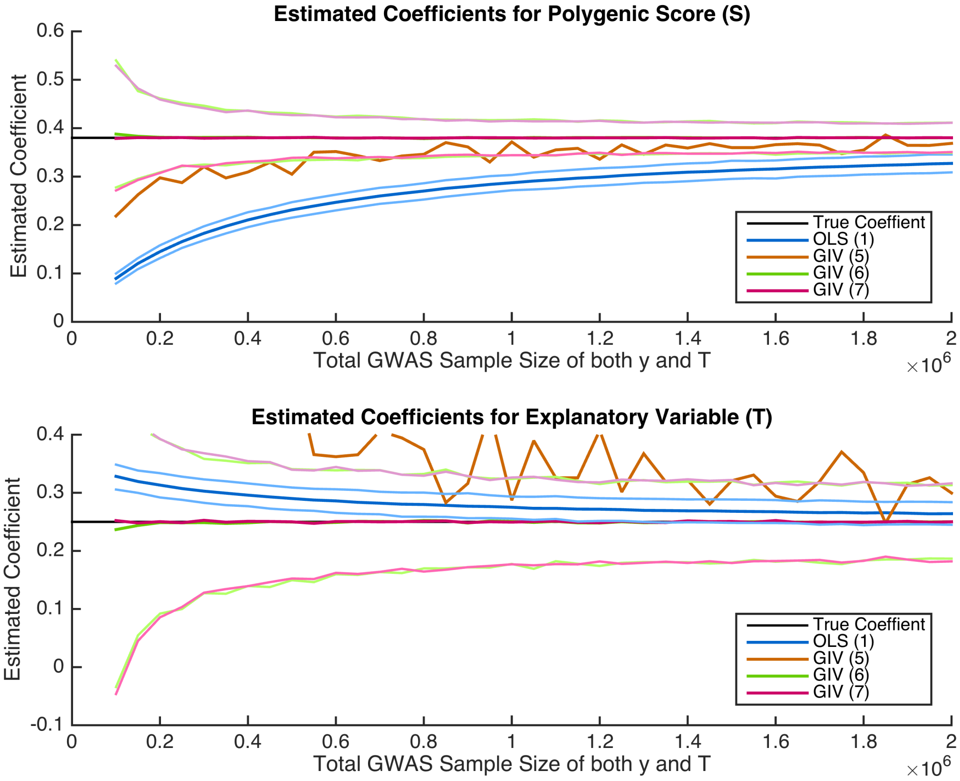
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T.

SI figure 2: Model with measurement errors and no additional endogeneity and methods 1,2,4,5 (see main text)



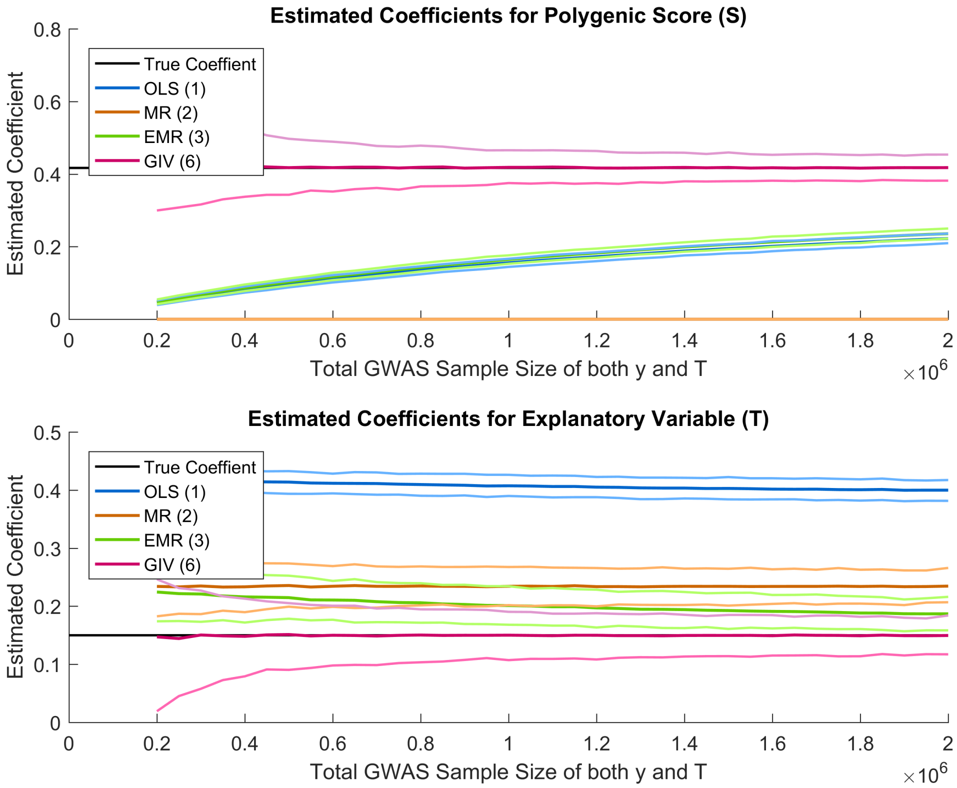
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. Note that for method 5 the confidence bounds are outside of the figure.

SI figure 3: Model with measurement errors and no additional endogeneity and methods 1,5,6,7 (see main text)



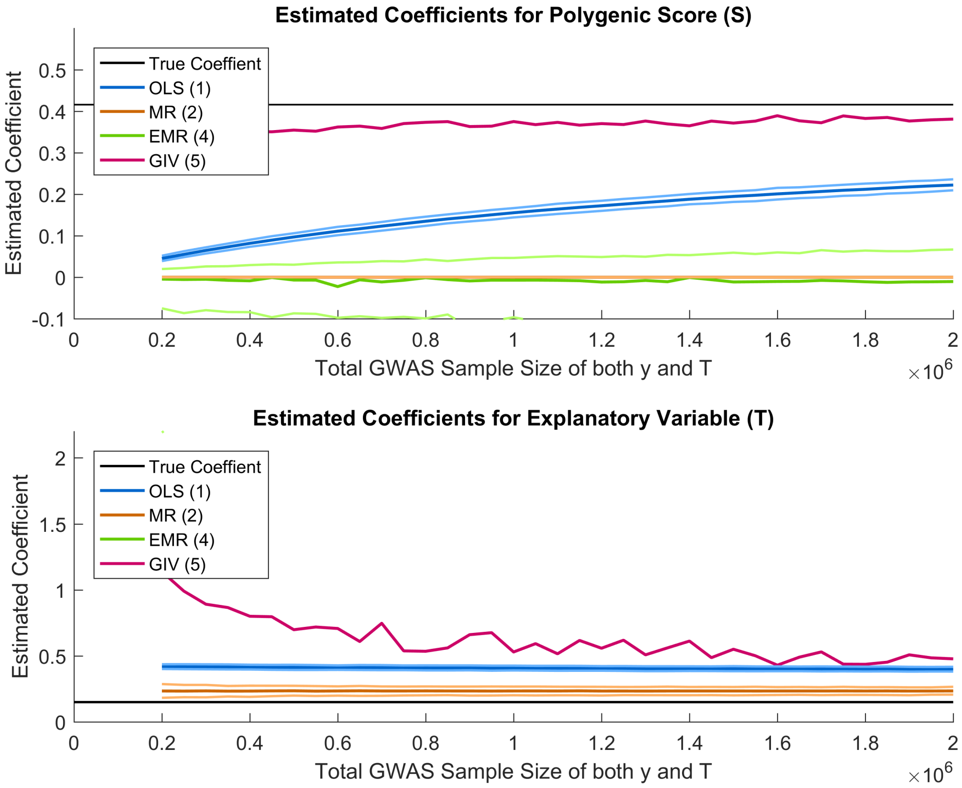
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. Note that for method 5 the confidence bounds are outside of the figure.

SI figure 4: Model with measurement errors, additional endogeneity, and methods 1,2,4,7 (see main text)



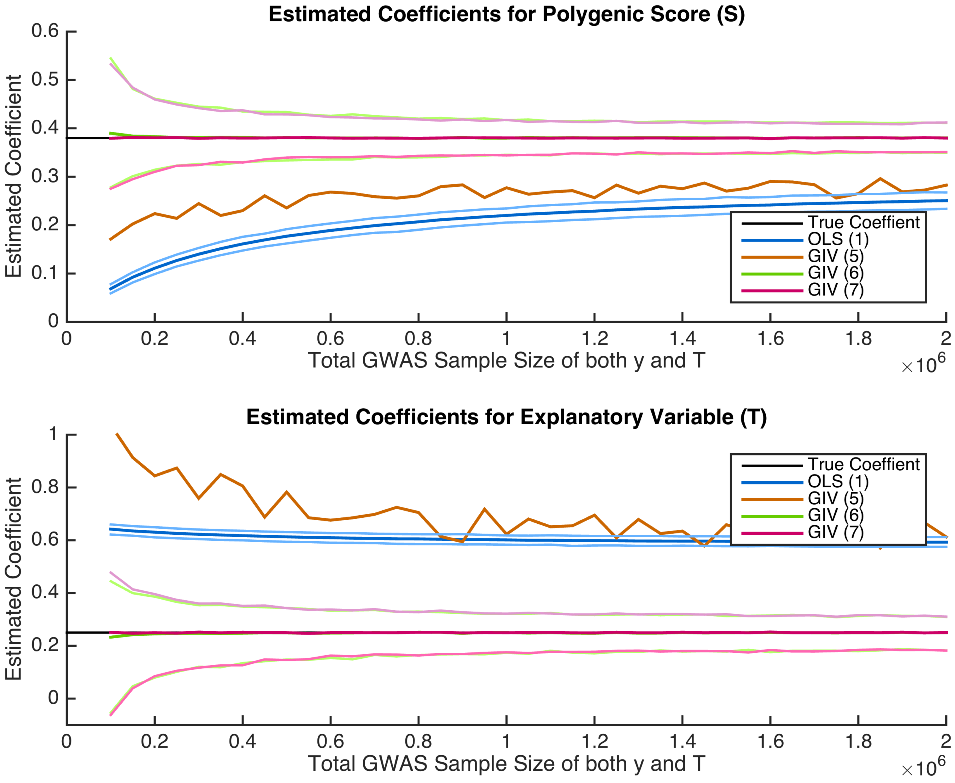
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T.

SI figure 5: Model with measurement errors, additional endogeneity, and methods 1,2,4,7 (see main text)



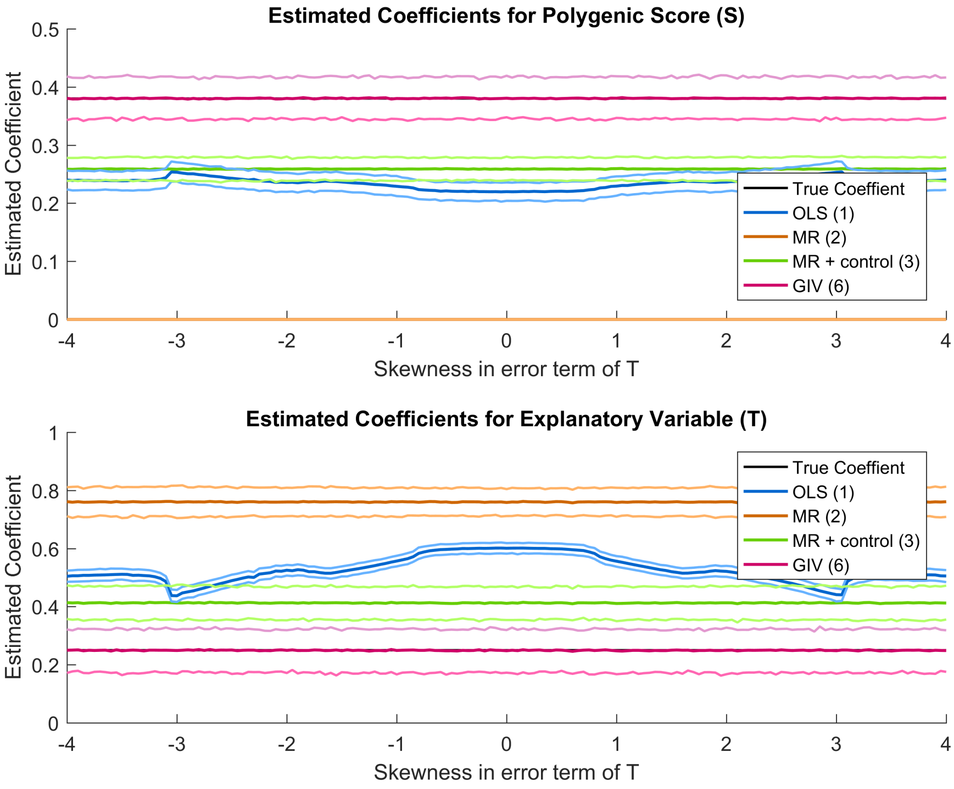
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,000, assuming 300,000 independent SNPs, a heritability of 0.2 for both traits, a genetic correlation of 0.6, a coefficient of 0.25 for T and a correlation of 0.4 between the error terms in T and y. Note that for method 5 the confidence bounds are outside of the figure.

SI figure 6: Model with measurement errors, additional endogeneity, and methods 1,5,6,7 (see main text)



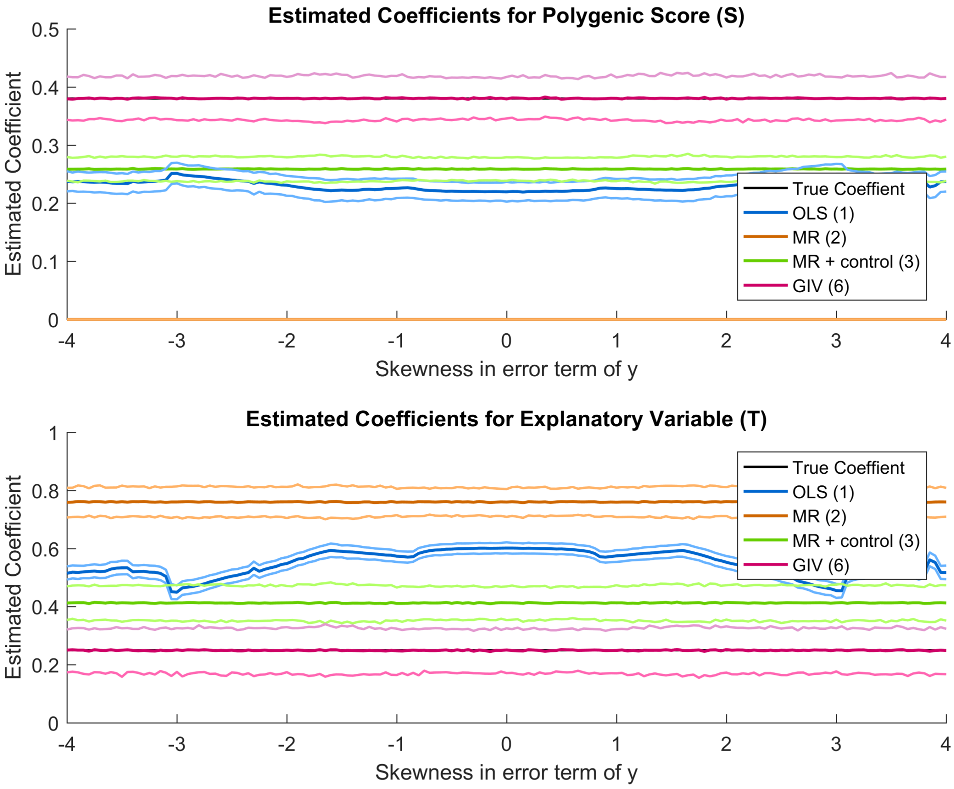
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. Note that for method 5 the confidence bounds are outside of the figure.

SI figure 7: Model with measurement errors and skewness in T



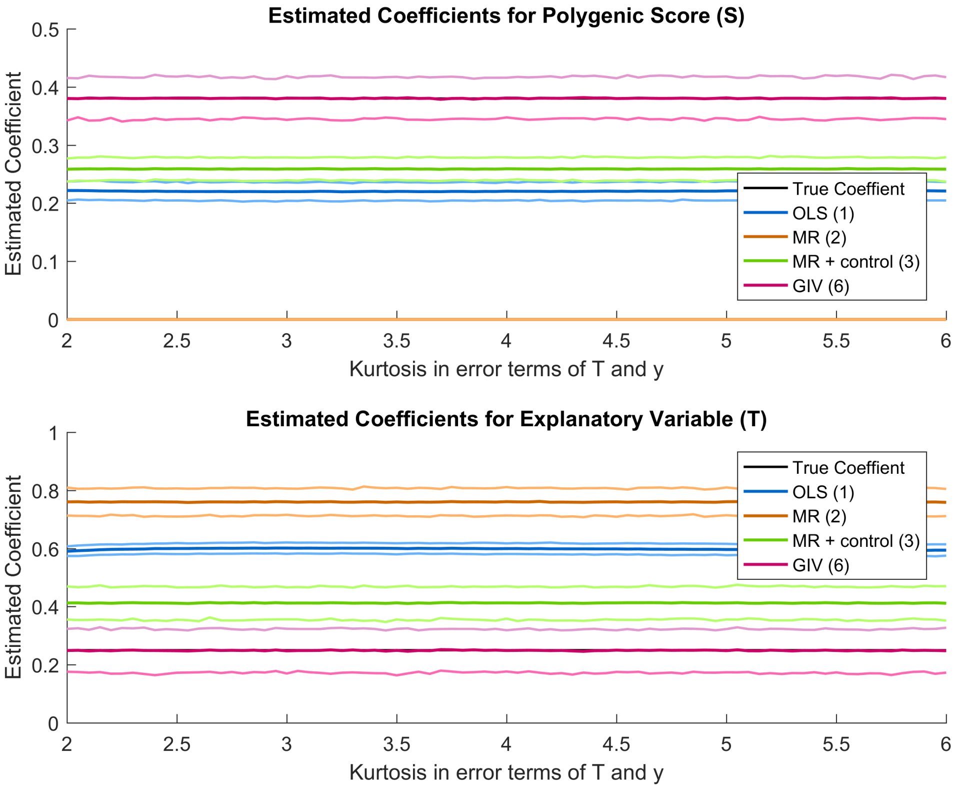
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. The total GWAS sample size is 1,000,000. Skewness is added to the error term in T.

SI figure 8: Model with measurement errors and skewness in y



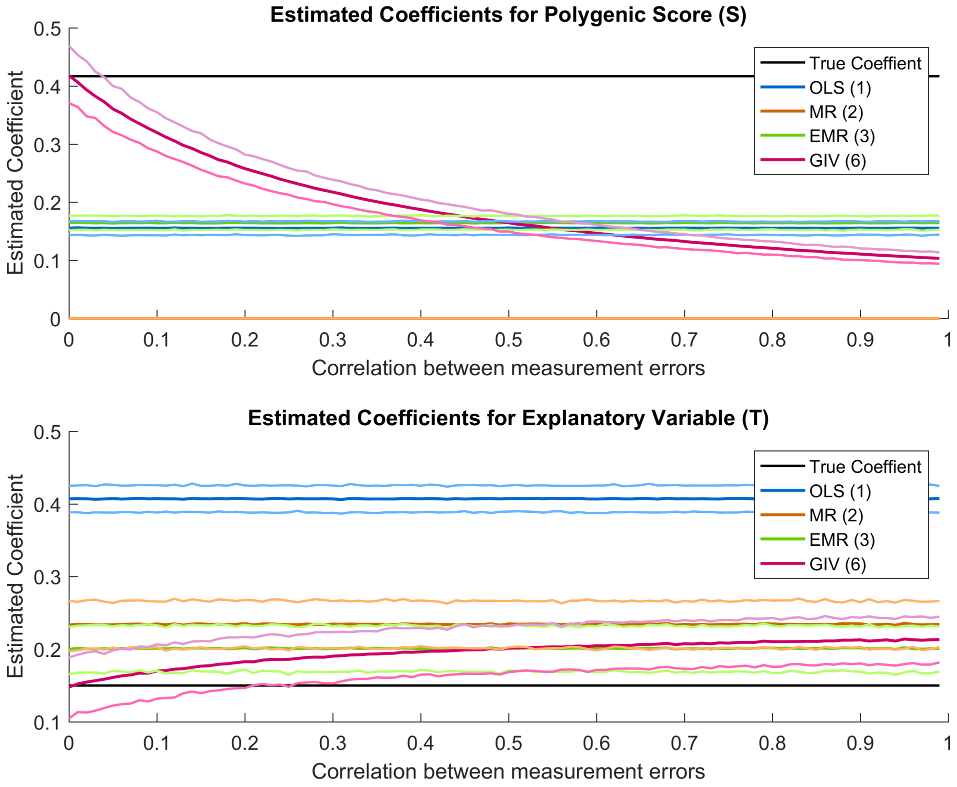
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. The total GWAS sample size is 1,000,000. Skewness is added to the error term in y.

SI figure 9: Model with measurement errors and kurtosis



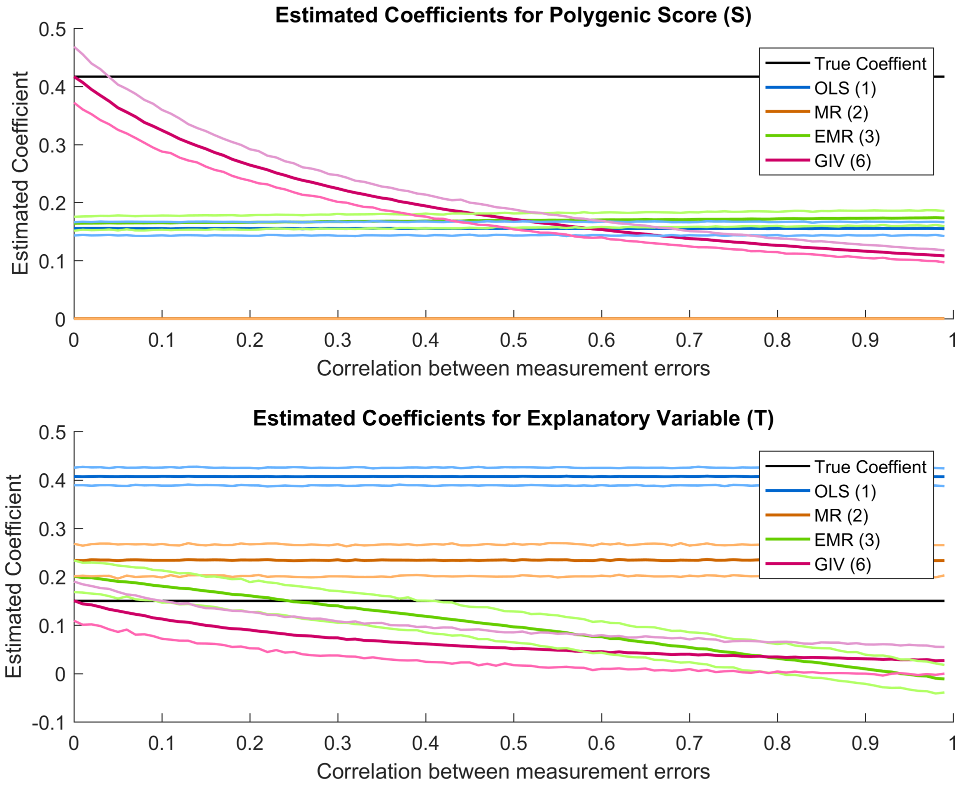
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. The total GWAS sample size is 1,000,000. Kurtosis is added to the error term in y and T.

SI figure 10: Model with dependent measurement errors between multiple indicators of S



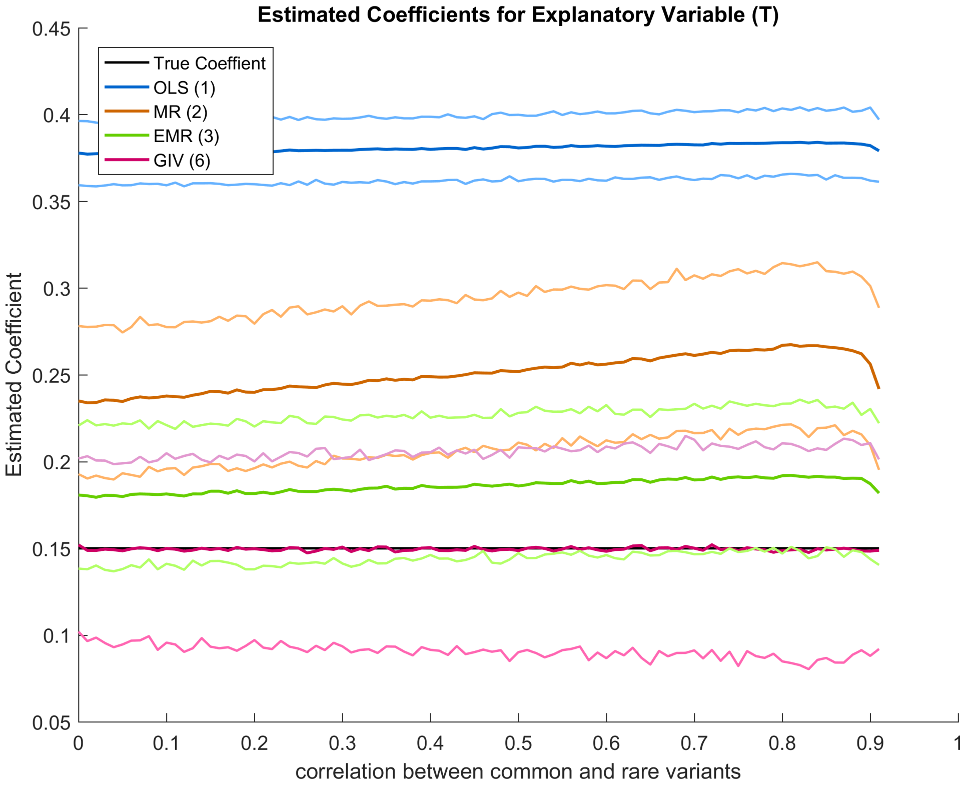
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. The total GWAS sample size was 1,000,000. The measurement errors in the PGS for y are correlated, the one in PGS for T is independent.

SI figure 11: Model with dependent measurement errors among T and multiple indicators of S (all correlations assumed equally large)



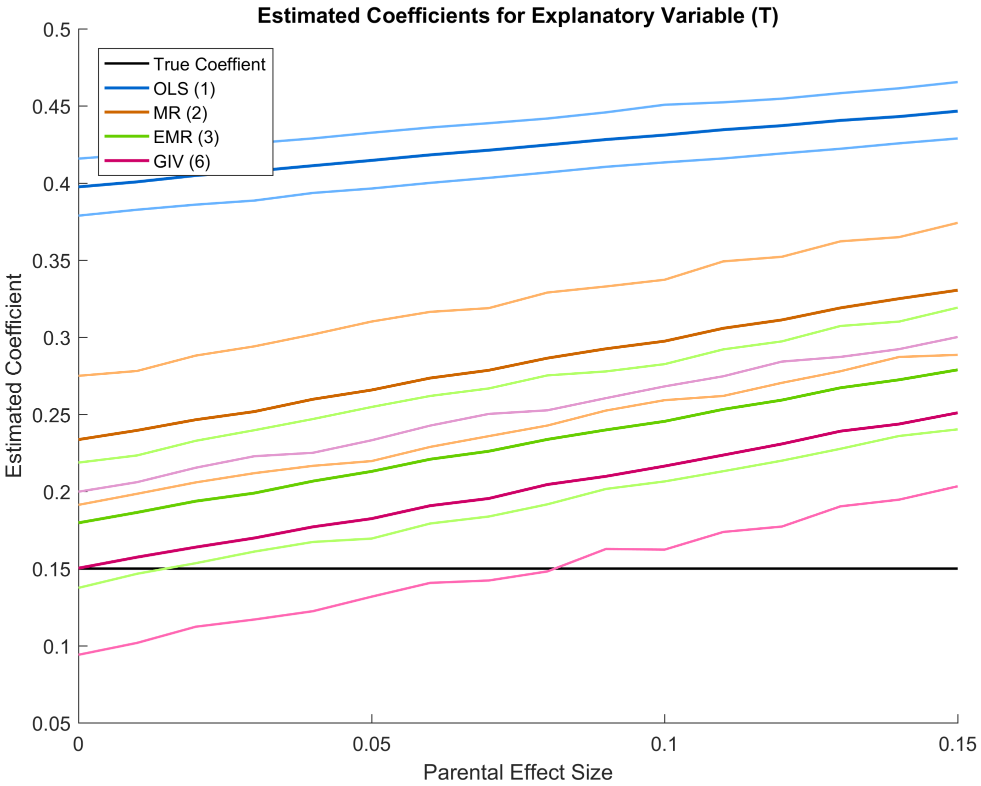
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. The total GWAS sample size was 1,000,000. The measurement errors in all PGS are correlated.

SI figure 12: Model with correlations between common and rare genetic variances



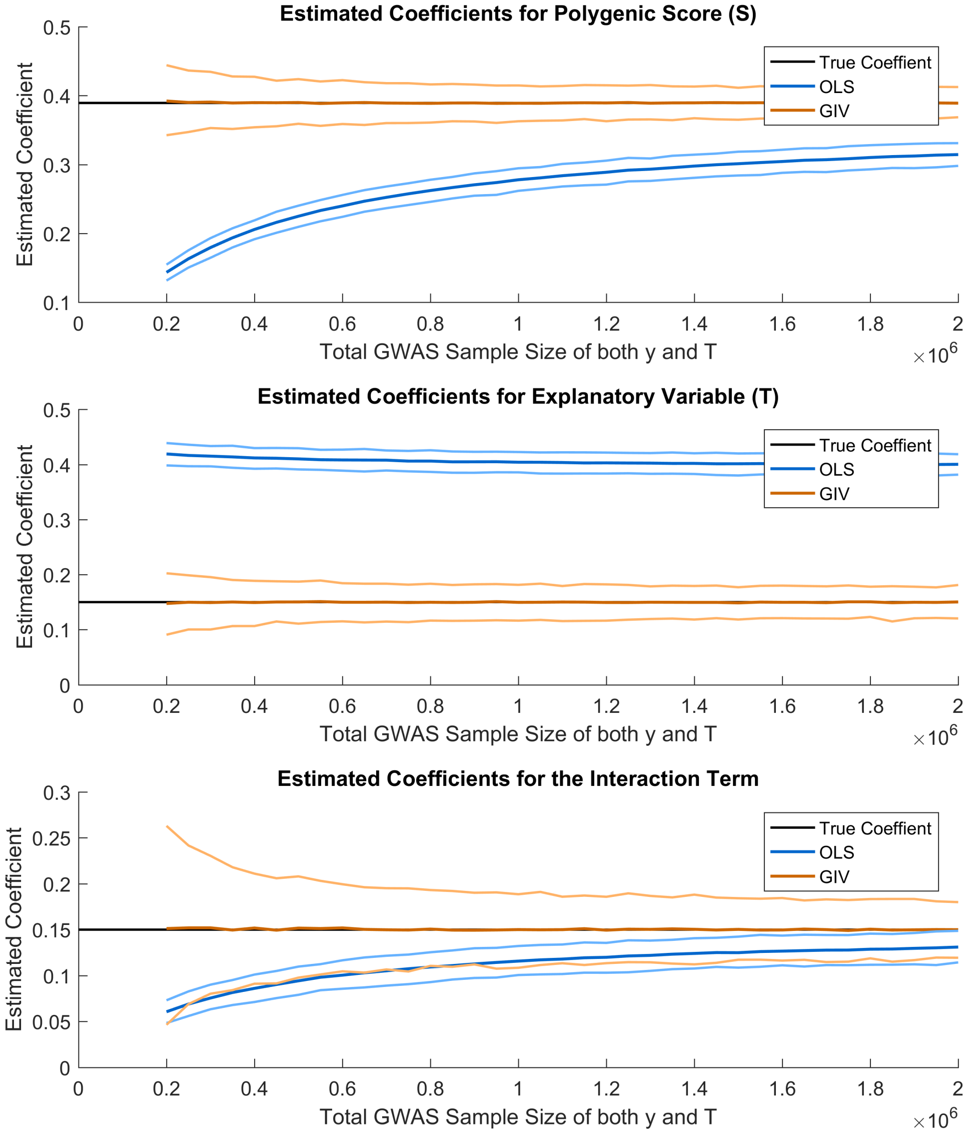
Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. The total GWAS sample size was 1,000,000. Unobserved causal genetic variants were added too model A.

SI figure 13: Model with parental effects



Estimated coefficients for model A, using various methods. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y and a correlation of 0.4 between the error terms in y and T. The total GWAS sample size was 1,000,000. Unobserved parental effects were added to model A.

SI figure 14: Model with interaction term



Estimated coefficients for model B, using OLS and GIV. The mean of the simulations for each method is shown, together with simulated 95% confidence interval is shown. The simulations are based on a sample of 8,600, assuming 300,000 independent SNPs, a heritability of 0.2 and 0.55 for y and T respectively, a genetic correlation of 0.15, a coefficient of 0.15 for effect of T on y, a coefficient of 0.15 for the interaction term and a correlation of 0.4 between the error terms in y and T. The total GWAS sample size was 1,000,000. Unobserved parental effects were added to model A.