Educational attainment, body mass index, and smoking initiation:

Using genetic path analysis to control for pleiotropy in a Mendelian randomization study

Frank D. Mann[1], Andrey A. Shabalin[2], Anna R. Docherty[2], & Robert F. Krueger[1]

[1]Department of Psychology, University of Minnesota

Department of Psychiatry, University of Utah

Correspondence to Frank D. Mann (fmann@umn.edu)

Abstract

When a randomized experimental study is not possible, Mendelian randomization studies use genetic variants or polygenic scores as instrumental variables to control for gene-environment correlation while estimating the association between an exposure and outcome. Polygenic scores have become increasingly potent predictors of their respective phenotypes, satisfying the relevance criteria of an instrumental variable. Evidence for pervasive pleiotropy, however, casts doubt on whether the exclusion criteria of an instrumental variable is likely to hold for polygenic scores of complex phenotypes, and a number of methods have been developed to adjust for pleiotropy in Mendelian randomization studies. Using multiple polygenic scores and path analysis we implement an extension of genetic instrumental variable regression, genetic path analysis, and use it to test whether educational attainment is associated with two health-related outcomes in adulthood, body mass index (BMI) and smoking initiation, estimating both gene-environment correlation and pleiotropy. Results provide compelling evidence for a complex set of gene-environment transactions that undergird the relation between educational attainment and health-related outcomes in adulthood.

Keywords: education; cognitive ability; polygenic risk; Mendelian randomization; pleiotropy

Educational attainment, body mass index, and smoking initiation: Using genetic path analysis to control for pleiotropy in a Mendelian randomization study

Mendelian randomization refers to the random assortment of genes that are given to children by their parents at the time of conception (Smith & Ebrahim, 2003). This results in distributions of genes that are independent of many factors that often confound associations documented in observational studies (Lawlor et al., 2008; Smith & Ebrahim, 2003; Smith & Ebrahim, 2004). Mendelian randomization studies use genetic variants or genetic propensity scores, also called polygenic risk scores, as instrumental variables to control for gene-environment correlation when testing a putatively casual relation between an exposure and outcome. The present study focuses on the use of polygenic scores to conduct Mendelian randomization studies, with emphasis placed on reviewing whether polygenic scores meet the criteria for a sound instrumental variable. We then present an extension of genetic instrumental variable regression (DiPrete, Burik, & Koellinger, 2018), genetic path analysis, to help overcome a limitation inherent to Mendelian randomization studies of complex phenotypes, specifically the high potential for pleiotropic effects on the exposure and outcome of interest. Using genetic path analysis, we then test whether educational attainment is associated with body mass index (BMI) and smoking initiation in a large sample of adults while estimating both gene-environment correlation and pleiotropy.

Gene-environment correlation refers to the non-random assortment of individuals into environments based on their genotype and is behaviorally manifest by individuals actively shaping and responding to their environments based, at least partly, on their heritable characteristics (Briley, Livengood, & Derringer, 2018; Jaffee & Price, 2007). This process results in heritable variation in measures of the environment (Kendler & Baker, 2007), which, in

turn, are thought to further reinforce the expression of relevant phenotypes. Importantly, without accounting for heritable variation in environmental exposures, one cannot know whether an association between an exposure and outcome reflects a true causal relation or, on the other hand, a niche-picking process (Scarr & McCartney, 1983). Auspiciously, as summary data from genome-wide association studies (GWASs) becomes readily available, it has become increasingly popular to use polygenic scores as instrumental variables for inferring causation in non-experimental studies (a.k.a. Mendelian randomization studies).

A polygenic score may be defined "as a single value estimate of an individual's propensity to a phenotype" (p. 1, Choi, Mak, & O'Reily, 2018) calculated by computing the sum of risk alleles corresponding to a phenotype in each individual, weighted by their effect size estimate from the most powerful GWAS on the phenotype. A polygenic score is typically calculated as $\mathrm{PGS}_k = \sum_i \beta_i \, \mathrm{SNP}_{ik}$, where PGS for individual *k* in the target sample is calculated by the summation of each SNP (measured for both the person *k* and passing a set association threshold in the discovery GWAS) multiplied by the effect size, β, of that SNP in the discovery GWAS. Thus, polygenic scores provide an index of an individual's genetic propensity for a given phenotype, or "an individual-level genome-wide genetic proxy" (p. 2, Choi, Mak, & O'Reily, 2018). Although polygenic scores may be used for a variety of purposes, a lot of emphasis has been placed on using polygenic scores as instrumental variables. However, as noted and addressed by others (Bowden et al. 2015; DiPrete, Burik, & Koellinger, 2018; van Kippersluis & Rietveld, 2017), it is not clear that polygenic scores meet the necessary criteria for a sound instrumental variable.

There are three criteria for a sound instrumental variable (Greenland, 2000). First, sometimes called the relevance criteria, the instrument must be related to the environmental

exposure. Second, according to the exclusion criteria, conditional on the relation between the exposure and outcome, there is no direct relation between the instrument and the outcome. Put differently, any relation between the instrument and outcome must be fully accounted for by its relation to the exposure. Third, the instrument should not be related to any unmeasured confounders. Note, however, that this third criteria, sometimes called the independence criteria, is not unique to using polygenic scores as instrumental variables, or instrumental variable analysis more generally, as this concern applies to all non-experimental studies for which an unmeasured confounder exists.

Nevertheless, as the size of GWASs continue to grow, polygenic scores have become increasingly potent predictors of their respective phenotypes, satisfying the relevance criteria. On the other hand, genetic correlations across related and seemingly unrelated phenotypes provides evidence for pleiotropic effects. This suggests that polygenic scores likely violate the exclusion criteria, and, therefore, casts doubt on their use as instrumental variables. In response to this concern, a number of methods have been developed to help correct for the presence of pleiotropy. For example, statistical techniques have been developed that are more robust to pleiotropic effects violating the exclusion criteria, including Egger regression (Bowden et al., 2015) and summary data-based multiple regression (Zhu et al., 2018), as well as pleiotropy-robust Mendelian randomization (Van Kippersluis & Rietveld, 2017) and genetic instrumental variable regression (DiPrete, Burik, & Koellinger, 2018). The present study intends to contribute to this body of work by integrating two existing methods, genetic instrumental variable regression and path analysis, to estimate and help control for pleiotropy in a Mendelian randomization study using multiple polygenic scores.

In a traditional Mendelian randomization study, two regressions are estimated simultaneously: the environmental exposure is regressed on the genetic instrument, and the outcome of interest is regressed on the environmental exposure. Unfortunately, due to pleiotropic effects, the association between the genetic instrument and the outcome is not fully mediated by the association between the genetic instrument and the exposure. Put differently, conditional on the association between the exposure and outcome, the genetic instrument is often predictive of both the environmental exposure *and* outcome, violating the exclusion criteria of a sound instrumental variable. However, as summary statistics from GWASs become available for a number of social, relational, and environmental exposures, in addition to outcomes of clinical and epidemiological interest, a path analysis using polygenic scores for an exposure *and* outcome can provide an estimate and control for pleiotropy when conducting a Mendelian randomization study.

An example of a path analysis using multiple polygenic scores is depicted in Figure 1. Similar to a traditional instrumental variable analysis, an environment or exposure (E) is regressed on a genetic instrument (PRS_E), which estimates and controls for gene-environment correlation. An outcome (Y) is then regressed on the exposure (E) free of genetic confounds that result from active and evocative gene-environment correlation. To estimate and control for the potential pleiotropic effects of the genetic instrument, a second genetic instrument is introduced (PRS_Y), which provides an index of polygenic liability for the outcome (Y). The correlation between the genetic instrument for the exposure (PRS_E) and the genic instrument for the outcome (PRS_Y) can be freely estimated, while simultaneously regressing the exposure (E) and outcome (Y) on the genetic instrument for the outcome (PRS_Y). These parameters provide a test and simultaneous control for pleiotropy, while also estimating and controlling for additional

gene-environment correlations that may not have been captured by the first genetic instrument. The correlation between the two genetic instruments sheds light on whether genetic liability for the exposure has pleiotropic effects on the outcome, and the regression of the outcome on its polygenic score provides a statistical control for pleiotropy. Finally, the regression of the exposure on the genetic instrument for the outcome tests for potential gene-environment correlations not fully accounted for by the genetic instrument for the exposure. Hereinafter, we provide a demonstration of this method focusing on the relationship between education and two important health-related outcomes: body mass index (BMI) and smoking initiation.

Method

Sample

The present study analyses data from the Study of Midlife Development in the United States (MIDUS; Brim, Ryff, & Kessler, 2004). Data was prepared for analyses with R version 3.5.2. Data was imported into R using the ‘Hmisc’ package (Harrell & Harrell, 2019), preprocessed, and then exported from R using the 'MplusAutomation' package version 0.7.1 (Hallquist & Wiley, 2018). Phenotype data and study materials are available on a permanent third-party archive, the 71 Inter-University Consortium for Political and Social Research (ICPSR). Requests to access data and study materials should be directed to the ICPSR[1]. For additional information regarding participant recruitment, compensation, and data collection, see Brim, Ryff, & Kessler (2004).

Only data from participants who were genotyped and predominantly of European ancestry were included in the present study (N = 1296). The average age of participants was approximately 54 years (median = 54 years, SD = 12.46 years, min. = 25 years, max. = 84 years),

[1] https://www.icpsr.umich.edu/icpsrweb

and approximately 51% of the sample was female (~ 49% male). There was considerable variation in highest level of education completed by participants (see Table 1).

Measures

The present study includes six focal constructs. (1) Educational attainment was measured using self-reports of the highest level of education completed by participants, rated on an ordinal scale. (2) BMI was calculated based on participants height and weight (mean = 28.79, median = 27.89, SD = 6.19, min. = 17.08, max. = 77.58)[2] (3) Smoking initiation was measured by asking participants whether they were ever a smoker or currently a smoker of cigarettes (No = 59%, Yes = 41%). (4-6) Polygenic scores for educational attainment, BMI, and smoking initiation were calculated using summary statistics from recent GWASs for each variable (Lee et al., 2018; Linnér et al., 2019; Locke et al., 2015).

Data Analytic Procedures

The ancestry of participants was estimated using Admixture software (Alexander, Novembre, & Lange, 2009) with a 1000 Genomes data (Phase 3) reference[3] using all 5 super-populations as a basis for estimation. To calculate ancestry component scores, genotype principal components analysis (PCA) was performed on participant genotypes, combined with 1000 Genome genotypes, after linkage disequilibrium (LD) pruning SNPs at a 0.2 R^2 threshold. Five ancestry component scores were calculated: European (EUR), East Asian (EAS), Ad-mixed American (AMR), Southeast Asian (SAS), and African (AFR).

To date, discovery GWASs have focused almost exclusively on participants of European ancestry. Consequently, the estimated effect sizes of individual SNPs are only known for

[2] There was a single outlier on BMI that was more than 5 standard deviations above than the mean. Results remain unchanged after excluding this observation.
[3] The Genome Project Consortium. A global reference for human genetic variation. (2015) *Nature*, 526, 68-78.

individuals of European ancestry, and the calculation of polygenic scores are only valid for participants of predominantly European ancestry. Therefore, to exclude ancestrally heterogeneous samples from the data, we excluded samples with less than 90% estimated European ancestry. Total of 1309 samples were included in imputation and polygenic risk scoring.

The Illumina OmniExpress arrays tag a sufficient number of variants on the X and Y chromosomes to determine biological sex (e.g. 17,707 SNPs on X chromosome and 1,367 on Y for array v. 1.1). Samples were excluded if self-reported sex did not match biological sex as determined by genotype (N = 13), as this may indicate either invalid self-reports, genotyping errors, or accidental I.D. swaps. After filtering out samples that did not pass ancestry- and sex-checks, PRSs were available for a final sample of N = 1296 participants. After MIDUS genotype samples were filtered via inclusion criteria, genotypes were imputed using minimac3 (Das et la., 2016) and Eagle (Loh et a., 2016) using the Haplotype Reference Consortium panel on the Michigan Imputation Server. SNPs with ambiguous strand orientation, >5% missing calls, or Hardy-Weinberg equilibrium $p < 0.001$ were excluded prior to imputation. After imputation, SNPs with minor allele frequency below 0.01 or an average call rate (AvgCall) below 0.9 were excluded. All genomic data were handled using Plink 1.9 (Purcell et al., 2007; Purcell & Chang, 2017).

Path analysis was conducted in Mplus version 8.1 (Muthén & Muthén, 2019), and missing data were handled using full-information maximum likelihood (Schafer & Graham, 2002). Because a subset of sibling- and twin-pairs are included in the current sample (N_{pairs} = 96), a family identification number was specified as a cluster variable (Muthén & Muthén, 2019) in path models to implement a Huber-White sandwich estimator, which adjusts the standard

errors of path coefficients for the non-independence of observations that results from a subset of participants being nested within the same family.

Age (centered at 54 years) and biological sex (coded female = 0, male = 1) were included as exogenous covariates of all focal study variables, in addition to the first five genomic principal component scores. Thus, we report results from fully-saturated models (i.e. df = 0). As the variance of certain PC scores approached zero, all PC scores were increased by a factor of 100 to avoid a singular observed covariance matrix of independent variables. The effects of age and biological sex on BMI and smoking initiation are included in Figures 2 and 3, but pathways from PC covariates were omitted from diagrams to ease visualization. As BMI and smoking initiation are continuous and binary outcomes, the estimated pathways to BMI and smoking initiation can be interpreted as linear and logistic[4] regression coefficients, with linear coefficients standardized and logistic coefficients exponentiated (i.e. reported as odds ratios). 99% confidence intervals are reported below their respective point estimates. Polygenic scores, self-reports of educational attainment, and BMI were standardized prior to fitting path models (M = 0, SD = 1).

Results

Results for educational attainment and BMI are reported in Figure 2. Results for educational attainment and smoking are reported in Figure 3. In both models, polygenic propensity for educational attainment was associated with educational attainment ($\beta = .27$, SE = .03, $p < .001$), providing evidence for gene-environment correlation. Providing evidence for pleiotropic effects, polygenic propensity for educational attainment was negatively correlated with polygenic risk for high BMI ($r = -.17$, SE = .03, $p < .001$) and polygenic risk for smoking initiation ($r = -.16$, SE = .03, $p < .001$). Providing a partial control for pleiotropic effects,

[4] All substantive results remain unchanged if smoking initiation is treated as a count variable with pathways estimated as Poisson regressions.

polygenic risk for high BMI was associated with BMI ($\beta = .23$, SE = .03, $p < .001$), and polygenic risk for smoking initiation was associated with smoking initiation (OR = 1.30, SE = .08, $p < .001$). After accounting for these direct associations, the pathway from polygenic propensity for educational attainment to BMI approached zero ($\beta = -.01$, SE = .03, $p = .691$), as did the pathway from polygenic propensity for educational attainment to smoking initiation (OR = 0.94, SE = .06, $p = .362$. These estimates suggest that the regression of BMI and smoking initiation on their respective polygenic scores provided an adequate statistical control for the pleiotropic effects of polygenic risk for educational attainment.

Notably, after regressing educational attainment on polygenic propensity for educational attainment, the direct association between polygenic propensity for BMI and education attainment was not strong or significantly different from zero ($\beta = -.03$, SE = .03, $p = .240$). However, even after regressing educational attainment on polygenic propensity for educational attainment, polygenic propensity for smoking initiation was directly and negatively associated with educational attainment ($\beta = –.08$, SE = .03, $p = .003$). This direct association between polygenic propensity for smoking initiation and educational attainment shows that the genetic instrument for educational attainment, by itself, only provided a partial control for gene-environment correlation. The regression of the exposure on polygenic risk for the exposure *and* outcome, however, provides an additional test and control for gene-environment correlation that has not traditionally been implemented in Mendelian randomization studies. Nevertheless, even after estimating pleiotropy and polygenic propensity for the exposure *and* outcome, there was still a protective association of educational attainment on BMI ($\beta = -.11$, SE = .03, $p < .001$) and smoking initiation (OR = 0.71, SE = .05, $p < .001$). Moreover, the association between polygenic propensity for educational attainment and BMI was statistically accounted for by educational

attainment (indirect effect = -.03, 99% bias-corrected bootstrapped C.I. = -.05, -.01, p = .001), as was the association between polygenic propensity for educational attainment and smoking initiation (indirect effect = -.09, 99% bias-corrected bootstrapped C.I. = -.15, -.04, p < .001).

Discussion

The present study proposed the integration of two existing methods, genetic instrumental variable regression and path analysis, to account for pleiotropy in Mendelian randomization studies using multiple polygenic scores. The method was then evaluated using a putatively important environmental exposure (educational attainment) and two outcomes that are of interest to clinicians and epidemiologists alike (BMI and smoking initiation). Importantly, the present study demonstrates that education has a protective association with BMI and smoking initiation, even when controlling for potential genetic confounds via Mendelian randomization *and* pleiotropic effects using multiple polygenic scores. Moreover, for the two phenotypes examined, controls for pleiotropy were effective, such that the direct pathways from polygenic propensity for education to BMI and smoking initiation approached zero, indicating that the proposed method is capable of addressing the exclusion criteria for a sound instrumental variable. In addition, polygenic risk for smoking initiation (but not BMI) was directly associated with educational attainment, even after accounting for polygenic propensity for educational attainment. This demonstrates that, at least for some phenotypes, traditional Mendelian randomization studies provide only a partial genomic control for the environmental exposure. The method proposed and implemented in the current study, however, provides an additional test and statistical control for potential gene-environment correlations, beyond what is typically accomplished in a Mendelian randomization study.

Of course, genetic path analysis is not without limitations. For one, it can only be applied to a Mendelian randomization study for which GWAS summary statistics are available for both the exposure and outcome. In addition, although polygenic scores have become potent predictors of their respective phenotypes, especially in comparison to single genetic variants, the arrays typically included in GWASs only tag point mutations (i.e. single nucleotide polymorphisms) and do not include insertion, deletions, and copy number variants. Further, the beta weights obtained from discovery GWASs are estimated with imprecision, and, consequently, polygenic scores provide only an imperfect proxy of genetic liability. Therefore, the strength of the proposed method depends on the size and overall quality of the discovery GWASs for the exposure and outcome of interest, though the quality of the GWASs for the phenotypes examined in the present study were reasonable by contemporary standards.

In many ways, the methodological integration that was proposed and implemented in the present study is an extension or specific instantiation of genomic structural equation modeling (Grotzinger et al., 2018). There are, however, important differences between genomic structural equation modeling and genetic path analysis as outlined in the present study. For example, genomic structural equation modeling is a technique that can be used to address a number of questions about the genetic architecture of complex phenotypes, including the search for SNPs not previously identified in a univariate GWAS. Alternatively, genetic path analysis using multiple polygenic scores was developed to address a limitation specific to Mendelian randomization studies and relies on the existence of discovery GWASs for the exposure and phenotype of interest. In addition, genomic structural equation modeling is based on genetic correlations estimated using a variant of LD-Score regression (Bulik-Sullivan et la., 2015), and genetic path analysis relies on multiple polygenic scores to estimate genetic correlations.

Genomic structural equation modeling also includes the estimation of latent variables that are not directly observed but, instead, are inferred indirectly from the data. Genetic path analysis, on the other hand, analyses associations between observed variables.

A remaining limitation to Mendelian randomization studies not addressed in the present study centers on the fact that, despite receiving a random assortment of genes from their parents, children's genotypes depend on their parents' genotype. Consequently, passive gene-environment correlations remain a possibility. Implementing genetic path analysis in a sample of siblings or twins would provide an additional control for this potential confound. Unfortunately, the sample analyzed in the present study did not include enough sibling-pairs to be adequately powered to fit the proposed path models to sibling-difference scores. Nevertheless, future studies may benefit from implementing genetic path analysis in larger samples of genotyped siblings with relevant exposures and outcomes measured. Finally, the present study did not address potential threats to the independence criteria for a sound instrument that is posed by any unmeasured confounder present in a non-experimental study. Despite these limitations, the present study provides compelling evidence for a complex set of gene-environment transactions that contribute to health-related outcomes in adulthood.

Acknowledgements

Special thanks to Colin G. DeYoung for helpful comments on an early presentation of this work. F.D.M.'s appointment was supported by a grant from the John Templeton Foundation, through the Genetics and Human Agency project. Since 1995 the MIDUS study has been funded by the following: John D. and Catherine T. MacArthur Foundation Research Network; National Institute on Aging (P01-AG020166); National institute on Aging (U19-AG051426).

Author Contributions

F.D.M. developed the proposed methodological integration, conducted analyses, and drafted the manuscript. A.A.S. & A.R.D. performed genotype calling, imputation, and polygenic scoring. R.F.K. contributed to the design of the study, obtained funding for the study, and supervised F.D.M. All authors provided critical revisions to manuscript and approved a final version.

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655-1664.

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, *44*, 512-525.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., ... & Daly, M. J. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, *47*, 1236-1241.

Brim, O. G., Ryff, C. D., & Kessler, R. C. (2004). *The MIDUS National Survey: An Overview*. University of Chicago Press.

Choi, S. W., Mak, T. S. H., & O'reilly, P. (2018). A guide to performing Polygenic Risk Score analyses. *BioRxiv*, 416545.

Briley, D. A., Livengood, J., & Derringer, J. (2018). Behaviour genetic frameworks of causal reasoning for personality psychology. *European Journal of Personality*, 202-220.

Das, S. *et al.* (2016) Next-generation genotype imputation service and methods. *Nature Genetics* 48, 1284-1287.

DiPrete, T. A., Burik, C. A., & Koellinger, P. D. (2018). Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. *Proceedings of the National Academy of Sciences*, *115*, E4970-E4979.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, *29*, 722-729.

Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., ... & Harden, K. P. (2018). Genomic SEM provides insights into the multivariate genetic architecture of complex traits. *BioRxiv*, 305029.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 621-638.

Harrell Jr, F. E., & Harrell Jr, M. F. E. (2019). Package 'Hmisc'. *CRAN2018*, 235-6.

Jaffee, S. R., & Price, T. S. (2007). Gene–environment correlations: A review of the evidence and implications for prevention of mental illness. *Molecular Psychiatry*, *12*, 432.

Kendler, K. S., & Baker, J. H. (2007). Genetic influences on measures of the environment: a systematic review. *Psychological Medicine*, *37*, 615-626.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, *27*, 1133-1163.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., ... & Fontana, M. A. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*, 1112-1121.

Linnér, R. K., Biroli, P., Kong, E., Meddens, S. F. W., Wedow, R., Fontana, M. A., ... & Nivard, M. G. (2019). Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics*, *51*, 245-255.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., ... & Croteau-Chonka, D. C. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197-206.

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., ... & Durbin, R. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*, 1443.

McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-303 (2010).

Muthén, L. K., & Muthén, B. (2019). Mplus. *The comprehensive modelling program for applied researchers: user's guide*, *5*.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*, 559-575.

Purcell, S., Chang, C., NIH-NIDDK Laboratory of Biological Modeling & Purcell Lab at Mount Sinai School of Medicine. (2017) PLINK 1.9 beta.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype→ environment effects. *Child Development*, 424-435.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147-177.

Shabalin, A., Clark, S., Hattab, M., Aberg, K. & van den Oord, E. (2017) *RaMWAS: Fast Methylome-Wide Association Study Pipeline for Enrichment Platforms* . R package version 1.2.0.

Smith, G.D., & Ebrahim, S. (2003). ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, *32*(1), 1-22.

Smith, G. D., & Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, *33*, 30-42.

van Kippersluis, H., & Rietveld, C. A. (2017). Pleiotropy-robust Mendelian randomization. *International Journal of Epidemiology*, *47*, 1279-1288.

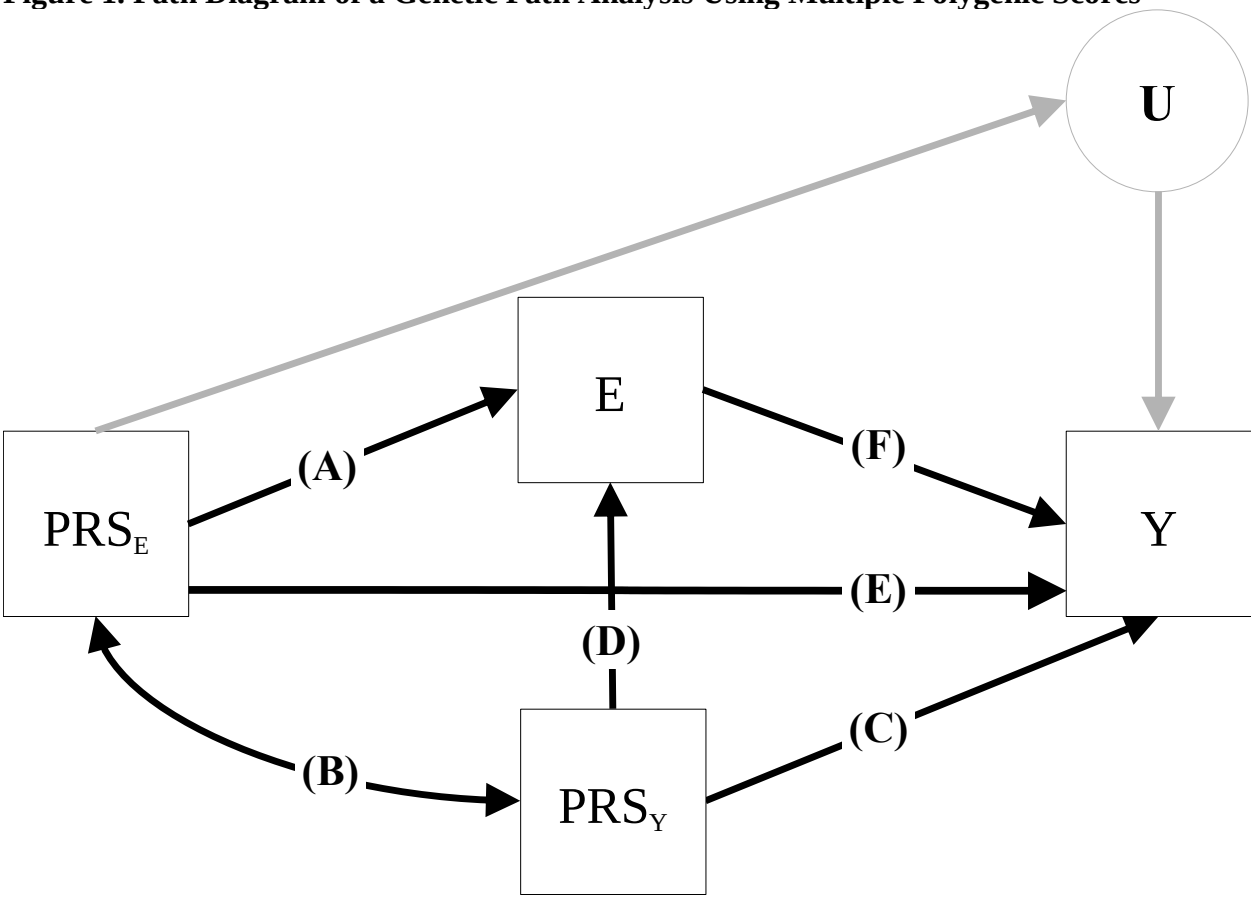
Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., ... & Yang, J. (2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications*, *9*, 224.

Table 1. Highest Level of Education Completed by Participants

				Number of Observations									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	NA
Frequency	1	6	26	10	199	195	59	110	325	54	232	70	6
Percent	< 1%	< 1%	~2%	< 1%	~15%	~15%	~5%	~9%	~25%	~4%	~18%	~5%	< 1%

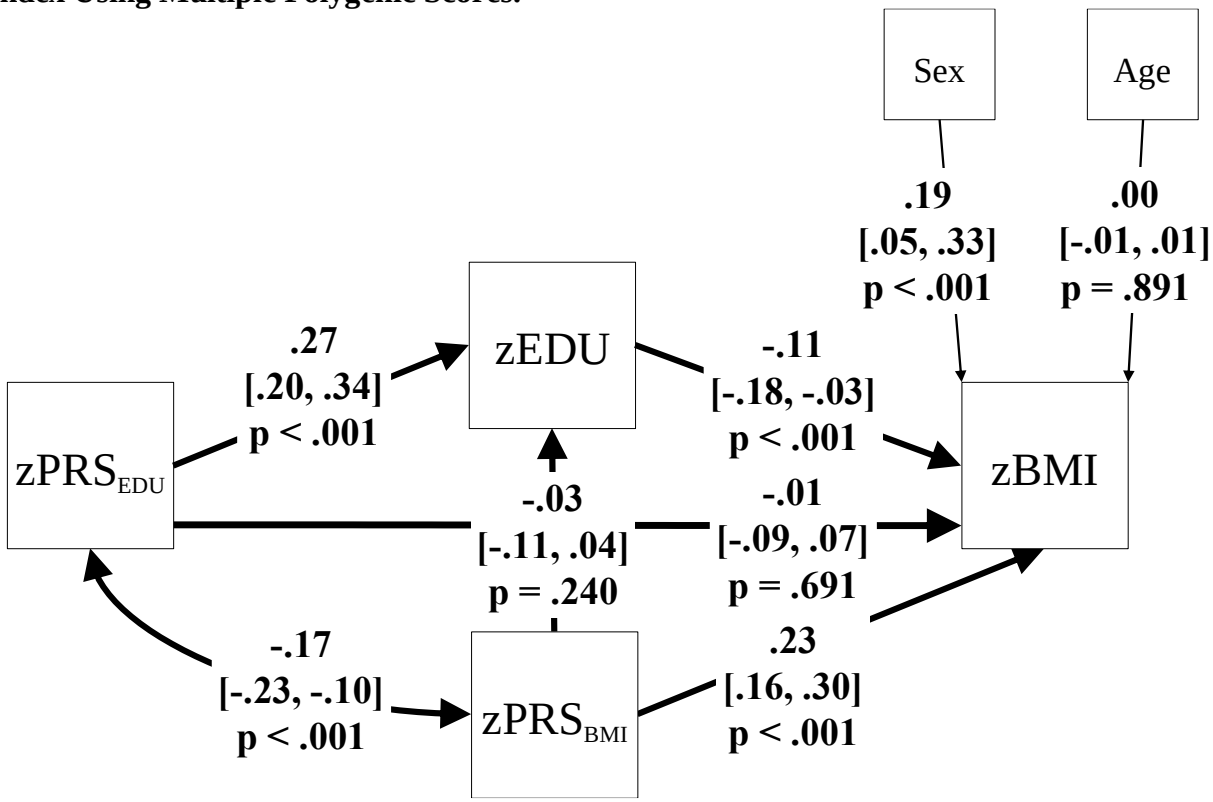
Notes. (1) = No school/some grade school (grades 1-6). (2) = Eighth grade/junior high school (grades 7-8). (3) = Some high school (grades 9-12, No Diploma or GED). (4) = GED (general education diploma). (5) = Graduated from high school. (6) = One to two years of college, no degree yet. (7) = Three or four years of college, no degree yet. (8) = Graduated from two years of college, vocational school, or obtained assoc. degree. (9) = Graduated from a four- or five-year college or obtained a bachelor's degree. (10) = Attended some graduate school, no graduate degree yet. (11) = Master's degree. (12) = PH.D., ED.D., MD, DDS, LLB, LLD, JD, etc. NA = missing values.

Figure 1. Path Diagram of a Genetic Path Analysis Using Multiple Polygenic Scores



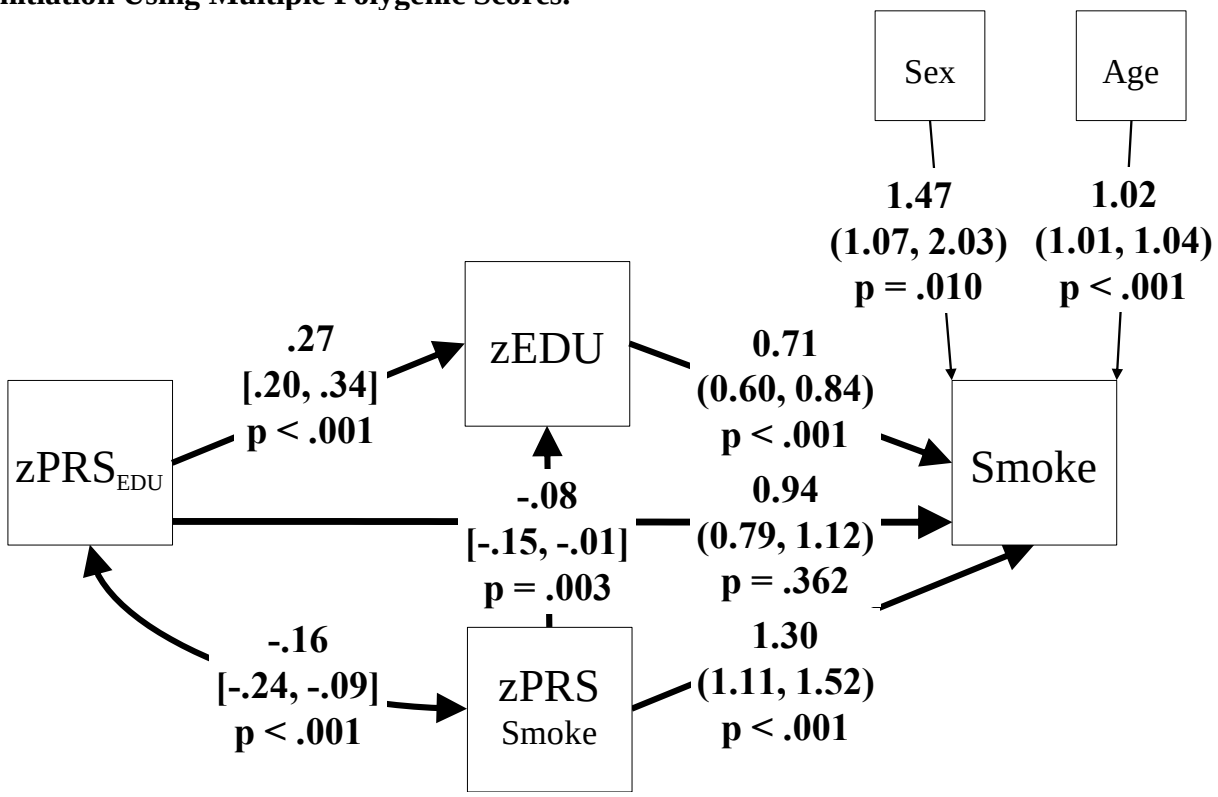
Notes. PRS = polygenic score. E = exposure. Y = outcome. (A) regression of the exposure on polygenic risk for the exposure- i.e. test of gene-environment correlation. (B) correlation between polygenic risk for the exposure and polygenic risk for the outcome- i.e. test of pleiotropy. (C) regression of the outcome on polygenic risk for the outcome. - i.e. statistical control for pleiotropic effects. (D) regression of the exposure on polygenic risk for the outcome- i.e. additional test and control for gene-environment correlation. (E) regression of the outcome on polygenic risk for the exposure, accounting for the exposure and polygenic risk for the outcome- i.e. a test of the control for pleiotropy. (F) regression of the outcome on the exposure, controlling for gene-environment correlations and pleiotropy.

Figure 2. Results of a Genetic Path Analysis of Educational Attainment and Body Mass Index Using Multiple Polygenic Scores.



Notes. The double-headed arrow represents a correlation. Single-headed arrows represent regressions. Unstandardized estimates are reported. All variables are standardized (M = 0, SD = 1), excluding biological sex (coded 0 = Female, 1 = Male) and age (years). Therefore, coefficents are intrepetted as the predicted standard deviation increase in BMI given a one unit increase in the predictor (e.g. a one standard deviation increase in polygenic risk or education, a one year increase in age, or being male instead of female). 99% confidence intervals are reported below parameter estimates. p = probability of the observed data if the null hypothesis is true. All focal variables were regressed on age, sex, and PC scores, but these pathways were omitted to ease visualization.

Figure 3. Results of a Genetic Path Analysis of Educational Attainment and Smoking Initiation Using Multiple Polygenic Scores.



Notes. The double-headed arrow represents a correlation. Single-headed arrows represent regressions. Unstandardized estimates are reported. All variables are standardized (M = 0, SD = 1), excluding smoking initiation (coded 0 = No, 1 = Yes), biological sex (coded 0 = Female, 1 = Male) and age (years). To help ease interpretation of results, estimates for pathways to smoking initiation are reported as odds ratios, intrepetted as the increased odds of having initiated smoking given a one unit increase in the predictor (i.e. a one standard deviation increase in polygenic risk or education, a one year increase in age, or being male instead of female). 99% confidence intervals for odds ratios and betas are reported in parentheses and brackets, respectively. p = probability of the observed data if the null hypothesis is true. All focal variables were regressed on age, sex, and PC scores, but these pathways were omitted to ease visualization.