Using multiple polygenic scores and 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

*Direct correspondence to Frank D Mann (fmann@umn.edu), University of Minnesota, Department of Psychology, Elliot Hall, 7 E River Road, Minneapolis, MN 55455

Abstract

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 potent predictors, satisfying the relevance criteria of an instrumental variable. Evidence for 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, we test whether educational attainment is associated with two health-related outcomes in adulthood, body mass index and smoking initiation, while estimating and controlling for both gene-environment correlations and pleiotropy. Results provide compelling evidence for a complex set of gene-environment transactions that undergird the relations between educational attainment and health-related outcomes in adulthood. Moreover, results are consistent with education having a protective effect on body mass index and smoking initiation. Although there is no perfect design for causal inference in epidemiological research, path analysis of polygenic scores for an exposure *and* outcome is capable of addressing the exclusion criteria for a sound instrumental variable and, consequently, has the potential to help advance Mendelian randomization studies.

Keywords: Mendelian randomization; pleiotropy; education; BMI; smoking;

Key Messages

- Path analysis of multiple polygenic scores is used to test and control for pleiotropy in a Mendelian randomization study.
- Results provide compelling evidence for a complex set of gene-environment transactions that undergird the relations between educational attainment and health-related outcomes in adulthood.
- Results are consistent with education having a protective environmentally-mediated effect on body mass index and smoking initiation
- The proposed path model is capable of addressing the exclusion criteria for an instrumental variable and, consequently, has the potential to help advance Mendelian randomization studies of complex outcomes.

Using multiple polygenic scores and 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 (1). This results in distributions of genes that are independent of many factors that often confound associations documented in observational studies (2,3). 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 a path analysis of multiple polygenic scores, a simple extension of genetic instrumental variable regression (4), 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 multiple polygenic scores and 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 (5,6). This process results in heritable variation in measures of the environment (7), 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 (8). 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" 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 (9). 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" (9). 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, it is not clear that polygenic scores meet the necessary criteria for a sound instrumental variable (4,10,11).

There are three criteria for a sound instrumental variable (12). 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 (10) and summary data-based multiple regression (13), as well as pleiotropy-robust Mendelian randomization (11) and genetic instrumental variable regression (4). The present study intends to contribute to this body of work by demonstrating how an existing method (i.e. path analysis), in combination with multiple polygenic scores, can be used to estimate and control for pleiotropy in a Mendelian randomization study.

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 pleiotropy, 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.

[FIGURE 1 HERE]

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 correlations. 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.

The potential for third-variable confounding (i.e. violation of the independence criteria) can be partially accounted for by introducing exogenous covariates to the model, specifically by regressing the exposure (E), outcome (Y), and both genetic instruments (PRS_E & PRS_Y) on a set of independent variables that are hypothesized to be related to two or more of the focal study constructs. Hereinafter, we provide a demonstration of this approach 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).(14). Data was prepared for analyses with R version 3.5.2. Data was imported into R using the 'Hmisc' package (15), preprocessed, and then exported from R using the 'MplusAutomation' package version 0.7.1 (16). Phenotype data and study materials are available on a permanent third-party archive, the 71 Inter-University Consortium for Political and Social Research (ICPSR). Additional information regarding participant recruitment, compensation, and data collection can be found elsewhere (14). 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), and approximately 51% of the sample was female (~ 49% male). There was considerable variation in highest level of education completed by participants and their parents, as well as total household income (see Table 1).

[TABLE 1 HERE]

Measures

The present study includes six focal constructs. Household income and educational attainment for participants and parents were measured using participant-reports of the highest level of education completed (rated on an ordinal scale) and total household income (in U.S. dollars). BMI was calculated based on participants height and weight (mean = 28.79, median = 27.89, SD = 6.19, min. = 17.08, max. = 77.58). There was a single outlier on BMI (z-score = 7.77); Effect sizes are similar, and results of null hypothesis significance tests remain unchanged after excluding this observation. Smoking initiation was measured by asking participants whether they were ever a smoker or currently a smoker of cigarettes (No = ~61%, Yes = ~39%). Polygenic scores for educational attainment, BMI, and smoking initiation were calculated using summary statistics from recent GWASs for each variable (17-19).

Data Analytic Procedures

The ancestry of participants was estimated using Admixture software (20) with a 1000 Genomes data (Phase 3) reference (21) 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 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.

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 = 1189 unrelated participants[1]. Genotypes were imputed using minimac3 (22) and Eagle (23) 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 genetic data were handled using Plink 1.9 (24-25). Finally, observations were excluded from analyses if data were missing for one or more exogenous covariate (i.e. age, sex, household income, and highest level of education completed by mothers and fathers), yielding a final analytic sample of (N = 1069) participants.

Path analysis was conducted in Mplus version 8.1 (26), and missing data were handled using full-information maximum likelihood (27). Age (centered at 54 years), biological sex (coded female = 0, male = 1), total household income, and the highest level of education completed by mothers and fathers were included as exogenous covariates of focal study variables, in addition to the first five genetic principal component scores. Thus, we report results from fully-saturated models (i.e. model degrees of freedom = 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. Polygenic scores, self-reports of

[1] Observations from a subset of biological siblings were omitted from the data.

educational attainment, parents' educational attainment, household income, and BMI were standardized before fitting path models (M = 0, SD = 1). As BMI and smoking initiation are continuous and binary outcomes, the estimated pathways to BMI and smoking initiation can be interpreted as linear and Poisson regression coefficients, with linear coefficients standardized and Poisson coefficients exponentiated (i.e. reported as risk ratios). 99% non-parametric bootstrapped confidence intervals are reported below their respective point estimates.

Results

[FIGURE 2 HERE]

Results for educational attainment and BMI are reported in Figure 2. Results for educational attainment and smoking initiation are reported in Figure 3. The effects of exogenous covariates are reported in Table 2. In both models, polygenic propensity for educational attainment was associated with educational attainment (β = .17, SE = .03, p < .001). Providing evidence for pleiotropy, polygenic propensity for educational attainment was negatively correlated with polygenic risk for high BMI (r = -.14, SE = .03, p < .001) and negatively correlated with polygenic risk for smoking initiation (r = -.14, SE = .03, p < .001). Providing a partial control for pleiotropy, polygenic risk for high BMI was associated with BMI (β = .22, SE = .03, p < .001), and polygenic risk for smoking initiation was associated with smoking initiation (RR = 1.15, SE = .04, p < .001). After accounting for these associations, the pathway from polygenic propensity for educational attainment to BMI approached zero (β = -.01, SE = .03, p = .808), as did the pathway from polygenic propensity for educational attainment to smoking initiation (RR = 0.99, SE = .04, p = .772). 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.

[FIGURE 3 HERE]

Notably, after regressing educational attainment on polygenic propensity for educational attainment, the association between polygenic propensity for BMI and education attainment approached zero (β = -.02, SE = .03, p = .467), but polygenic propensity for smoking initiation was negatively associated with educational attainment (β = -.07, SE = .03, p = .008). 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, and controlling for the effects of age, biological sex, household income, mother's education, father's education, and the first five genetic principal components, there was still a protective association of educational attainment with BMI (β = -.08, SE = .03, p = .027) and smoking initiation (RR = 0.88, SE = .04, p = .002).

[TABLE 2 HERE]

Discussion

The present study proposed using path analysis of multiple polygenic scores to account for pleiotropy in Mendelian randomization studies. The proposal was then evaluated using a putatively important environmental exposure and two outcomes that are of interest to clinicians and epidemiologists alike. Importantly, the present study demonstrates that education has a protective association with BMI and smoking initiation, even when controlling for demographic variables (e.g. age, biological sex, socio-economic status, etc.) and potential genetic confounds.

Moreover, for the two phenotypes examined, statistical controls for pleiotropy were effective, such that the direct pathways from polygenic propensity for education to BMI and smoking initiation approached zero, indicating that path analysis is capable of addressing the exclusion criteria for a sound instrumental variable when polygenic scores can be calculated for *both* the exposure and outcome of interest. 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, genetic instrumental variable regression may provide only a partial genetic control for an environmental exposure. The path analysis 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 traditional genetic instrumental variable regression.

Of course, the present study is not without limitations. For one, the path model in the presents study is only useful when GWAS summary statistics are available for both an exposure and associated outcome. 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 common point mutations (i.e. single nucleotide polymorphisms) and do not include rare variants, 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 common 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 addition, it is imperative that any sample of genotyped participants that is used to conduct a Mendelian randomization study not be included in initial GWAS discovery efforts for the exposure or outcome.

A remaining limitation to Mendelian randomization studies not addressed here centers on the fact that, despite receiving a random assortment of genes from their parents, children's genotypes depend on their parents' genotype. This pathway was captured in the current study (at least in part) by regressing polygenic scores for educational attainment on the highest level of education completed by parents, both mothers and fathers, which evinced positive and statistically significant associations. Nevertheless, passive gene-environment correlations remain a possibility. Implementing a path analysis of multiple polygenic scores in a sample of siblings or twins would provide an additional control for this pathway. 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.

Although the size of the present study is small for the purpose of conducting a discovery GWAS, the present study was adequately powered to fit the relevant path models according to conventional standards (28). Finally, the present study did not fully address potential threats to the independence criteria for a sound instrumental variable that is posed by an *unmeasured* confounder. However, this limitation is not specific to Mendelian randomization studies but is a more general limitation that applies to all non-experimental studies. The present study did control for a number of *measured* confounders, including, age, biological sex, total household income, mothers' and fathers' education, and the first five genetic principal components.

Controlling for these covariates, the present study found evidence for a complex set of gene-environment transactions that contribute to important health-related outcomes in adulthood. Crucially, data were consistent with education having a direct environmentally-mediated protective impact on BMI and smoking initiation. Although there is no perfect design for unambiguous causal inference in epidemiological research, the proposed path model offers an advance to Mendelian randomization studies of complex outcomes.

Funding

The work 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

FDM developed the idea for the study, conducted analyses, and drafted the manuscript. AAS & ARD performed genotype calling, imputation, and polygenic scoring. RFK contributed to the design of the study, obtained funding for the study, and supervised FDM. All authors provided critical revisions to manuscript and approved a final version.

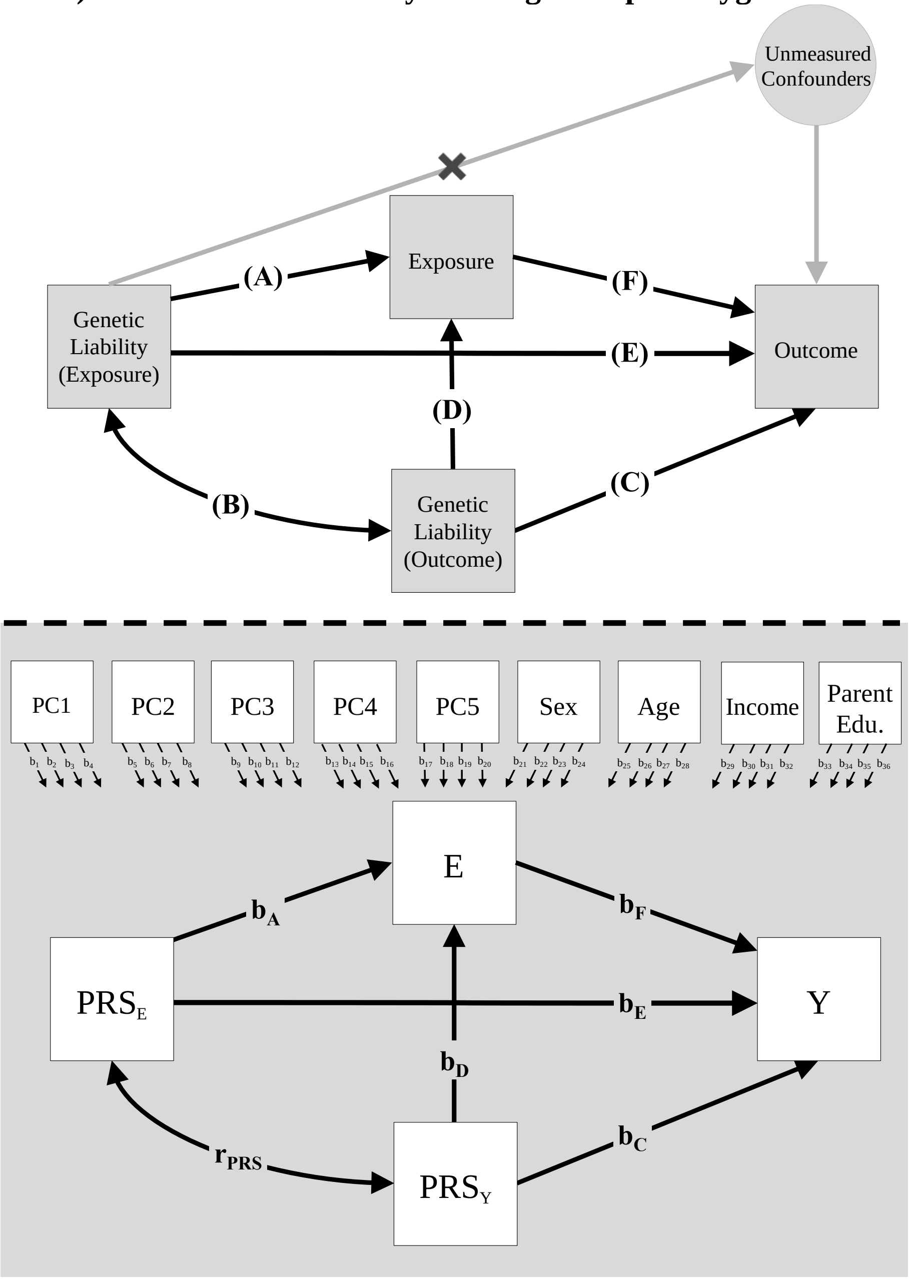
Table 1. Total Household Income and Highest Level of Education Completed by Participants & Parents

Total	N	NA	Mean	Median	SD	Minimum	Maximum
Income	1158	40	$86,274	$72,875	$61,866.41	$0	> $300,00

Level of					Order-Categorical Response									
Education		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	NA
Participant:	Frequency	1	6	22	10	182	175	57	101	296	51	215	68	5
	Percent	< 1%	< 1%	~2%	< 1%	~15%	~15%	~5%	~9%	~25%	~4%	~18%	~6%	< 1%
Mother:	Frequency	28	95	128	15	440	108	10	107	161	6	55	9	27
	Percent	~2%	~8%	~11%	~1%	~37%	~9%	~1%	~9%	~13%	< 1%	~5%	< 1%	~2%
Father:	Frequency	68	141	112	10	301	98	14	71	162	9	78	44	81
	Percent	~6%	~12%	~9%	~1%	~25%	~8%	~1%	~6%	~14%	~1%	~7%	~4%	~7%

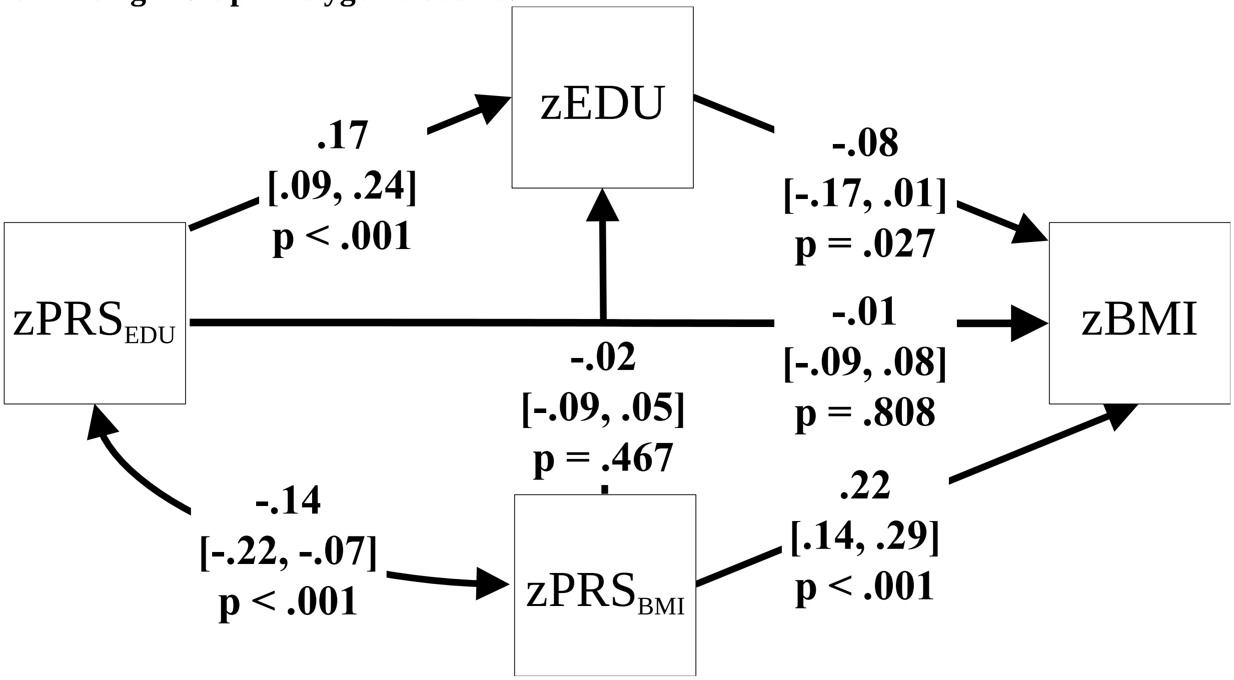
Notes. (1) = No school/some grade school (grades 1-6). (2) = Eighth grade/junior high school (grades 7-8). (3) = Some high school (grades 9-12, No Diploma or GED). (4) = GED (general education diploma). (5) = Graduated from high school. (6) = One to two years of college, no degree yet. (7) = Three or 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. Conceptual Diagram (Top Panel) and Path Diagram (Bottom Panel) of a Genetic Path Analysis Using Multiple Polygenic Scores



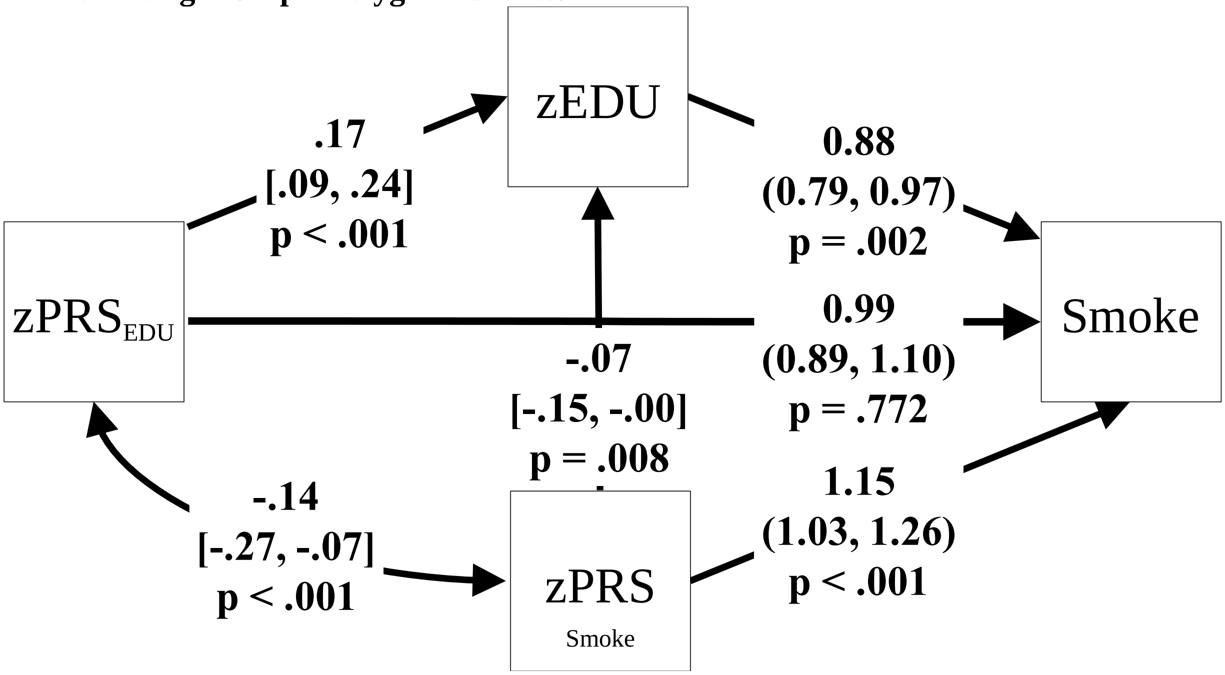
Notes. *Top panel*: (A) test of gene-environment correlation. (B) test of pleiotropy. (C) statistical control for pleiotropy. (D) additional test for gene-environment correlation. (E) test of statistical control for pleiotropy. (F) test of quasi-causal effect of the exposure. The "X" on the pathway to unmeasured confounders reflects the independence criteria of a sound instrument. *Bottom panel*: PRS = polygenic score. E = measure of exposure. Y = measure of outcome. PC = principal component. $b_1 – b_i$ = effects of covariates on focal variables truncated to ease presentation. b_A = regression of exposure on polygenic risk for the exposure. r_{PRS} = correlation between polygenic risk for the exposure and polygenic risk for the outcome. b_C = regression of the outcome on polygenic risk for the outcome. b_D = regression of the exposure on polygenic risk for the outcome. b_E = regression of the outcome on polygenic risk for the exposure. b_F = regression of the outcome on the exposure.

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. All focal variables were standardized (M = 0, SD = 1). Therefore, coefficents are intrepetted as the predicted standard deviation increase in BMI given a standard deviation increase in polygenic risk or education. 99% bias-corrected bootstrapped confidence intervals are reported below parameter estimates. p = probability of the observed data if the null hypothesis is true (i.e. β = 0). All focal variables are regressed on age, sex, and PCs, but these pathways are omitted to ease visualization. See Table 2 for the effects of exogenous covariates.

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. All focal variables are standardized (M = 0, SD = 1). To help ease interpretation of results, estimates for pathways to smoking initiation are reported as risk ratios, intrepetted as the increased risk of having initiated smoking given a one unit increase in the predictor (i.e. a standard deviation increase in polygenic risk or education). 99% bias-corrected bootstrapped confidence intervals for risk ratios (RR) and betas [β] are reported in parentheses and brackets, respectively. p = probability of the observed data if the null hypothesis is true (i.e. β = 0 or RR = 1). All focal variables are regressed on age, sex, and PCs, but these pathways are omitted to ease visualization. See Table 2 for the effects of exogenous covariates

Table 2. Effects of Exogenous Covariates on Focal Study Variables

	Outcome = BMI											
	PRS$_E$			PRS$_Y$			Exposure (E)			Outcome (Y)		
	b	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*
Age	.01	.00	.030	-.01	.00	.013	.00	.00	.135	-.00	.00	.496
Sex	.06	.06	.327	-.09	.06	.125	.14	.05	.009	.17	.06	.004
MEDU	.13	.04	< .001	-.06	.04	.112	.15	.03	< .001	.06	.04	.099
FEDU	.11	.04	.004	-.07	.04	.044	.21	.03	< .001	-.13	.04	.001
Income	.09	.03	.002	.00	.03	.909	.23	.02	< .001	-.01	.03	.624
PC1	-.51	.84	.537	.839	.76	.267	-.22	.69	.746	-.12	.65	.856
PC2	-.96	.79	.228	-2.29	.76	.003	.13	.66	.840	.73	.75	.329
PC3	-.07	.11	.513	.22	.12	.062	.08	.09	.368	.29	.10	.005
PC4	-.06	.16	.697	-.11	.15	.474	-.00	.14	.992	.02	.16	.883
PC5	.01	.03	.777	-.15	.03	< .001	-.04	.03	.171	.03	.02	.188

	Outcome = Smoking											
	PRS$_E$			PRS$_Y$			Exposure (E)			Outcome (Y)		
	b	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*RR*	*SE*	*p*
Age	.01	.00	.030	.00	.00	.504	.00	.00	0.09	1.02	.00	< .001
Sex	.06	.06	.327	-.02	.06	.723	.14	.05	.008	1.22	.09	.010
MEDU	.13	.04	< .001	-.08	.04	.046	.14	.03	< .001	0.97	.04	.574
FEDU	.11	.04	.004	.00	.04	.998	.21	.03	< .001	0.96	.04	.343
Income	.09	.03	.002	-.04	.03	.280	.23	.02	< .001	0.93	.04	.103
PC1	-.51	.83	.537	.99	.78	.204	-.17	.679	.801	2.89	3.66	.226
PC2	-.96	.79	.228	.73	.71	.301	.23	.644	.724	2.78	5.47	.260
PC3	-.07	.11	.513	.03	.11	.806	.08	.093	.382	1.00	.15	.987
PC4	-.06	.16	.697	-.10	.16	.538	-.01	.136	.969	0.93	.17	.687
PC5	.01	.03	.777	-.03	.03	.226	-.04	.027	.168	0.97	.03	.343

Notes. *b* = multiple regression coefficient. RR = risk ratio. *SE* = standard error. *p* = probability of the observed data if the null hypothesis is true (i.e. *b* = 0). MEDU = highest level of education completed my mother. FEDU = highest level of education completed by father. Income = total household income. PC1 – PC5 = first five genetic principal components.

References

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

[2] Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statistics in Medicine. 2008 Apr 15;27(8):1133-63.

[3] Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. International Journal of Epidemiology. 2004 Feb 1;33(1):30-42.

[4] DiPrete TA, Burik CA, Koellinger PD. Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. Proceedings of the National Academy of Sciences. 2018 May 29;115(22):E4970-9.

[5] Briley DA, Livengood J, Derringer J. Behaviour genetic frameworks of causal reasoning for personality psychology. European Journal of Personality. 2018 May;32(3):202-20.

[6] Jaffee SR, Price TS. Gene–environment correlations: A review of the evidence and implications for prevention of mental illness. Molecular Psychiatry. 2007 May;12(5):432.

[7] Kendler KS, Baker JH. Genetic influences on measures of the environment: a systematic review. Psychological Medicine. 2007 May;37(5):615-26.

[8] Scarr S, McCartney K. How people make their own environments: A theory of genotype→ environment effects. Child Development. 1983 Apr 1:424-35.

[9] Choi SW, Mak TS, O'reilly P. A guide to performing Polygenic Risk Score analyses. BioRxiv. 2018 Jan 1:416545.

[10] Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology. 2015 Apr 1;44(2):512-25.

[11] Rietveld CA. Pleiotropy-robust Mendelian randomization. International Journal of Epidemiology. 2018 Aug;47(4):1279-88.

[12] Greenland S. An introduction to instrumental variables for epidemiologists. International Journal of Epidemiology. 2000 Aug 1;29(4):722-9.

[13] Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, Robinson MR, McGrath JJ, Visscher PM, Wray NR, Yang J. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nature Communications. 2018 Jan 15;9(1):224.

[14] Ryff OG, Kessler RC. How healthy are we?: A national study of well-being at midlife. University of Chicago Press; 2004 Jan 15.

[15] Harrell Jr FE, Harrell Jr MF. Package 'Hmisc'. CRAN2018. 2019 Jan 25:235-6.

[16] Hallquist MN, Wiley JF. MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in M plus. Structural equation modeling: a multidisciplinary journal. 2018 Jul 4;25(4):621-38.

[17] Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, Nguyen-Viet TA, Bowers P, Sidorenko J, Linnér RK, Fontana MA. Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. Nature Genetics. 2018 Aug;50(8):1112.

[18] Linnér RK, Biroli P, Kong E, Meddens SF, Wedow R, Fontana MA, Lebreton M, Tino SP, Abdellaoui A, Hammerschlag AR, Nivard MG. 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. 2019 Feb;51(2):245.

[19] Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, Powell C, Vedantam S, Buchkovich ML, Yang J, Croteau-Chonka DC. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015 Feb;518(7538):197.

[20] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009 Sep 1;19(9):1655-64.

[21] Consortium GP, Auton A, Brooks LD. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

[22] Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D. Next-generation genotype imputation service and methods. Nature genetics. 2016 Oct;48(10):1284

[23] Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R. Reference-based phasing using the Haplotype Reference Consortium panel. Nature genetics. 2016 Nov;48(11):1443.

24 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007 Sep 1;81(3):559-75.

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

[26] Muthén LK, Muthén B. Mplus. The comprehensive modelling program for applied researchers: user's guide. 2019 Mar 24;5.

[27] Enders CK, Bandalos DL. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Structural equation modeling. 2001 Jul 1;8(3):430-57.

[28] Jackson DL. Revisiting sample size and number of parameter estimates: Some support for the N: q hypothesis. Structural equation modeling. 2003 Jan 1;10(1):128-41.