PENN INSTITUTE for ECONOMIC RESEARCH
UNIVERSITY of PENNSYLVANIA

The Ronald O. Perelman Center for Political Science and Economics (PCPSE)
133 South 36th Street
Philadelphia, PA 19104-6297

pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper
20-024

Genetic Risks, Adolescent Health and Schooling Attainment

VIKESH AMIN
Central Michigan University

JERE R. BEHRMAN
University of Pennsylvania

JASON M. FLETCHER
University of Wisconsin-Madison

CARLOS A. FLORES
California Polytechnic State

ALFONSO FLORES-LAGUNES
University Syracuse University

HANS-PETER KOHLER
University of Pennsylvania

June 28, 2020

https://ssrn.com/abstract=3638250

Genetic Risks, Adolescent Health and Schooling Attainment

Vikesh Amin[1]; Jere R. Behrman[2]; Jason M. Fletcher[3]; Carlos A. Flores[4]; Alfonso Flores-Lagunes[5]; Hans-Peter Kohler[6]

[1] Department of Economics, Central Michigan University
[2] William R. Kenan, Jr. Professor of Economics and Sociology, University of Pennsylvania
[3] Department of Sociology and La Follette School of Public Affairs, University of Wisconsin-Madison, NBER, and IZA
[4] Department of Economics, California Polytechnic State University
[5] Department of Economics, Syracuse University, IZA, and GLO
[6] Fredrick J. Warren Professor of Demography, University of Pennsylvania

Abstract:
We provide new evidence on the effect of adolescent health behaviors/outcomes (obesity, depression, smoking, and attention deficit hyperactivity disorder (ADHD)) on schooling attainment using the National Longitudinal Study of Adolescent to Adult Health. We take two different approaches to deal with omitted variable bias and reverse causality. Our first approach attends to the issue of reverse causality by using health polygenic scores (PGSs) as proxies for actual adolescent health. Second, we estimate the effect of adolescent health using sibling fixed-effects models that control for unmeasured genetic and family factors shared by siblings. We use the PGSs as additional controls in the sibling fixed-effects models to reduce concerns about residual confounding from sibling-specific genetic differences. We find consistent evidence across both approaches that being genetically predisposed to smoking and smoking regularly in adolescence reduces schooling attainment. We find mixed evidence for ADHD. Our estimates suggest that having a high genetic risk for ADHD reduces grades of schooling, but we do not find any statistically significant negative effects of ADHD on grades of schooling. Finally, results from both approaches show no consistent evidence for a detrimental effect of obesity or depression on schooling attainment.

Key Words: adolescent health; polygenic scores; education

JEL Codes: I21; I10

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis. The authors acknowledge research funding from NIH grant number 1R01HD094011-01. We thank Daniel Eisenberg and Petri Böckerman for useful comments, and Christian Failla for research assistance.

1. Introduction

There is a large empirical literature showing that poor health in childhood and adolescence is associated with lower schooling attainment (see Prinz et al. 2018 for an extensive literature review). However, it is difficult to establish whether there is a causal relationship because of (1) unobserved genetic and early-life factors that affect both health and schooling attainment, which confounds estimates, and (2) reverse causality insofar as childhood and adolescent schooling affect health. Several studies (reviewed in section 2) have contributed to the literature by attempting to control for confounding using methods such as sibling fixed-effects, while leaving unresolved issues of reverse causality. The results from these fixed-effect studies suggest that there may be a causal effect of poor childhood and adolescent mental health on schooling attainment, whereas there is not much evidence for a causal effect of poor physical health.

We contribute to the literature by taking two different approaches to provide new evidence on the effects of adolescent health behaviors/outcomes (obesity, depression, smoking, and attention deficit hyperactivity disorder (ADHD)) on adult schooling attainment for a sample of European-ancestry individuals in the National Longitudinal Study of Adolescent to Adult Health (Add Health). Our first approach focuses on the issue of reverse causality by using health polygenic scores (PGSs), which are summary measures of an individual's genetic predisposition for a given trait, to be proxies for the corresponding adolescent health behavior/outcome. We use Ordinary Least Square (OLS) regressions to estimate the effect of these health PGSs on schooling, conditional on an educational attainment PGS and other controls. These estimates provide new evidence for the effect of adolescent health that are free of reverse causality, as later schooling outcomes cannot shape the PGSs, which are fixed at conception. They can be informative as to whether a causal effect exists. For example, if there were no causal effect of adolescent obesity on schooling attainment, then we would expect that having a high genetic risk of being obese does not affect schooling attainment. OLS estimates of the effect of PGSs, however, may not reflect a causal effect because of the unobserved family environment. We therefore also use the available sibling sample in Add Health to estimate the effect of the health PGSs using sibling fixed-effects regressions. The sibling fixed-effects estimates control for parental genetics and unobserved family factors, and are more likely to provide causal estimates than between-family estimates. Our second approach provides complementary evidence by estimating the effect of actual adolescent health using OLS regressions, while exploiting the richness of the Add Health data to control for sources of unobserved heterogeneity. We use the health and educational attainment PGSs to control for unobserved genetic heterogeneity, proxy for unobserved family characteristics (e.g., variables relating to mother's schooling, mother's health, and parental investments), and control for community environmental factors through community fixed effects. Like previous studies, we also compare these OLS estimates to sibling fixed-effects estimates that control for unmeasured genetic and family factors shared by siblings.[1] Including PGSs as additional controls is an improvement upon sibling fixed-effects estimates in previous studies, which have not been able to control for unobserved sibling-specific genetic heterogeneity. Using the PGSs as additional controls also makes the sibling fixed-effects approach closer in spirit to twins fixed-effects models with MZ (monozygotic; identical) twins.

[1] We could also estimate the effect of adolescent health by using the PGSs as instruments. We do not take this approach because of the difficulty in defending the exclusion restriction assumption requiring that PGSs only affect schooling through their relationship with adolescent health. For example, the exclusion restriction could be violated due to pleiotropy; that is, genes that affect adolescent health could also affect schooling either through other genes or other traits (von Hinke et al. 2016; van Kippersluis & Rietveld 2018; Fletcher 2018). Another possibility is to estimate sibling fixed-effects instrumental variable regressions using the PGSs as instruments. However, the first stage is not sufficiently powerful in the data.

We find consistent evidence across both approaches that adolescent smoking may reduce schooling attainment. OLS and sibling fixed-effects estimates show that being genetically predisposed to smoking has a negative effect on schooling attainment, with the latter estimates suggesting that a one-unit increase in the ever-smoke PGS results in 0.12 of a grade less of schooling and a lower probability of graduating from college of 8 percentage points. The negative effects of ever smoking regularly in adolescence result in 0.63 of a grade less of schooling and a lower probability of graduating from college of 20 percentage points, and are robust to controlling for unmeasured family and genetic factors. We find mixed evidence for ADHD. Our estimates show that having a high genetic risk of ADHD reduces schooling attainment, but the effect of actual ADHD during adolescence is small and statistically insignificant. Consistent with the literature, we find little evidence of negative effects of adolescent obesity in our two approaches.

2. Literature Review

Estimating causal effects of childhood/adolescent health on human capital outcomes is challenging. These processes are bi-directional and dynamic, which suggests the likely presence of many generic empirical challenges, including reverse causality and confounding from a variety of sources (environmental factors, family factors, individual factors). There are few examples of good instruments for child health—many influences on health cannot be validly excluded from human capital outcomes. This has left many researchers to focus on reducing confounding as a key way of advancing our knowledge in this area. To this end, most studies have used sibling fixed-effects models, which control for all unobserved family, genetic and environmental factors shared by siblings.

Sibling fixed-effects estimates from several studies show that measures of mental health (e.g., ADHD, conduct problems, depression) have statistically significant negative effects on schooling attainment. For example, the sibling fixed-effects estimates in Currie & Stabile (2006) indicate that a one-unit increase in hyperactivity scores increases the probability of grade retention by 10-12% in a sample of children aged 4-12 from the Canadian and American National Longitudinal Survey of Youth datasets. Using administrative data on 50,000 children and adolescents aged 18 and younger born in Manitoba, Canada, Currie et al. (2010) find that a diagnosis of ADHD or conduct disorders at ages 14-18 decreases the probability of being in grade 12 by age 17 by 19%. Using Add Health, Fletcher & Wolfe (2008) find that ADHD in childhood increases the probability of grade repetition, but does not affect grades of schooling at age 21. This suggests that the negative short-term consequences of ADHD may not lower schooling attainment in the longer run. Also using Add Health, Fletcher (2010) finds that one standard deviation higher adolescent depressive symptoms increases the probability of dropping out of high school by 25-30%. Smith & Smith (2010), using the Panel Study of Income Dynamics, find that having a psychological problem before age 17 decreases grades of schooling by about 0.33. Salm & Schunk (2012) use administrative data from school entrance medical examinations in Osnabruck, Germany, and find that childhood mental health problems lower cognitive ability test scores at age 6 by 10%. Although sibling fixed effects is a powerful approach to control for unobserved confounders shared by siblings, it still suffers from residual confounding from unobserved sibling-specific differences such as in innate ability or health endowments.[2]

[2] In an attempt to account for this residual confounding, Fletcher & Lehrer (2009, 2011) use Add Health and employ a sibling fixed-effects instrumental-variable design, where they use genetic markers to instrument for sibling differences in adolescent health. Although there may be concerns about the validity of the exclusion restriction (e.g., von Hinke et al. 2016; van Kippersluis & Rietveld 2018; Fletcher 2018), in their 2009 paper they find that having higher inattentive symptoms decreases grades of schooling at age 21 by 3.5 grades. They also find large negative effects for ADHD (over 2 grades) and depression (over 1 grade) on grades of schooling, but the standard errors are too large to rule out null effects. In their 2011 paper, they find that ADHD has a statistically significant negative effect on cognitive test scores in adolescence.

While there is consistent and robust evidence across studies and datasets that poor mental health in childhood and adolescence has negative effects on schooling attainment, the evidence for poor physical health is much weaker. Sabia & Rees (2015) look at the effect of adolescent body mass index (BMI) on college graduation in Add Health. They instrument BMI with the BMI of the biological sibling, which controls for sibling-specific but not shared-sibling unobserved factors, and whether the mother reports being obese, which controls for some shared-sibling factors. Their estimates indicate that a higher adolescent BMI lowers the probability of completing college for both men and women. The studies by Currie et al. (2010), and Salm & Schunk (2012) also examine the effects of physical health measures on educational outcomes. Sibling fixed-effects estimates, which control for sibling-shared but not sibling-specific unobserved factors, in Currie et al. (2010) and Salm & Schunk (2012) show that there is no statistically significant effect of asthma on cognitive ability. Lundborg et al. (2011) use twins fixed effects, which control for twins-shared but not twins-specific unobserved factors, on a large sample of MZ Swedish male twins, and find no effect of global health at age 18 on schooling attainment.

This paper differs from previous papers in the literature in three main respects. First, we attend to the issue of reverse causality by leveraging genetic measures of child health. Second, we employ sibling fixed-effects models with controls for PGSs, which allow a "genetic lottery" interpretation of our genetic measures. Third, we compare and contrast models that use genetic vs. standard survey measures of a variety of health outcomes/behaviors. Finally, we note that our results are not directly comparable to previous studies that have used Add Health, as our use of genetic information requires that we focus on European-ancestry individuals, while previous studies include all races.

3. Data

We use Add Health, which is a nationally-representative sample of 20,745 students in grades 7 through 12 (aged 12-21) in 1994-95 (wave 1). Adolescents were surveyed from 132 schools that were selected to ensure representativeness with respect to region, urbanicity, school size and type, and ethnicity. In wave 1, data were collected from adolescents, their parents, siblings, friends, relationship partners, fellow students, and school administrators. The adolescents have been followed after 1 year (wave 2, 1996), 6 years (wave 3, 2001-2002), 13 years (wave 4, 2008), and 20 years (wave 5, 2016-2018). An important aspect of Add Health is that the original design included oversamples of more than 3,000 pairs of individuals with genetic resemblance, including twins, full/half siblings and unrelated siblings in the same household. We make use of the full biological sibling sample to control for shared family, genetic and environmental factors that may confound standard OLS estimates.

At wave 4, 96% of participants consented to providing saliva samples. Approximately 12,200 (80% of those participants) consented to long-term archiving and were consequently eligible for genome-wide genotyping. Genotyping was done on two Illumina platforms, with approximately 80% of the sample genotyping performed with the Illumina Omni1-Quad BeadChip and 20% genotyped with the Illumina Omni2.5-Quad BeadChip. After quality-control procedures, genotyped data are available for 9,974 individuals (7,917 from the Omni1 chip and 2,057 from the Omni2 chip) with 609,130 single nucleotide polymorphisms (SNPs) common across both genotyping platforms. Using these data, Add Health has released PGSs for 9,129 individuals. A PGS is a summary measure of an individual's genetic predisposition for a given trait, and is constructed using results from Genome-Wide Association Studies (GWAS). In a GWAS, hundreds of thousands of SNPs are tested for associations with an outcome. As an example, Locke et al. (2015) conducted a GWAS on a sample of 339,224 individuals and identified 97 SNPs as genome-wide significant predictors ($p<5x10^{-8}$) of BMI, which explain about 2.7% of the variation in BMI. A PGS for individual *i* (equation 1) is a weighted average across the total number of SNPs (m) for a given trait, of the number of reference alleles A (0, 1 or 2) at each SNP (k) multiplied by the corresponding beta estimate from the GWAS analysis. The construction of PGSs

is conceptually simple, but in practice involves several decisions such as whether to use genome-wide significant SNPs or all SNPs. We refer readers to Braudt & Harris (2018) for a detailed description of how the PGSs were constructed in Add Health.

(1) $PGS_i = \sum_{k=1}^{m} \beta_k \ A_{ik}$.

Our main analysis is based on a sample of 5,728 European-ancestry individuals. We concentrate on individuals of European ancestry because most GWAS studies are for this population, and the PGSs for other ethnic groups may not have the same predictive power (Martin et al. 2017). The sibling fixed-effects analysis is based on a sub-sample of 788 full biological siblings (576 full siblings and 212 fraternal twins) of European-ancestry.[3] This sub-sample consists of 373 families with 2 siblings and 14 families with 3 siblings. A detailed description of the variables used is given in appendix A.

4. Empirical Models

Our first approach focuses on eliminating reverse causality concerns in the relationship between health and schooling with the use of health PGSs as proxies for adolescent health. We estimate the effect of PGSs using OLS for equation (2), where the schooling attainment for individual i ($Schooling_i$) is related to the PGSs for BMI, depression, ever smoked, ADHD and educational attainment, a vector of control variables (X_i) and a stochastic error term (u_i). The control variables—listed in appendix A—include the first 20 principal components of the genetic data, which helps control for population stratification.[4]

$$(2)\ Schooling_i = \beta_0 + \beta_1 PGS\_BMI_i + \beta_2 PGS\_DEP_i + \beta_3 PGS\_EverSmoked_i + \beta_4 PGS\_ADHD_i + \beta_5 PGS\_Edu_i + X_i'\gamma + u_i$$

The coefficients on the PGSs may not necessarily reflect pure genetic effects, as they may be confounded by family environment. For example, parental genetics may influence the family environment provided to children, which in turn may affect child outcomes. The PGSs therefore may reflect the influence of both genes and family environment. Equation (2) attempts to account for this to an extent by controlling for self-reported parental health measures to capture parental genetics, and community fixed-effects to control for community factors during adolescence.

In order to more fully control for parental genotypes and unmeasured family factors, we also estimate a sibling fixed-effects model (equation 3) where the schooling of sibling i in family j ($Schooling_{ij}$) is related to the PGSs, control variables(X_{ij}), sibling fixed effects (μ_j), and an error term (ε_{ij}).

$$(3)\ Schooling_{ij} = \alpha_0 + \alpha_1 PGS\_BMI_{ij} + \alpha_2 PGS\_DEP_{ij} + \alpha_3 PGS\_EverSmoked_{ij} + \alpha_4 PGS\_ADHD_{ij} + \alpha_5 PGS\_Edu_i + X_{ij}'\gamma \ + \mu_j + \varepsilon_{ij}$$

The within-sibling variation in the PGS is considered to be quasi-exogenous because differences in genotypes of full biological siblings are the outcomes of a genetic lottery (Fletcher & Lehrer

[3] Of the 5,728 European-ancestry individuals, 1,063 are from the sibling sub-sample. We drop non-related siblings (214 observations), half-siblings (181 observations), twins with undetermined zygosity (24 observations) and identical twins (180 observations). We omit a further 275 observations because the co-sibling is missing.

[4] Population stratification is a situation where the distribution of genes systematically differs by population subgroups such as by ethnicity (von Hinke et al. 2016).

2011). The sibling fixed-effects approach also controls for any parental, neighborhood, or school factors that are shared by siblings.

Our second approach directly estimates the effect of adolescent health. We first use OLS to estimate equation (4), which relates the schooling attainment of individual i to a series of dummy variables for the incidence of adolescent obesity, depression, ever smoking regularly, ADHD and a vector of control variables (X_i). Importantly, we include in X_i the health and educational PGSs to control for unobserved genetic heterogeneity, as well as variables relating to mother's education, mother's health, parental investments (e.g. breastfeeding) to proxy for unobserved family characteristics, and community fixed effects to control for unobserved community-level environmental factors.

$$(4)\ Schooling_i = \beta_0 + \beta_1 Obese_i + \beta_2 Depressed_i + \beta_3 EverSmoke_i + \beta_4 ADHD_i + X_i'\delta + v_i$$

Finally, we follow previous studies and estimate the effect of adolescent health using sibling fixed-effects regressions (equation 5). Although the sibling fixed-effects approach controls for factors that are shared by siblings, it does not control for sibling-specific factors such as innate ability and health endowments. Therefore, contrary to previous papers in the literature, we also include the education and health PGSs as additional covariates to proxy for innate ability and health endowments.

$$(5)\ Schooling_{ij} = \alpha_0 + \alpha_1 Obese_{ij} + \alpha_2 Depressed_{ij} + \alpha_3 EverSmoke_{ij} + \alpha_4 ADHD_{ij} + X_{ij}'\delta + \mu_j + v_{ij}$$

5. Results

5.1 Descriptive Statistics

Descriptive statistics for the main sample are shown in Table 1 columns 1-3. The summary statistics show that 53% are female with an average age of 16 years at wave 1, and 29 years at wave 4. In terms of adolescent health, 10% were obese, 7% were depressed, 26% had ever smoked regularly and 6% were diagnosed with ADHD. On average, individuals have 14.6 grades of schooling and 32% are college graduates. The PGSs are distributed with a mean of 0 and standard deviation of 1.[5] Summary statistics for the sibling sample are given in columns 4-6. Although, the sibling sample is substantially smaller than the main sample, the summary statistics do not reveal any major differences. For example, in the sibling sample 10% were obese, 8% were depressed, 24% had ever smoked regularly and 5% were diagnosed with ADHD. Average grades of schooling is also 14.6 grades as in the main sample. There is a fair amount of within-sibling variation, which is needed to identify the sibling fixed-effects estimates. Figure 1 shows the distribution of the within-siblings differences in the PGSs, which is approximately normally distributed. Table 2 shows that 13% of siblings are discordant on obesity, 16% on depression, 25% on ever smoking regularly, and 9% on ADHD. The absolute value of the mean difference in grades of schooling is 0.87, and 26% are discordant on college graduation.

5.2 Effect of Health PGSs on Schooling Attainment

[5] Appendix Table B1 shows the correlation coefficients among the PGSs. The educational attainment PGS is negatively correlated with the health PGSs, as one would expect. The educational attainment PGS is most strongly correlated with the ADHD PGS (-0.25), and least correlated with the ever-smoked PGS (-0.10). The PGSs for BMI, depression, ever-smoked and ADHD are positively correlated amongst each other. The PGSs for BMI, depression and ever-smoked are most strongly correlated with the ADHD PGS.

Table 3 provides estimates of the effect of health PGSs on schooling attainment. The first five columns of Table 3 consider the effects on grade of schooling, and the last three consider the effects on college graduation. Among the columns considering the effects on grades of schooling (respectively, college graduation), the first three (one) employ the main sample and the last two employ the sibling sample. In addition, some of the estimated models in Table 3 use a basic set of control variables, while others use a full set of controls. The specific variables included in each of these two sets of controls are listed in the notes to Table 3.

The estimates in column 1 of Table 3 show that all the health PGSs have statistically significant negative effects on grades of schooling when conditioning on the basic controls. The ADHD PGS has the largest effect, with a one-unit (i.e., a one-standard-deviation) increase in the ADHD PGS decreasing grades of schooling by 0.23. The ever-smoked PGS has the smallest effect, with an estimate of -0.10. The estimated effects of the health PGSs decrease in magnitude as we control for family and community factors (the full controls), and the educational PGS in column 3. In this case, the BMI and depression PGSs have smaller estimated effects of -0.03 and -0.05, respectively, and are statistically insignificant. The ADHD and ever-smoked PGSs, however, still have statistically significant negative effects. A one-unit increase in the ADHD (respectively, ever-smoked) PGS decreases grades of schooling by 0.08 (0.06) of a grade. In addition, the results in column 3 indicate that a one-unit increase in the educational attainment PGS increases grades of schooling by 0.34. The sibling fixed-effects estimates in column 5 again show the importance of innate genetic ability. A one-unit increase in the educational PGS increases grades of schooling by 0.38, which is similar to the OLS estimate from the full specification in column 3. The sibling fixed-effects estimates for all the health PGSs are statistically insignificant but the magnitudes are not small. They suggest negative effects of having a high genetic risk of ADHD and ever smoking, if we consider that the smaller sibling sample leads to low precision. For example, in column 5 a one-unit increase in the ADHD PGS decreases grades of schooling by 0.06. This estimate is similar in magnitude to the OLS estimate of -0.08 from the full specification in column 3, but its standard error is about five times larger. The sibling fixed-effects estimate for the ever-smoked PGS is -0.12, even larger than the OLS estimate in column 3. This is a considerable effect, given that the estimated effect of the educational PGS is 0.38.

The sibling fixed-effects estimates for college graduation in column 8 provide robust evidence that being genetically predisposed to smoking lowers the probability of college graduation, but there is no strong evidence for a negative effect of having a high genetic risk of ADHD. In particular, a one-unit increase in the ever-smoked PGS reduces the probability of college graduation by 8 percentage points, and this estimate is statistically significant at the 1% level. Moreover, the estimated effect of the ever-smoked PGS is about the same size as the educational PGS, which has an estimated coefficient of 0.073. In comparison, the sibling fixed-effects estimate for the ADHD PGS is close to zero (0.0027) and substantially smaller than the corresponding OLS estimate of 0.0405 from the sibling sample in column 7. This suggests that the OLS estimate for the ADHD PGS is confounded by parental genetics and family environment. We also estimated the effects of the health PGSs on the probability of (1) being a high school dropout, (2) being a high school graduate, and (3) having some college education. We did not find consistent evidence that the health PGSs predict these outcomes, except for the ever-smoke PGS when considering the outcome some college education (see appendix table B2).

The results from Table 3 suggest that having a high genetic risk of ADHD or ever smoking may have negative impacts on schooling attainment. It is possible that the effects of the ADHD and ever-smoked PGSs operate through risky behaviors. Adolescents with a high genetic risk of ADHD or ever smoking may engage in risky behaviors, which may lead to lower schooling attainment. To examine this possibility, we re-estimated the regressions controlling for a PGS for risk tolerance. The risk tolerance PGS is based on a GWAS by Linnér et al. (2019), who analyzed the genetic architecture of risk tolerance, adventurousness, and risky behaviors in driving,

drinking, smoking and sexual domains. Results are shown in Table 4. As this PGS is available for 4,755 unrelated European-ancestry individuals, columns 1 and 3 report estimates without controlling for the risk tolerance PGS. The estimates show that a one-unit increase in the ever-smoked (respectively, ADHD) PGS decreases grades of schooling by 0.06 (0.10) of a grade and decreases the probability of college graduation by 2.11 (2.98) percentage points. These estimates are similar to the corresponding estimates in Table 3 columns 3 and 6. The magnitude of these estimates is virtually unchanged when we control for the risk tolerance PGS in columns 2 and 4. This suggests that the effects of the ever-smoked and ADHD PGSs are not driven by risk tolerance. It is interesting to note that the risk tolerance PGS does not predict grades of schooling—the estimated coefficient is small (-0.0335) and statistically insignificant.[6] It does, however, have a statistically significant negative effect on college graduation. A one-unit increase in the risk tolerance PGS reduces the probability of college graduation by 1.28 percentage points, which is about a fifth of the effect of the educational PGS (6.26 percentage points).

Finally, we have also estimated the effect of the health PGSs on adolescent human capital outcomes from wave 1. The outcomes we use are: (1) percentile rank on the Add Health Peabody Picture Vocabulary Test (AHPPVT), (2) grade point average based on grades in math, English, science and history, and (3) dummy variable equal to 1 if the individual repeated either of grades 6 through 12. The results are given in appendix table B3. The OLS and sibling fixed-effects estimates do not show any consistent evidence of negative effects of the health PGSs on these outcomes. This suggests that the effect of the ever-smoked and ADHD PGSs on adult schooling attainment do not strongly operate through adolescent educational achievement.

5.3 Effect of Adolescent Health on Schooling Attainment

The analysis up till now provides evidence on the effect of adolescent health on school attainment that is free from concerns of reverse causality by using the health PGSs as proxies for actual adolescent health. To provide additional evidence, we present results of the effect of actual adolescent health on schooling attainment in Table 5. The OLS estimates for the main sample in columns 1 and 5, which condition on the full set of controls, show statistically significant negative effects of all adolescent health measures, with the largest effect being for ever-smoking regularly. Individuals who ever smoked regularly in adolescence have 0.88 of a grade less of schooling and are 20 percentage points less likely to graduate college compared to individuals who did not. This finding is robust to controlling for unobserved family and genetic factors. The sibling fixed-effects estimate for grades of schooling in column 3 indicates that individuals who ever smoke regularly in adolescence have 0.72 of a grade less of schooling. This difference drops to 0.63 of a grade when adding the PGSs as controls in column 4. For college graduation, the sibling fixed-effects estimate in column 7 indicates that individuals who ever smoked regularly in adolescence have a 22 percentage-point lower probability of graduating from college, which drops slightly to 20 percentage points when controlling for the PGSs in column 8. The results also provide suggestive evidence that there may not be a causal effect of adolescent obesity and depression on schooling attainment. While OLS estimates for both the main and sibling samples show large differences in schooling attainment by obesity and depression status, the sibling fixed-effects estimates are substantially smaller and statistically insignificant. For example, the OLS estimates in columns 2 and 6 for the sibling sample show that individuals who were depressed in adolescence have on average 1 grade less of schooling and are 17 percentage points less likely to be college graduates than those who were not depressed. The sibling fixed-effects estimates in columns 4 and 8, however, indicate much smaller differences and are statistically insignificant. Individuals who were depressed in adolescence have on average 0.12 grade less of schooling and are only 0.6 percentage points less likely to be college graduates than those who were not depressed.

[6] When we regress grades of schooling on the risk tolerance PGS, full controls and community fixed effects, the estimated coefficient (standard error) on the risk tolerance PGS is -0.0463 (0.0287).

The sibling fixed-effect models with PGSs as additional controls are close in spirit to MZ twins fixed-effect models. Table 6 gives OLS and twins fixed-effects estimates from a sample of 378 MZ twins taken from the full Add Health sibling dataset. The results from using MZ twins are similar to those from the sibling fixed-effects regressions with PGSs as additional controls in Table 5. Although the MZ twins fixed-effects estimates have large standard errors, likely due to the small sample size, the magnitudes of the estimate for ever smoked regularly is considerable. For example, the twins fixed-effects estimates in column 2 indicate that twins who ever smoked regularly in adolescence have 0.75 of a grade less of schooling, which is similar to the sibling fixed-effect estimate of 0.63 with PGSs as additional controls (column 4 Table 5). In column 4, the MZ twins fixed-effects estimates show that twins who ever smoked regularly in adolescence are 11 percentage points less likely to be college graduates. This is about half the size of the sibling fixed-effects estimate of 20 percentage points in column 8 of Table 5, but the OLS estimate for MZ twins (-0.1090) is different than the OLS estimate for the sibling sample (-0.2813).

Our findings may not be generalizable to the entire Add Health cohort, as they are based on a sample of European-ancestry individuals who agreed to have DNA samples taken. For reference, Table 7 presents OLS and sibling fixed-effects estimates using the entire Add Health sibling dataset for whites and non-whites. Similar to our previous results in Table 5, the sibling fixed-effects estimates for whites and non-whites in Table 7 both show that ever smoking regularly in adolescence has a statistically significant negative effect on schooling attainment, whereas the effects of adolescent obesity, depression, and ADHD are not statistically significant. The difference in grades of schooling by adolescent smoking status is larger for non-whites. For non-whites, the average difference in grades of schooling between individuals who ever smoked regularly in adolescence and those that did not is 0.90 of a grade (column 4). The corresponding difference for whites is 0.66 of a grade (column 2). However, there is only a 11 percentage point difference in the probability of college graduation by adolescent smoking status for non-whites (column 8), whereas there is a 21 percentage point difference for whites (column 6). This suggests that our results are likely to be generalizable for whites in Add Health, but probably not for other racial groups.

6. Summary

It is extremely difficult to establish credible research designs to estimate the causal effect of childhood/adolescent health on schooling attainment. This difficulty is primarily due to reverse causality and confounding from genetic, family, and environmental factors. Given these challenges, previous studies have by in large used sibling fixed-effects models to assess whether standard associations are robust to controlling for unobserved genetic, environmental, and family factors shared by siblings. This study makes use of genetic data from Add Health in two innovative ways to increase our understanding of whether there is a causal effect of adolescent health on schooling attainment. First, we estimate the effect of health PGSs on schooling attainment. While these estimates do not tell us the effect of actual adolescent health, they have the advantage of being free from reverse causality and can be informative as to whether a causal effect exists. Results from OLS regressions indicate that having a high genetic risk for smoking is detrimental to schooling attainment. This finding is robust in sibling fixed-effects models that some control for unobserved shared family and environmental factors. Second, like previous studies, we estimate the effect of adolescent health using sibling fixed-effects models, with the difference that our analyses additionally use the PGSs to control for possible residual confounding from sibling-specific genetic differences. This is close in spirit to using a MZ twins fixed-effects approach.

Our main findings can be summarized as follows. First, OLS and sibling fixed-effects estimates show that being genetically predisposed to ever smoking has a negative effect on schooling attainment, with the latter estimates suggesting that a one-standard-deviation increase in the ever-smoke PGS results in 0.12 of a grade less of schooling and a lower probability of graduating from college of 8 percentage points. Interestingly, our analyses suggest that this result:

(i) is not being driven by being genetically predisposed to risk tolerance, adventurousness, and risky behaviors in driving, drinking, smoking and sexual domains; and (ii) does not strongly operate through adolescent educational achievement. Second, the negative effects of ever smoking regularly in adolescence result in 0.63 of a grade less of schooling and a lower probability of graduating from college of 20 percentage points, and are robust to controlling for unmeasured family and genetic factors. Third, our estimates show that having a high genetic risk of ADHD reduces schooling attainment, but the effect of actual ADHD during adolescence is small and statistically insignificant. Finally, across both our approaches, we find very little evidence of a detrimental effect of obesity and depression on schooling attainment, consistent with findings in the literature.

References

Braudt, D. B., & Harris, K. M. (2018). Polygenic Scores (PGSs) in the National Longitudinal Study of Adolescent to Adult Health (Add Health)–Release.

Currie, J., & Stabile, M. (2006). Child mental health and human capital accumulation: the case of ADHD. *Journal of Health Economics*, *25*(6), 1094-1118.

Currie, J., Stabile, M., Manivong, P., & Roos, L. L. (2010). Child health and young adult outcomes. *Journal of Human Resources*, *45*(3), 517-548.

Fletcher, J. M. (2010). Adolescent depression and educational attainment: results using sibling fixed effects. *Health Economics*, *19*(7), 855-871.

Fletcher, J. M. (2018). Economics and Genomics. In *Oxford Research Encyclopedia of Economics and Finance*.

Fletcher, J. M., & Lehrer, S. F. (2009). The effects of adolescent health on educational outcomes: Causal evidence using genetic lotteries between siblings. In *Forum for Health Economics & Policy* (Vol. 12, No. 2). De Gruyter.

Fletcher, J. M., & Lehrer, S. F. (2011). Genetic lotteries within families. *Journal of Health Economics*, *30*(4), 647-659.

Fletcher, J., & Wolfe, B. (2008). Child mental health and human capital accumulation: the case of ADHD revisited. *Journal of Health Economics*, *27*(3), 794-800.

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*(2), 245-257.

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.

Lundborg, P., Nilsson, A., & Rooth, D. O. (2011). Does early life health predict schooling within twin pairs?. IZA Discussion Paper No. 5803

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., ... & Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, *100*(4), 635-649.

Prinz, D., Chernew, M., Cutler, D., & Frakt, A. (2018). *Health and economic activity over the lifecycle: Literature review*. National Bureau of Economic Research Working Paper No w24865.

Sabia, J. J., & Rees, D. I. (2015). Body weight, mental health capital, and academic achievement. *Review of Economics of the Household*, *13*(3), 653-684.

Salm, M., & Schunk, D. (2012). The relationship between child health, developmental gaps, and parental education: Evidence from administrative data. *Journal of the European Economic Association*, *10*(6), 1425-1449.

Smith, J. P., & Smith, G. C. (2010). Long-term economic costs of psychological problems during childhood. *Social Science & Medicine*, *71*(1), 110-115.

von Hinke, S., Smith, G. D., Lawlor, D. A., Propper, C., & Windmeijer, F. (2016). Genetic markers as instrumental variables. *Journal of Health Economics*, *45*, 131-148.

van Kippersluis, H., & Rietveld, C. (2018). Pleiotropy-robust Mendelian Randomization. *International Journal of Epidemiology*, *47*, 1279-1288.

Table 1: Descriptive Statistics

	Main Sample			Sibling Sample		
	Obs (1)	Mean (SD) (2)	Min (Max) (3)	Obs (4)	Mean (SD) (5)	Min (Max) (6)
Basic Demographics						
Female	5728	0.53 (0.50)	0 (1)	788	0.52 (0.50)	0 (1)
Age-wave 1	5726	16.01 (1.74)	12 (21)	788	16.04 (1.67)	13 (19)
Age-wave 4	5726	29.01 (1.75)	25 (34)	788	29.03 (1.68)	25 (33)
Birth order	5718	1.77 (1.03)	1 (12)	787	2.12 (1.15)	1 (9)
Adolescent health						
Obese	5593	0.10 (0.29)	0 (1)	768	0.10 (0.29)	0 (1)
Depressed	5713	0.07 (0.25)	0 (1)	787	0.08 (0.27)	0 (1)
Ever smoked regularly	5726	0.26 (0.44)	0 (1)	788	0.24 (0.42)	0 (1)
ADHD	5728	0.06 (0.23)	0 (1)	788	0.05 (0.22)	0 (1)
Family background						
Mother's grades of schooling	5728	13.22 (2.01)	8 (17)	788	13.31 (1.94)	8 (17)
Mother's schooling missing	5728	0.15 (0.36)	0 (1)	788	0.14 (0.34)	0 (1)
Mother obese	5728	0.20 (0.37)	0 (1)	788	0.27 (0.42)	0 (1)
Mother obese missing	5728	0.16 (0.37)	0 (1)	788	0.11 (0.32)	0 (1)
Mother excellent/v.good health	5728	0.60 (0.45)	0 (1)	788	0.62 (0.45)	0 (1)
Mother excellent/v.good health missing	5728	0.15 (0.35)	0 (1)	788	0.13 (0.33)	0 (1)
Mother smoke	5728	0.32 (0.43)	0 (1)	788	0.32 (0.43)	0 (1)
Mother smoke missing	5728	0.15 (0.36)	0 (1)	788	0.13 (0.34)	0 (1)
Child health insurance	5728	0.06 (0.23)	0 (1)	788	0.07 (0.24)	0 (1)
Child health insurance missing	5728	0.10 (0.30)	0 (1)	788	0.07 (0.26)	0 (1)
Months breastfed	5728	3.27 (5.06)	0 (24)	788	3.53 (5.35)	0 (24)
Months breastfed missing	5728	0.12 (0.32)	0 (1)	788	0.08 (0.27)	0 (1)
Parents married	5728	0.79 (0.38)	0 (1)	788	0.81 (0.38)	0 (1)
Parents married missing	5728	0.10 (0.30)	0 (1)	788	0.09 (0.28)	0 (1)
Genetics						
BMI PGS	5728	0.00 (1.00)	-3.67 (3.82)	788	0.05 (1.01)	-3.37 (2.65)
Depression PGS	5728	0.00 (1.00)	-3.25 (6.39)	788	-0.06 (0.96)	-3.25 (3.91)
Ever-Smoke PGS	5728	0.00 (1.00)	-4.64 (3.49)	788	-0.07 (0.98)	-3.71 (3.49)
ADHD PGS	5728	0.00 (1.00)	-3.82 (3.56)	788	-0.01 (1.06)	-3.81 (3.47)
Educational PGS	5728	0.00 (1.00)	-4.13 (3.39)	788	0.04 (0.99)	-2.34 (3.39)

Adolescent Achievement- wave 1						
Verbal Ability Percentile Rank	5478	58.67 (25.80)	0 (100)	757	57.29 (24.95)	0 (100)
Grade Repetition	5728	0.05 (0.23)	0 (1)	788	0.06 (0.23)	0 (1)
GPA	5594	2.64 (0.89)	0 (4)	767	2.64 (0.88)	0 (4)
Adult Schooling Attainment- wave 4						
Grades of Schooling	5728	14.56 (2.19)	8 (20)	788	14.56 (2.25)	8 (20)
High school dropout	5728	0.08 (0.26)	0 (1)	788	0.09 (0.28)	0 (1)
High school graduate	5728	0.16 (0.37)	0 (1)	788	0.15 (0.35)	0 (1)
Some college education	5728	0.44 (0.50)	0 (1)	788	0.43 (0.49)	0 (1)
College graduate	5728	0.32 (0.47)	0 (1)	788	0.34 (0.47)	0 (1)

Figure 1: Within-Sibling Variation in PGSs

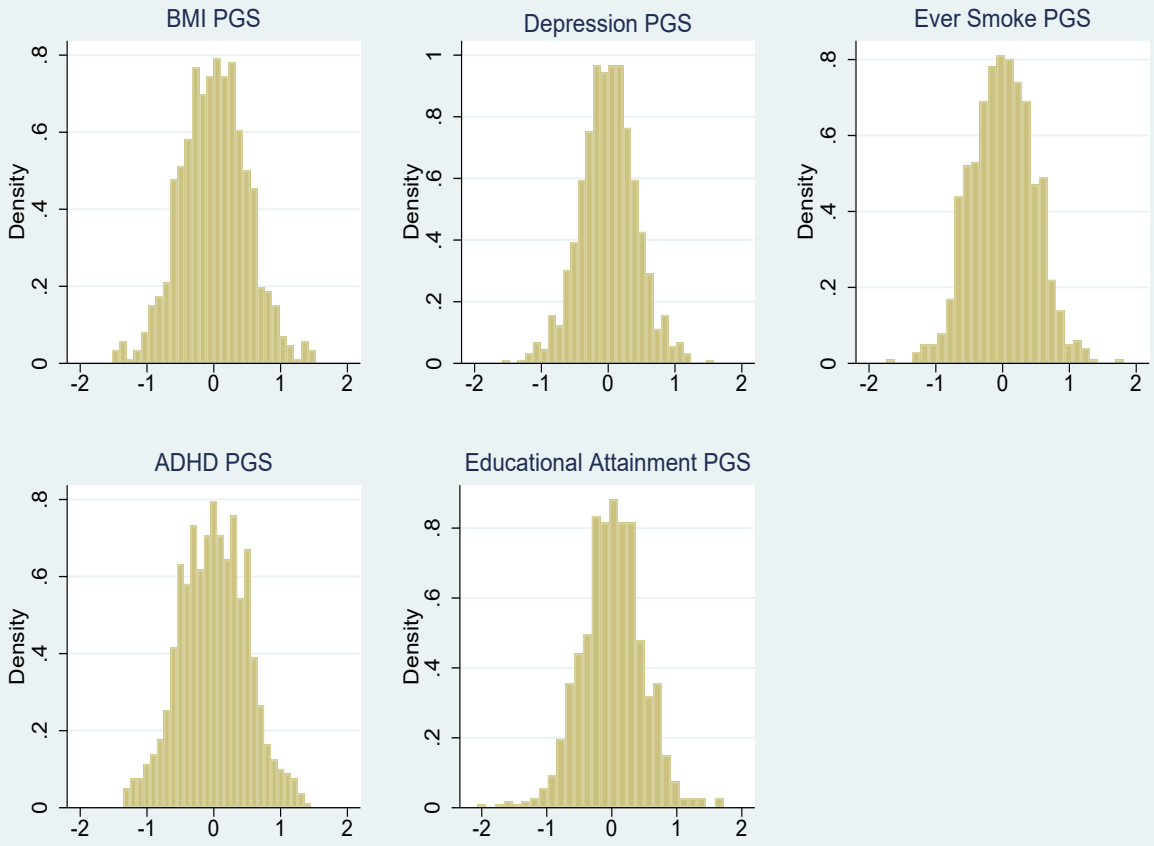

Table 2: Within-Sibling Variation in Adolescent Health, Adolescent Cognition and Adult Schooling Attainment

Adolescent Health	
% of families discordant on	
Obesity	12.50% [370]
Depression	15.65% [386]
Ever Smoked Regularly	24.87% [387]
ADHD	9.26% [387]
Schooling Attainment	
Mean (standard deviation) of absolute within-family difference in grades of schooling	0.87 (0.86) [387]
% of families discordant on college graduation	26% [387]

Notes: Number of families in [.].

Table 3: Effect of Health PGSs on Schooling Attainment

Outcome	Grades of Schooling (1)	Grades of Schooling (2)	Grades of Schooling (3)	Grades of Schooling (4)	Grades of Schooling (5)	College Graduate (6)	College Graduate (7)	College Graduate (8)
Sample	**Main**	**Main**	**Main**	**Sibling**	**Sibling**	**Main**	**Sibling**	**Sibling**
Controls	**Basic**	**Full**	**Full**	**Basic**	**Basic**	**Full**	**Basic**	**Basic**
Fixed Effects	**None**	**Community**	**Community**	**None**	**Sibling**	**Community**	**None**	**Sibling**
BMI PGS	-0.1253***	-0.706***	-0.0309	-0.0546	-0.0379	-0.0113*	-0.0308*	0.0095
	(.0309)	(.0243)	(.0247)	(.0827)	(.1146)	(.0061)	(.0176)	(.0265)
Depression PGS	-0.1582***	-0.0883**	-0.0491	-0.0319	-0.0814	-0.0195**	-0.0207	-0.0170
	(.0391)	(.0353)	(.0356)	(.1054)	(.1515)	(.0077)	(.0238)	(.0289)
Ever-Smoke PGS	-0.0999***	-0.0849***	-0.0641**	-0.1668*	-0.1212	-0.0211***	-0.0361*	-0.0798***
	(.0291)	(.0255)	(.0249)	(.0922)	(.1385)	(.0057)	(.0192)	(.0247)
ADHD PGS	-0.2330***	-0.1402***	-0.0830***	-0.1521*	-0.0647	-0.0289***	-0.0405**	0.0027
	(.0297)	(.0278)	(.0277)	(.0898)	(.1339)	(.0062)	(.0189)	(.0275)
Educational PGS			0.3366***	0.6077***	0.3734***	0.0610***	0.1214***	0.0737**
			(.0251)	(.0904)	(.1356)	(.0065)	(.0184)	(.0292)
F-Statistic	36.06	15.08	5.70	2.41	0.45	9.26	4.11	2.91
P-Value	0.0000	0.0000	0.0004	0.0486	0.7740	0.0000	0.0028	0.0214
N	5,716	5,716	5,716	786	786	5,716	786	786
# of Families				386	386		366	366

Notes: Basic controls consist of age, gender, birth order and the first 20 principal components of the genetic data. Full controls consist of the basic controls and controls for the number of months the respondent was breastfed, mother's grades of schooling, dummy variables equal to 1 if (i) the mother reports being obese, (ii) the mother reports being a smoker, (iii) the mother reports being in excellent or very good health, (iv) the parents are married, (v) if the child had no health insurance. Missing values are imputed with the sample mean and controlled for with dummy variables for missing data. Standard errors for the main sample are clustered at the community level. Standard errors for the sibling sample are clustered at the family level. The F-statistic and p-value are for the joint test of significance for the health PGSs. ***significant at 1% **significant at 5% *significant at 10%.

Table 4: Effect of Health PGSs on Schooling Attainment Controlling for a Risk Tolerance PGS

Outcome	**Grades of Schooling (1)**	**Grades of Schooling (2)**	**College Graduate (3)**	**College Graduate (4)**
Sample	**Main**	**Main**	**Main**	**Main**
Controls	**Full**	**Full**	**Full**	**Full**
Fixed Effects	**Community**	**Community**	**Community**	**Community**
BMI PGS	-0.0257	-0.0209	-0.0109	-0.0109
	(.0268)	(.0281)	(.0074)	(.0074)
Depression PGS	-0.0560	-0.0549	-0.0197**	-0.0197**
	(.0382)	(.0379)	(.0086)	(.0086)
Ever-Smoke PGS	-0.0565*	-0.0571*	-0.0211***	-0.0214***
	(.0303)	(.0303)	(.0070)	(.0070)
ADHD PGS	-0.1000***	-0.0965***	-0.0298***	-0.0284***
	(.0362)	(.0360)	(.0064)	(.0063)
Educational PGS	0.3394***	0.3414***	0.0619***	0.0626***
	(.0285)	(.0285)	(.0071)	(.0071)
Risk Tolerance PGS		-0.0335		-0.0128**
		(.0288)		(.0061)
F-Statistic	4.18	3.72	9.67	8.71
P-Value	0.0038	0.0072	0.0000	0.0000
N	4,664	4,664	4,664	4,664

Notes: Basic controls consist of age, gender, birth order and the first 20 principal components of the genetic data. Full controls consist of the basic controls and controls for the number of months the respondent was breastfed, mother's grades of schooling, dummy variables equal to 1 if (i) the mother reports being obese, (ii) the mother reports being a smoker, (iii) the mother reports being in excellent or very good health, (iv) the parents are married, (v) if the child had no health insurance. Missing values are imputed with the sample mean and controlled for with dummy variables for missing data. Standard errors for the main sample are clustered at the community level. The F-statistic and p-value are for the joint test of significance for the health PGSs. ***significant at 1% **significant at 5% *significant at 10%.

Table 5: The Effect of Adolescent Health on Schooling Attainment

Outcome	Grades of Schooling (1)	Grades of Schooling (2)	Grades of Schooling (3)	Grades of Schooling (4)	College Graduate (5)	College Graduate (6)	College Graduate (7)	College Graduate (8)
Sample	**Main**	**Sibling**	**Sibling**	**Sibling**	**Main**	**Sibling**	**Sibling**	**Sibling**
Controls	**Full**	**Basic**	**Basic**	**Basic**	**Full**	**Basic**	**Basic**	**Basic**
Fixed Effects	**Community**	**None**	**Sibling**	**Sibling**	**Community**	**None**	**Sibling**	**Sibling**
Obese	-0.1988**	-0.7235**	0.2607	0.3230	-0.0419**	-0.1614***	0.0498	-0.0441
	(.0932)	(.3101)	(.3566)	(.3544)	(.0186)	(.0560)	(.0767)	(.0762)
Depressed	-0.4215***	-0.9895***	-0.1401	-0.1157	-0.0837***	-0.1710***	-0.0060	-0.0064
	(.1118)	(.3065)	(.3583)	(.3578)	(.0170)	(.0554)	(.0567)	(.0572)
Ever Smoked	-0.8847***	-1.2176***	-0.7191***	-0.6295**	-0.1978***	-0.2813***	-0.2189***	-0.2005***
Regularly	(.0603)	(.1928)	(.2597)	(.2602)	(.0140)	(.0358)	(.0543)	(.0545)
ADHD	-0.4625***	0.1215	0.0672	-0.0041	-0.0880***	0.0830	0.0233	0.0127
	(.1108)	(.4305)	(.5011)	(.5037)	(.0212)	(.0795)	(.1094)	(.1065)
BMI PGS	-0.0138			0.0084	-0.0073			0.0255
	(.0235)			(.1200)	(.0059)			(.0270)
Depression PGS	-0.0456			-0.0879	-0.0168**			-0.0207
	(.0342)			(.1540)	(.0075)			(.0281)
Ever-Smoke PGS	-0.0413			-0.0550	-0.0163***			-0.0630**
	(.0255)			(.1400)	(.0057)			(.0250)
ADHD PGS	-0.0673**			-0.0700	-0.0259***			-0.0022
	(.0260)			(.1375)	(.0060)			(.0276)
Educational PGS	0.3067***			0.3801**	0.0550***			0.0766***
	(.0245)			(.1367)	(.0065)			(.0279)
N	5,568	748	748	748	5,568	748	748	748
# of families		368	368	368		368	368	368

Notes: Basic controls consist of age, gender, birth order and the first 20 principal components of the genetic data. Full controls consist of the basic controls and controls for the number of months the respondent was breastfed, mother's grades of schooling, dummy variables equal to 1 if (i) the mother reports being obese, (ii) the mother reports being a smoker, (iii) the mother reports being in excellent or very good health, (iv) the parents are married, (v) if the child had no health insurance. Missing values are imputed with the sample mean and controlled for with dummy variables for missing data. Standard errors for the main sample are clustered at the community level. Standard errors for the sibling sample are clustered at the family level. The F-statistic and p-value are for the joint test of significance for the health PGSs. ***significant at 1% **significant at 5% *significant at 10%.

Table 6: Effect of Adolescent Health on Schooling Attainment, MZ Twin Sample

Outcome	**Grades of Schooling (1)**	**Grades of Schooling (2)**	**College Graduate (3)**	**College Graduate (4)**
Sample	**MZ Twins**	**MZ Twins**	**MZ Twins**	**MZ Twins**
Controls	**Basic**	**Basic**	**Basic**	**Basic**
Fixed Effects	**None**	**Twin**	**None**	**Twin**
Obese	-0.4623	0.1413	-0.2647***	-0.1991
	(.2935)	(.3178)	(.0590)	(.1226)
Depressed	-1.0319***	-0.4819	-0.2277***	-0.0634
	(.3753)	(.3097)	(.0734)	(.1015)
Ever-Smoked	-0.6450*	-0.7510	-0.1090	-0.1076
Regularly	(.3393)	(.4865)	(.0669)	(.6753)
ADHD	-0.4055	0.3755	-0.0303	0.0538
	(.5963)	(.5211)	(.1474)	(.0462)
N	378	378	378	378
# of twin pairs	189	189	189	189

Notes: Basic controls consist of age and gender. Standard errors for the main sample are clustered at the family level. ***significant at 1% **significant at 5% *significant at 10%.

Table 7: The Effect of Adolescent Health on Schooling Attainment for Whites and Non-Whites

Outcome	**Grades of Schooling (1)**	**Grades of Schooling (2)**	**Grades of Schooling (3)**	**Grades of Schooling (4)**	**College Graduate (5)**	**College Graduate (6)**	**College Graduate (7)**	**College Graduate (8)**
Sample	**White Siblings**	**White Siblings**	**Non-White Siblings**	**Non-White Sibling**	**White Sibling**	**White Sibling**	**Non-White Sibling**	**Non-White Sibling**
Controls	**Basic**	**Basic**	**Basic**	**Basic**	**Basic**	**Basic**	**Basic**	**Basic**
Fixed Effects	**None**	**Sibling**	**None**	**Sibling**	**None**	**Sibling**	**None**	**Sibling**
Obese	-0.6064***	0.0530	-0.3407*	-0.0027	-0.1170***	0.0207	-0.0940***	-0.0180
	(.1697)	(.2223)	(.1808)	(.2087)	(.0337)	(.0445)	(.0292)	(.0350)
Depressed	-1.1431***	-0.2844	-0.9044***	-0.2987	-0.1549***	-0.0128	-0.1449***	-0.0521
	(.2077)	(.2165)	(.1960)	(.2244)	(.0327)	(.0372)	(.0277)	(.0366)
Ever Smoked	-1.2882***	-0.6630***	-1.1529***	-0.9005***	-0.2846***	-0.2068***	-0.1589***	-0.1052***
Regularly	(.1141)	(.1488)	(.2019)	(.2299)	(.0219)	(.0305)	(.0276)	(.0343)
ADHD	-0.1538	-0.2655	-1.3598**	-0.4838	-0.0264	-0.0419	-0.0361	0.1106
	(.2437)	(.2784)	(.6347)	(.6627)	(.0462)	(.0565)	(.0801)	(.0915)
N	2,030	2,030	1,505	1,505	2,030	2,030	1,505	1,505
# of Families	986	986	715	715	986	986	715	715

Notes: Basic controls consist of age, gender, and birth order. Standard errors are clustered at the family level. ***significant at 1% **significant at 5% *significant at 10%.

Appendix A: Variable Descriptions

This appendix provides information on the construction of the key variables used in the analysis.

Schooling Attainment Measures:

Grades of Schooling: This is based on responses to the question "what is the highest level of education that you have achieved to date?" at wave 4. Response options and their assigned grades of schooling (in parentheses) were: eighth grade or less (8), some high school (10), high school graduate (12), some vocational/technical training (13), completed vocational/technical training (14), some college (14), completed college (16), some graduate school (17), completed a master's degree (18), some graduate training beyond a master's degree (19), completed a doctoral degree (20), some post-baccalaureate professional education (18), and completed post-baccalaureate professional education (19).

College Graduate: This is a binary indicator equal to 1 if the individual reports completing college or higher when asked about the highest level of education achieved at wave 4.

Adolescent Health Measures

Obesity: We first calculate BMI using self-reported height and weight at wave 1. Adolescents are classified as obese if their BMI is greater or equal to the 95th percentile of the BMI distribution adjusted for age and sex. The 95th percentile cutoffs are only available for individuals up to age 20 and are obtained from https://www.cdc.gov/growthcharts/html_charts/bmiagerev.htm

Depression: At wave 1, 19 of the 20 items of the Center for Epidemiologic Studies Depression Scale (CES-D) are available, which is used to measure depressive symptoms. The CES-D scale is widely used to measure symptoms of depression, and has been tested in multiple settings for validity and reliability. The CES-D is created by summing responses (ranging from 0 to 3) to questions such as "how often in the last week were you bothered by things not normally bothersome". Depression is defined as a CES-D score greater than or equal to 22 for male adolescents, and 24 for females.

Ever Smoked Regularly: This is an indicator variable equal to 1 if the adolescent at wave 1 replies yes to the question "have you ever smoked cigarettes regularly, that is, at least 1 cigarette every day for 30 days?" Adolescents who reply no and adolescents who have never smoked are coded as 0.

ADHD: ADHD is measured retrospectively from wave 4. At wave 4, respondents were asked "has a doctor, nurse, or other health care professional ever told you that you have or had: attention problems or ADD or ADHD". Respondents that reply yes were then asked "how old were you when the doctor nurse, or other health care professional diagnosed you with attention problems or ADD or ADHD?" Using responses to these questions, we create an indicator variable for ADHD in adolescence if respondents report that they were diagnosed with attention problems or ADD or ADHD at age 20 or younger.

It is also possible to measure ADHD retrospectively in early childhood from wave 3 as done in Fletcher & Wolfe (2008) and Fletcher & Lehrer (2009, 2011). At wave 3, respondents were asked to think back when they were between 5 and 12 and report how often they performed a set of behaviors (e.g., squirmed in their car seat, had difficulty sustaining attention in tasks). In total, there are 17 questions that are keyed to the Diagnostic and Statistical Manual of Mental Disorders. We do not measure ADHD based on wave 3 responses, because there only 4,846

European-ancestry individuals who have no missing data on the ADHD scale in wave 3, whereas all 5,728 respondents answered the ADHD questions at wave 4.

Adolescent Achievement Measures:

Adolescent cognition: This is measured using the individual's percentile rank (ranging from 0-100) on the Add Health Peabody Picture Vocabulary Test (AHPPVT) at wave 1. The AHPPVT is a computerized, abridged version of the Peabody Picture Vocabulary Test–Revised (PPVT-R). The AHPPVT is a test of hearing vocabulary, designed for persons 2.5 to 40 years old who can see and hear reasonably well and who understand standard English to some degree. The test scores are standardized by age.

GPA: At wave 1, individuals reported grades in English, math, science, and history. We assigned 0 to "took subject/wasn't graded"; 1 to "D or lower"; 2 to "C", 3 to "B"; and 4 to "A". The GPA was calculated as the average grade across these four subjects.

Grade 6-12 repetition: This is a dummy variable equal to 1 if the individual reports that they repeated a grade from 6-12 and 0 if otherwise.

Control Variables:

Basic controls: These consist of age, a dummy variable for being female, birth order and the first 20 principal components of the genetic data.

Full controls: We use responses from the parental questionnaire to control for the number of months the respondent was breastfed, mother's grades of schooling, and a series of dummy variables equal to 1 if (i) the mother reports being obese, (ii) the mother reports being a smoker, (iii) the mother reports being in excellent or very good health, (iv) the parents are married, (v) the child had no health insurance. In order to increase the sample size, missing values are imputed with the sample mean and controlled for with dummy variables for missing data.

Appendix B: Additional Tables

Appendix Table B1: Correlation between the PGSs

	Education	BMI	Depression	Ever-Smoked	ADHD
Education	1.000	-0.1712***	-0.1519***	-0.1043***	-0.2461***
		(.0132)	(.0130)	(.0135)	(.0129)
BMI	-0.1712***	1.000	0.0347***	0.1143***	0.2164***
	(.0132)		(.0131)	(.0130)	(.0129)
Depression	-0.1519***	0.0347***	1.000	0.0992***	0.1993***
	(.0130)	(.0131)		(.0134)	(.0126)
Ever-Smoked	-0.1043***	0.1143***	0.0992***	1.000	0.1412***
	(.0135)	(.0130)	(.0134)		(.0129)
ADHD	-0.2461***	0.2164***	0.1993***	0.1412***	1.000
	(.0129)	(.0129)	(.0126)	(.0129)	

Notes: N=5,728. Robust standard errors in parentheses. ***significant at 1% **significant at 5% *significant at 10%

Appendix Table B2: The Effect of Health PGSs on Different Levels of Schooling Attainment

Outcome	High School Dropout (1)	High School Dropout (2)	High School Dropout (3)	High School Graduate (4)	High School Graduate (5)	High School Graduate (6)	Some College Education (7)	Some College Education (8)	Some College Education (9)
Sample	**Main**	**Sibling**	**Sibling**	**Main**	**Sibling**	**Sibling**	**Main**	**Sibling**	**Sibling**
Controls	**Full**	**Basic**	**Basic**	**Full**	**Basic**	**Basic**	**Full**	**Basic**	**Basic**
Fixed Effects	**Community**	**None**	**Sibling**	**Community**	**None**	**Sibling**	**Community**	**None**	**Sibling**
BMI PGS	0.011	0.0014	0.0175	-0.0005	-0.0056	-0.0063	0.0107	0.0349*	-0.0235
	(.0041)	(.0103)	(.0173)	(.0040)	(.0144)	(.0250)	(.0071)	(.0185)	(.0344)
Depression PGS	0.0024	0.0034	0.0078	0.0063	-0.0002	-0.0051	0.0108	0.0175	0.0061
	(.0038)	(.0144)	(.0198)	(.0065)	(.0156)	(.0287)	(.0081)	(.0249)	(.0442)
Ever-Smoke PGS	0.001	0.0145	0.0191	0.0056	0.0008	-0.0308	0.0147**	0.0208	0.0917**
	(.0036)	(.0098)	(.0211)	(.0045)	(.0151)	(.0247)	(.0071)	(.0193)	(.0373)
ADHD PGS	0.0053	-0.0013	-0.0021	0.0022	0.0224	0.0270	0.0214***	0.0194	-0.0283
	(.0034)	(.0111)	(.0209)	(.0056)	(.0138)	(.0265)	(.0061)	(.0195)	(.0343)
Educational PGS	-0.0126***	-0.0305**	-0.0057	-0.0386***	-0.0463***	-0.0326	-0.0098	-0.0446**	-0.0344
	(.0038)	(.0123)	(.0222)	(.0048)	(.0127)	(.0254)	(.0069)	(.0195)	(.0391)
F-Statistic	1.43	0.66	0.49	0.63	0.72	0.61	6.26	1.95	1.67
P-Value	0.2301	0.6230	0.7426	0.6451	0.5816	0.6546	0.0002	0.1020	0.1559
N	5,716	786	786	5,716	786	786	5,716	786	786
# of Families		366	366		366	366		366	366

Notes: Basic controls consist of age, gender, birth order and the first 20 principal components of the genetic data. Full controls consist of the basic controls and controls for the number of months the respondent was breastfed, mother's grades of schooling, dummy variables equal to 1 if (i) the mother reports being obese, (ii) the mother reports being a smoker, (iii) the mother reports being in excellent or very good health, (iv) the parents are married, (v) if the child had no health insurance. Missing values are imputed with the sample mean and controlled for with dummy variables for missing data. Standard errors for the main sample are clustered at the community level. Standard errors for the sibling sample are clustered at the family level. The F-statistic and p-value are for the joint test of significance for the health PGSs. ***significant at 1% **significant at 5% *significant at 10%.