How Smoking, Drugs, and Obesity Affect Education, Using Genes as Instruments

Edward C. Norton
Professor, Department of Health Management and Policy, and
Professor, Department of Economics
University of Michigan

Euna Han
Post-doctoral Fellow, Institute for Health Research and Policy
University of Illinois at Chicago

December 2008

Please address correspondence to:

Edward C. Norton
Department of Health Management and Policy
M3108 SPH II
109 S. Observatory St., Room M3108
University of Michigan
Ann Arbor, MI 48019-2029
USA

Tel: 734-615-5738
Fax: 734-764-4338
E-mail: ecnorton@umich.edu

Keywords: Education, genetics, smoking, drinking, Add Health

SUMMARY

One of the fundamental questions in health economics is the relationship between two types of human capital—health and education. A persistent research problem, though, is measuring a causal effect, because health is endogenous in models that predict educational attainment. We use genetic information as instrumental variables for obesity and substance use. We analyze data from the National Longitudinal Study of Adolescent Health (Add Health) because a subset of respondents contributed DNA samples. Six genes identified from the DNA samples were originally chosen specifically because they are known to be related to dopamine or serotonin, neurotransmitters related to satiation, obesity, substance abuse, and other behaviors. Four additional genes were recently added to the public data set. The genetic information provides strong instrumental variables, with large incremental effects. Several genes predict changes in BMI of more than three BMI units, in number of cigarettes smoked by more than two per day, and in the probability of using illegal drugs by more than 30 percentage points. While simple cross-sectional results show a strong negative effect of smoking during adolescence on eventual high school education, two-stage least squares results show no statistically significant effect. As predicted in the conceptual framework, the endogeneity bias is negative. This result is consistent with a story of omitted variables such as the discount rate.

Acknowledgements

The authors have no financial or personal relationships between themselves and others that might bias their work.

This research uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). No direct support was received from grant P01-HD31921 for this analysis.

INTRODUCTION

One of the fundamental questions in health economics is the relationship between two types of human capital—health and education. In cross-sectional data, there is generally a positive correlation between measures of health and education. Adults with higher levels of education typically report being in better health. Economists have studied this positive relationship over the life-cycle, and find that it holds from adolescent years through retirement.

A persistent research problem, though, is measuring a causal effect, because health and education affect each other. Reverse causality and omitted confounding variables inevitably lead to endogeneity bias in empirical work. Therefore, the importance of the topic means that finding good instrumental variables would advance the literature. Because achieving a high school education is the most important milestone in education, we focus on how certain behavioral measures of health and health behaviors affect high school graduation. Obesity, smoking, and illegal drug use during adolescence all may lower the probability of earning a high school degree. Or, they may just be related to underlying differences in the unobserved rate of time preference. Such an omitted variable would bias a simple regression coefficient in the negative direction. It is our aim to correct that bias.

We use genetic information as instrumental variables. Specifically, we use genotypes as instruments for phenotypes and behaviors. Two other economic paper use genetic information as instrumental variables to study how health affects education (Ding et al., 2006; Fletcher and Lehrer, 2008). Certain genes are known to be related to obesity and substance use. Instrumental variables created from genetic information allow us to control for the endogeneity of obesity and substance use and to obtain consistent estimates of their causal effects on education. In other related literature, economists have used siblings and twins as controls or instruments by arguing that biological siblings and twins share many genes. Genetic information can greatly improve this approach if the genes are targeted to the endogenous variables, as they are in this study. We test for whether the instruments can be excluded from the main equation, and generally follow our approach in a related prior paper about obesity and labor market outcomes (Norton and Han, 2008).

We analyze data from the National Longitudinal Study of Adolescent Health (Add Health). In Wave III of Add Health, a subset of the Add Health respondents contributed DNA samples. Six genes identified from the DNA samples were originally chosen specifically because they are believed to be related to obesity or other behaviors, and to have a relatively high prevalence in the population. Polymorphisms in these genes have been linked to obesity through behavior (Blundell 1977; Hoebel et al., 1989). Four additional genes were recently added to the public data set. All affect how the central nervous system regulates satiation. Several pass all specification tests for good instruments.

While simple cross-sectional results show a strong negative effect of obesity and substance use during adolescence on eventual high school education, two-stage least squares results show no statistically significant effect. As predicted in the conceptual framework, the endogeneity bias is negative. This result is consistent with a story of omitted variables such as the discount rate. The genetic information provides strong instrumental variables.

BACKGROUND

Substance use during early years of human capital formation is particularly important because adverse effects may linger over the life-cycle. These adverse effects include addiction and a detrimental effect on wages (Yamada, Kendix, and Yamada, 1996). Illicit drug use is prevalent among American adolescents, and the prevalence rate has increased steadily over the last few decades. In 1991, approximately 12% and 16% of 10th and 12th grade students, respectively, reported using any illicit drugs in the past month. Those rates increased to 18% and 23% for the 10th and 12th graders, respectively, in 2004 (Johnston, O'Malley, and Bachman, 2002).

Substance use matters for educational attainment for several reasons (Register, Williams, and Grimes, 2001). First, substance use may adversely affect cognitive skills. Second, substance users may have a higher rate of time preference, which diminishes the importance of education to the substance user. Third, substance use may change a person's time preference due to the addictive nature of the substances. Fourth, substance users may prefer current income rather than future income and thus drop out of school to support their addictive habit. On the other hand, substance use may also relieve stress and therefore help increase academic productivity. In sum, we expect a negative effect, if any, of substance use on educational attainment.

The effect of illicit drug use (particularly, marijuana) on education outcomes is confounded by the reverse causality from schooling to illicit drug use and from omitted variables correlated with both education and illicit drug use. Students who used illicit drugs also may have selected to do so because they were likely to drop out before their senior year (Pacula, Ringe, Ross, 2003). Most of economic literature asking how substance use affects educational accomplishment acknowledges the endogeneity of substance use and addresses the issue using state-level legal factors or additional controls for school characteristics (Yamada, Kendix, Yamada, 1996; Chatterji, 2006; Pacula, Ringe, Ross, 2003; Register, Williams, and Grimes, 2001), whereas a few do not address the issue (Mensch and Kandel, 1988; Bray et al., 2000).

Yamada, Kendix, and Yamada (1996) examine the contemporaneous effects of substance use among adolescents in the twelfth grade during the 1981-1982 academic year on high school graduation using the NLSY. For instruments to identify alcohol consumption and marijuana use they use legal factors, such as marijuana decriminalization and the minimum drinking age for beer and liquor, and the beer tax rate and liquor price. Their results show that frequent marijuana use (using marijuana in each of the ten months during the academic year) lowers the probability of high school graduation by 5.6 percent. However, more recent studies criticized this study because of its small sample size (*n*=1035), it did not distinguish late high school graduation from on-time graduation, and it did not control for school factors that are likely to affect both school completion and substance use (Pacula, Ringe, Ross, 2003). Finally, their study does not address the long-term effects of adolescent drug use on educational attainment beyond the high school level (Register, Williams, and Grimes, 2001).

Chatterji (2006) estimates the effect of past illicit drug use during high school on the number of years of schooling completed when most respondents reach 26 years old. Using data from the National Education Longitudinal Study, the author controls for the endogeneity of illicit drug use by state drug policies and school characteristics as instrumental variables. In a reduced form model, the study reports a 0.2–0.3 year and 0.2–0.4 year decrease in the years of schooling at age 26 for marijuana users and cocaine users, respectively, in 10th or 12th grades. However, the IV estimates are not statistically significant, although a negative association remains.

Pacula, Ringel, and Ross (2003) explore the causal effect of marijuana use on cognitive impairment, measured as 10th graders' performance on standardized tests, using the National Education Longitudinal Survey. This study is unique because it estimates the effect on test scores. It also uses a continuous measure of marijuana use measured from the number of times a person used marijuana in the past 30 days, 12 months, and in their lifetime. Additionally, this study extensively controls for covariates that would be correlated with both illicit drug use and the test score. It controls for school-level variables such as high school program types (college prep, vocational, or regular), proportion of white students, proportion of students receiving free or reduced-price lunches, proportion of dropout among 10th graders, the minimum salary paid to teachers, type of school (catholic, other private, public), and the urbanicity of the school location. They also simultaneously control for current use of alcohol, frequency of binge drinking, initiation of cigarette smoking in 8th grade, 8th grade GPA, hours spent on homework in 8th grade, time spent working at a job, an index of negative behaviors (such as getting sent to the office or disrupting class) in 8th grade, the number of stressors in the past two years (including parental divorce, death in the family, school change, serious illness), whether students had ever been offered drugs at school. They also try to identify marijuana use using state level price information including geometric mean price of an ounce of commercial grade marijuana, the maximum fine and minimum jail time statutorily imposed for marijuana possession offences involving 10 grams of marijuana, and state decriminalization status. They do not find statistically significant negative effect of marijuana use on composite, math, and reading standardized test scores in their reduced form model. However, their first-differencing model shows a smaller but statically significant negative effect of marijuana use and test scores: initiating marijuana use at between 10th and 12th grade lowers math test score by 0.65 points.

Register, Williams, and Grimes (2001) estimates the effect of early illicit drug use on subsequent educational attainment as years of education using males in the 1984 and 1992 waves of NLSY79. They estimate drug use on all drugs, hard drugs, and marijuana. They identify illicit drug use at 1984 by using variations in state in the decriminalization of marijuana use at age 14. A statistically significant negative effect of early illicit drug use on the education attainment is found for the total sample (1.1, 0.8, and 0.8 less years of education for all drugs, hard drugs, and marijuana users, respectively, compared to non-users).

Mensch and Kandel (1988) investigate the effect of various substance use (including cigarettes, alcohol, marijuana, and other illicit drugs) on high school dropout status using the 1984 wave of NLSY79. In their study, cigarettes, marijuana, and other illicit drug use is associated with the propensity of high school dropout. However, no significant association between prior alcohol use and high school dropout is found. Because this study does not address the endogeneity of the illicit drug use in their estimation, their study results can not be interpreted with causal implications.

Bray and colleagues (2000) examines the effect of the onset age of cigarettes, alcohol, marijuana, and other illicit drug (cocaine or crack, hallucinogens, stimulants, sedatives or inhalants) use prior to age 16, 17 or 18 on the likelihood of high school dropout by those three ages. They use data from four longitudinal surveys of students in a southeastern US pubic school system during 1985 and 1994. Students were in 6th to 8th grade in the first survey. In their survey they could identify the reason the students left school including graduation, dropout of school, or transfer to another school. The authors argue that estimating separate models by age could be important if the types of dropout are different by age. That is, some students may drop out from high school to get a job assuming continuation of education will not increase their lifetime earnings. Other dropout type would include students who drop out because they lack academic ability to complete high school regardless of their belief about the impact of continuing education on their lifetime earnings. It is possible that later dropout (say at age 18) is likely for the latter reason, while earlier dropout could be a mix of those two reasons. Their estimation shows that initiation of marijuana use prior to school dropout has a statistically significant positive effect on the likelihood of dropping out of the school at age 16 (18 percentage point increase) and age 18 (12 percentage point increase) when controlling for onset age of cigarettes, alcohol, and other illicit drug use. However, contrary to the expectation, they find a statistically significant negative effect of initiation of cigarette use prior to school dropout on the probability of dropout at age 17 (13 percentage point decrease). It is not clear why they find this contradictory effect for initiation of cigarette use.

The association of alcohol use on schooling, such as such as high school completion, college matriculation, and college graduation, is also well studied in the economic literature (Cook and Moore 1993; Yamada, Kendix, and Yamada 1996; Bray et al. 2000; Koch and Ribar 2001; Dee and Evans 2003; Wolaver, 2002; Mullay and Sindelar, 1994). Early studies finds that alcohol use is associated with reduced educational attainment including a reduced probability of graduating either high school or college, less years of schooling completed, or grades. However, more recent studies (Koch and Ribar 2001; Dee and Evans 2003) suggests a modest or near zero negative effect of alcohol use on educational attainment.

Smoking is endogenous in its effect on education. There are numerous previous studies reporting the positive effect of education on smoking initiation decision (de Walque, 2007; Currie and Moretti, 2003), cessation decision for once smokers (de Walque, 2007; Sander, 1995), or the probability of smoking or the probability of smoking regularly (Grimard and Parent 2006). In addition, unobserved family structure (family's attitude toward illicit substance use during high school) or individual capacity would be correlated with both smoking decision and years of education. Therefore, controlling for endogeneity of smoking in estimating its effect on educational attainment is essential to identify causality.

Cook and Moore (1993) report that heavy smokers (smoking more than 15 cigarettes per day yearly) among high school seniors had 0.68 year less highest year completed compared to non-smokers, whereas light smokers (1−15 cigarettes per day yearly) did not show a statistical difference in the highest completed year compared to non-smokers. For the sample aged from 14 years both light smoking and heavy smoking significantly reduced highest completed year of education by 0.4 (light smoking) and 0.9 years (heavy smoking). Because they do not address the endogeneity of teen smoking in their estimation, their results can not imply any causality.

Cook and Hutchinson's (2006) study seeks to explain the predictive effect of youthful smoking on educational accomplishment not only by the time preference but also the social status sensitivity among peers. For example, high-aptitude students or students who highly commit to their school activities may take larger costs for choosing smoking than low-aptitude students or students with low-commitment to their schools (Aloise-Young and Hennigan, 1996). They use NLSY97 and restrict their sample to high school juniors across waves. They find that smoking (smoked at least one cigarette in the previous 30 days from the time of interview) in 11th grade predicts high school completion and four year college graduation among high school graduates. Specifically, an increase of one standard deviation on ASVAB score increases the odds of graduating high school and four-year college by 1.7 and 3.4 times, respectively, for men, and 2.0 and 2.8 times, respectively, for women. However, they do not address the endogeneity of smoking.

Genetics

A *gene* is a unit of heredity, consisting of a string of DNA. Functionally, genes regulate the production of proteins. Because human chromosomes, which are strings of genetic material divided into genes, come in pairs, each gene has two copies. *Alleles* are variations of the same gene. Humans have two alleles of each gene. The *genotype* is the specific genetic makeup of an individual. The genotype combined with environmental factors produce observable characteristics called *phenotypes*, such as whether or not a person has blue eyes or is obese. An individual's two alleles may be the same, but often differ. When two alleles differ, one may be dominant and one recessive, such as for eye or hair color. Or the phenotype may depend on the combined alleles in another manner, possibly also depending on the environment.

A number of genes have been linked to obesity in the biomedical literature (Comuzzie and Allison, 1998; Snyder et al., 2004). However, it is not as simple as there being a "fat" gene or a "skinny" gene. Instead there is a complex relationship between neurotransmitters in the brain, genes, and obesity. Neurotransmitters, including dopamine and serotonin, regulate food intake, and are thus related to obesity (Guo, North, and Choi, 2006). Certain genes interact with these neurotransmitters. Furthermore, the interaction depends on the exact genotype, with certain polymorphisms of genes related to high obesity and others to low obesity.

The Add Health data set has information on ten relevant genes chosen because they have a high prevalence in the population and a direct role in either dopamine or serotonin. Furthermore, each gene has been shown in epidemiologic literature to be related to obesity (Guo, North, and Choi, 2006). In the Add Health data, we created genetic variables based on whether an individual had a particular genotype in either allele. Several of these variables are highly correlated to obesity and behaviors. Because genes related to neurotransmitters may also affect other behaviors besides those leading directly to obesity, we were not surprised to find that some of our genetic variables were also related to smoking and drug use.

METHODS

Our main model predicts educational attainment, measured as a high school degree (or the equivalent), as a function of obesity, smoking, and illegal drug use, while controlling for other observable factors. The prior economic literature suggests that obesity and substance use among adolescents reduces educational attainment. The simple correlations between obesity and education and between substance use and education are negative. However, we are interested in measuring the causal relationship. Measuring a causal relationship requires controlling for endogeneity, which is discussed below.

The main outcome is a dichotomous measure of completing 12 years of education, which is generally required for a high school degree. We estimate the main model with a linear probability model. This makes the interpretation of the estimated coefficients easy, and does not substantively change the conclusions when compared to results from a logit model. The main equation is therefore

$$\Pr(HS_i) = \alpha_0 + \alpha_1 BMI_i + \alpha_2 Smoke_i + \alpha_3 Drugs_i + X_i \alpha_x + \varepsilon_i$$

where *i* indexes individuals, *HS* indicates completing 12 years of education, *BMI* is the continuous measure of obesity, *Smoke* is a continuous measure of smoking, *Drugs* is a dichotomous measures of illegal drug use, the vector *X* includes exogenous explanatory variables, ε is the i.i.d. error term, and the parameters to be estimated are α. The goal is to estimate consistent estimates of α_1, α_2, and α_3.

Endogeneity

The primary econometric concern is how to control for the endogeneity of BMI and of substance use (both smoking and drugs). One way to think of why BMI and substance use may be endogenous is through reverse causality. A person with higher education may be better informed about the health risks of being overweight, of smoking, and of drugs. The long-term health consequences of obesity and substance use are well documented in the medical literature and the popular press. A person with higher education may also be better informed about the labor market consequences. Women who are overweight or obese have been shown to have worse labor market outcomes. Illegal drug use can result in job termination. A person with more education may have better access to information, or be better able to process the information that they encounter. The reverse causality story suggests that the coefficient in a simple regression is biased downwards.

Another way to think about the endogeneity problem is that it is due to omitted variables that affect education, obesity, and substance use. One example is a person's discount rate. A person with a high discount rate will tend to invest less in education, as well as less in health capital. They are more likely to consume fatty foods, smoke cigarettes, and use illegal drugs. This explanation also implies that the estimated coefficient in a simple model to predict education is biased downwards.

We use variation in genes associated with obesity and behavior to provide natural variation in obesity and in substance use. This natural variation identifies the effects of obesity and substance use on education. We create variables based on genetic information to form instruments in the two-stage least squares models. At conception some people are naturally predisposed to have a higher or lower BMI than other people. Others are genetically predisposed to be more or less likely to smoke or consume illegal drugs when adults. The first-stage obesity equation is

$$BMI_i = \beta_0^{BMI} + GenIV_i^{BMI} \beta_1^{BMI} + X_i \beta_X^{BMI} + \upsilon_i^{BMI}$$

where *GenIV* is a vector of genetic information that are valid instruments, ν is the error term, and the βs are parameters to be estimated. The equations for smoking and drug use are similar, but obviously have a different dependent variable and may use a different set of genetic instruments.

$$Smoke_i = \beta_0^{Smoke} + GenIV_i^{Smoke}\beta_1^{Smoke} + X_i\beta_X^{Smoke} + \upsilon_i^{Smoke}$$

$$Drugs_i = \beta_0^{Drugs} + GenIV_i^{Drugs}\beta_1^{Drugs} + X_i\beta_X^{Drugs} + \upsilon_i^{Drugs}$$

We control for heteroskedasticity and report robust standard errors for all models with clustering at the school level (Norton et al., 1996).

We conduct a number of robustness checks. For the continuous variables of obesity and smoking, we also estimate models with dichotomous measures of whether the adolescent was obese, and whether the adolescent had ever smoked. We estimate models with all three potentially endogenous variables separately, as well as together.

DATA

The National Longitudinal Study of Adolescent Health (Add Health) is a nationally-representative study of how health-related behaviors in adolescents affect various outcomes in early adulthood. The first wave, which began in 1994, collected individual-, school-, and community-level information on respondents in grades 7 through 12. We analyze data from the first three waves.

Among the 20,745 participants in the Wave I of the Add Health data, DNA information was collected for 2,489 in Wave III. The final sample of 2,027 respondents was obtained among those 2,489 individuals with DNA information, after deleting observations with missing values in all relevant variables (74 respondents missing in DNA information; 1 missing in education; 58 missing in measured BMI; 100 missing in amount of smoking; 28 missing in drug use; 345 missing in school clustering information; 3 missing in other covariates).

Dependent variables

Our dependent variable is the stock of education, which was dichotomously measured as high school or more education, meaning 12 years or more of education. A majority of the final sample (88%) reported to have at least 12 years of education (see Table 1). The stock of education was measured in Wave III, when all respondents are in their 20s and presumably through with non-advanced education.

Explanatory variable of interest

Three explanatory variables of primary interest are body weight status, smoking, and drug use. First, we used body mass index (BMI) as weight in kilograms divided by height in meters squared in Wave 1 to measure body weight status. We used measured—not self-reported—height and weight because of the known biases in self-reported data. In our final sample, BMI ranges from 13 to 46 (mean of 22.5). We also used a dichotomous measure for being obese for individuals whose BMI is 30 or larger. Approximately seven percent of the overall sample is obese as an adolescent.

Second, the number of cigarettes smoked was defined as the number of cigarettes smoked each day during the past 30 days, on the days the respondent smoked. Sample respondents smoked about two cigarettes each day. Persons were coded as ever smoked if they reported that they have ever smoked cigarettes regularly, that is, at least one cigarette every day for 30 days. The proportion of the ever-smoked respondents was

17.6%. Third, we coded a person as having ever used drugs if they reported using marijuana, cocaine, or inhalants during the past 30 days at least once. Slightly more than half of the final sample reported to have ever used drugs (see Table 1). For all the substance use variables, we used collected information from Wave I for the respondents who were older than 14 years at Wave I, and replaced for the information from Wave II for the other respondents who were 14 years or younger at Wave I.

Other control variables

Age at the time of interview in Wave I ranges from 13 to 19 in our sample, and we created a series of seven dummy variables for age 13 to age 19. It is especially important to control for age in the first-stage regressions that predict BMI and substance use, because BMI and substance use tend to rise with age. We use the same age variables for the main equation to predict education, even though education is measured at Wave III (there is an eight year difference for all respondents, so changing the variables would not change the age coefficients). About 17 percent of the sample were non-Hispanic Black and another 15 percent were Hispanic. Less than one percent of the final sample were married (0.02%) at the time of interview. The vast majority of the sample (67%) report being in excellent or very good health status. Four regional areas (northeast, west, south, and Midwest) are represented fairly evenly. Approximately one-third of the final sample reported that cigarettes and alcohol were available in their house, and the proportion of the sample reported availability of alcohol and guns in their house is nearly 26% and 23%, respectively. A much smaller proportion (2.8%) reported any drug availability in the house.

In the estimation of the level of education on body weight status, we additionally control for the hours watching TV, video, or playing video games in an average week, and the extent of physical exercise, such as jogging, walking, karate, jumping rope, gymnastics or dancing during the past week from the time of interview (mild extent if 1−2 time, moderate extent if 3–4 times, and heavy extent if 5 or more time during the past week). The amount of hours watching TV, video, and playing video games were averaged at 16 hours (with a range between 0 to 99 hours), four hours (with a range between 0 to 99 hours), and three hours (with a range between 0 to 60 hours), respectively. Slightly more than half of the final sample engaged in mild (32.8%) or moderate (25.1%) exercise.

Instruments and specification tests

We used a slightly different set of genetic variables for each of the explanatory variables of interest. All the genetic variables were based on whether either of the two alleles for each gene had a specific genotype. A value of one indicates that either one of the alleles (or both) showed a specific polymorphism. Because multiple polymorphisms of one gene may be related to obesity, we sometimes created more than one variable per gene. The genetic instruments come from two of the genes (see Table 2). The 48-bp repeat polymorphism of the Dopamine D4 Receptor (DRD4 gene) and the Monoamine Oxidase A-uVNTR (MAOA gene) are common instruments for an indicator for ever smoked, ever used drug, and the amount of smoke. Also, the RS13280604 single nucleotide polymorphism of the Neuronal nicotinic cholinergic receptors (nAChRs) beta-3 subunit is used for an additional instrument for indicators of ever smoked and ever used drugs. We use only the Dopamine D4 Receptor (DRD4 gene) for BMI. For specific instrument for each substance use variable of interest, see Table 2.

Specification tests confirm that our instruments are strong for all the variable of interest. The instrumental variables are jointly statistically significant at the 1% level in the first-stage regressions, with *F*-statistics ranging between 9.01 and 17.12. These results confirm that the genetic variables are good instruments. Regression-based Hausman tests did not reject the null hypothesis of the exogeneity of BMI at the 5% level. LM tests of the exclusion restrictions did not reject the null hypothesis for all the variables of interest (see Table 3).

RESULTS

First-stage results

The first-stage results are important for establishing the strength of the instruments and providing evidence that the models make sense. Here we interpret the magnitude of the instrumental variables. First, three instruments are used to predict continuous BMI. The magnitudes of these three coefficients are 4.7, 3.6, and –3.8. To put this in perspective, a one-unit increase in BMI for a five-foot person corresponds to an increase of over five pounds; for a six-foot tall person a one-unit increase in BMI corresponds to an increase of over seven pounds. Therefore, genetic variation alone leads to considerable exogenous changes in weight for a person of average height. A change in any of these three genes leads to exogenous changes in weight of at least 15 pounds for most people, and for some more than 30 pounds. These coefficients are of meaningful magnitude, even after controlling for other covariates.

For the number of cigarettes smoked, the six genetic instruments also had large and statistically significant effects. Five of the six were larger than 2.0 in absolute value, and two were greater than 3.0 in absolute value. Genetic variation determined at conception is highly predictive of smoking behavior as an adolescent. Some genes appear to be protective against smoking, while others appear to be positively related. It is interesting that the only two other coefficients of comparable magnitude are being African-American (coefficient = −2.0) and having cigarettes easily available at home (coefficient = 2.2). Therefore, having certain genes provide the same effect on smoking as leaving cigarettes lying around at home, or hiding them all if they are available.

The third endogenous variable in the model is illegal drug use ever. This dichotomous variable was modeled in a linear probability model. Again, the genes predict strongly with high absolute magnitude. Five of the nine instruments had coefficients greater than 0.30 in absolute value. For a linear probability model this implies that having a certain gene changes the probability of ever doing drugs as an adolescent by at least 30 percentage points. Nothing else in the model swings the needle, so to speak, anywhere near as much.

In sum, the genetic instruments not only pass the statistical tests for instruments, but also are large in magnitude. They provide strong exogenous variation in the explanatory variables of interest.

Main results

For the main results, showing the effect of obesity and substance use on the probability of completing a high school degree after controlling for endogeneity, we show results two ways. First, we run three separate models controlling for just one of the three endogenous variables. Then we include all three together. This allows us to see the separate and joint effects.

For obesity, in simple cross section there is no correlation between BMI and the probability of completing high school. Controlling for the endogeneity of BMI lowers the coefficient to –0.03 (p<.05), and controlling for all three endogenous variables changes it back to about –0.02 (not statistically significant).

For smoking the pattern was quite different. Simple regression finds strong negative correlation of smoking and educational attainment. Those who smoked more cigarettes as adolescents were significantly less likely to ever get a high school degree. The estimated coefficient of –0.0067 (p<.05) implies that every additional cigarette smoked during a typical smoking day is correlated with a lower probability of ever graduating by a little less than one percentage point. However, after controlling for endogeneity, the estimated coefficient is no longer statistically significant.

For drug use, the simple correlation is positive. A person who reports ever using illegal drugs as an adolescent is about four percentage points more likely to get a high school degree. The positive correlation is surprising. However, again after controlling for endogeneity the coefficient becomes statistically insignificant.

When all three endogenous variables are included jointly, there is no statistical significance found. Other results from other covariates make sense. Hispanics are less likely to have a high school degree. Those in excellent or very good health are more likely to have a degree. Age does not matter, and this makes sense given that we are looking at a sample of persons in their mid twenties. The results are somewhat robust to alternative specifications.

CONCLUSION

We address an important question in applied microeconomics—what is the causal relationship between health and education, for certain health behaviors among adolescents? Because of the clear endogeneity of obesity and substance use among adolescents, it is important to find strong instrumental variables. We use genetic variation to identify the model. The genes in our data set are known from the biological literature to be related to obesity, substance use, and other behaviors related to gratification and satiation. The genetic variables used as instruments not only pass the standard statistical tests for instruments, they have large coefficients in the first-stage results. These genes predict large swings in BMI, number of cigarettes smoked, and the probability of ever using drugs. The strength of the instruments gives us confidence in the results.

The finding that after controlling for endogeneity, the measures of obesity and substance abuse do not influence the probability of completing a high school degree is important. Although there are large cross-sectional correlations between substance use and education, these disappear after controlling for endogeneity. The genetic component of health and healthy behaviors is strong.

REFERENCES

Aloise-young PA, Hennigan KM. 1996. Self-image, the smoker stereotype and cigarette smoking: developmental patterns from fifth to eighth grade. *Journal of Adolescence*. 163-177.

Blundell JE. 1977. Is there a role for serotonin (5-hydroxytryptamine) in feeding? *International Journal of Obesity* 1(1): 15-42.

Bray JW. 2005. Alcohol Use, Human Capital, and Wages. *Journal of Labor Economics*, 23(2): 279-312.

Bray, Jeremy W; Zarkin, Gary A; Ringwalt, Chris; and Junfeng Qi. 2000. The Relationship between Marijuana Initiation and Dropping out of High School. *Health Economics* 9(1): 9-18.

Chatterji P. 2006. Illicit drug use and educational attainment. *Health Economics* 15: 489–511.

Comuzzie AG, Allison DB. 1998. The search for human obesity genes. *Science* 280 (5368): 1374–1377.

Cook PJ, Hutchinson R. 2006. Smoke signals: adolescent smoking and school continuation. NBER working paper 12472.

Cook PJ and Moore MJ. 1993. Drinking and schooling. *Journal of Health Economics* 12, 411-429.

Currie J., Moretti E. 2003. Mother's education and the intergenerational transmission of human capital: evidence from college openings and longitudinal data. *Quarterly Journal of Economics* 34, 1495-1532.

De Walque D. 2007. Does education affect smoking behaviors? Evidence using the Vietnam draft as an instrument for college education. *Journal of Health Economics* 26, 877-895.

Dee TS, Evans WN. 2003. Teen drinking and educational attainment: evidence from two-sample instrumental variable estimates. *Journal of Labor Economics* 21(1): 178-209.

Ding W, Lehrer SF, Rosenquist JN, Audrain-McGovern J. 2006. The impact of poor health on education: New evidence using genetic markers. NBER working paper 12304.

Fletcher JM, Lehrer SF. 2008. Using Genetic Lotteries within Families to Examine the Causal Impact of Poor Health on Academic Achievement. Yale University working paper.

Grimard F, Parent D. 2007. Education and smoking: were Vietnam war draft avoiders also more likely to avoid smoking? *Journal of Health Economics* 26, 896-926.

Guo G, North, Choi. 2006. DRD4 gene variant associated with body mass: The National Longitudinal Study of Adolescent Health. *Human Mutation* 27(3): 236-241.

Hoebel BG, Hernandez L, Schwartz DH, Mark GP, Hunter GA. 1989. Microdialysis studies of brain norepinephrine, serotonin, and dopamine release during ingestive behavior. Theoretical and clinical implications. *Annals of the New York Academy of Sciences* 575:171-191.

Koch SF, Ribar DC. 2001. A siblings analysis of the effects of alcohol consumption onset on educational attainment. *Contemporary Economic Policy* 19: 162–174.

Johnston LD, O'Malley PM, Bachman JG. Ecstasy use among American teens drops for the first time in recent years, and overall drug and alcohol use also decline in the year after 9/11. University of Michigan News and Information Services: Ann Arbor, MI, 16 December 2002 [On-line]. Available: www.monitoringthefuture. org, accessed 9/30/03.

Mensch B.S. and D.B. Kandel. 1988. Dropping out of high school and drug involvement. *Sociology of Education* 61: 95-113.

Mullahy, J. and Sindelar, J. L. 1994. The direct and indirect effects of alcoholism on income. *Milbank Quarterly* 72: 359–376.

Norton EC, Bieler GS, Ennett ST, Zarkin GA. 1996. Analysis of Prevention Program Effectiveness with Clustered Data Using Generalized Estimating Equations. *Journal of Consulting and Clinical Psychology* 64(5): 919–926.

Norton EC, Han E. 2008. Genetic Information, Obesity, and Labor market outcomes. *Health Economics* 17(9):1089–1104.

Pacula RL, Ringel J, Ross KE. 2003. Does marijuana use impair human capital formation? NBER working paper 9963.

Register, Charles A.; Williams, Donald R. and Paul W. Grimes. 2001. Adolescent Drug Use and Educational Attainment. *Education Economics* 9(1): 1-18.

Sander W. 1995. Schooling and quitting smoking. *Review of Economics and Statistics* 77, 191-199.

Snyder EE, Walts B, Perusse L, Chagnon YC, Weisnagel SJ, Rankinen T, Bouchard C. 2004. The human obesity gene map: the 2003 update. *Obesity Research* 12:369–439.

Wolaver, Amy M. 2002. Effects of heavy drinking in college on study effort, grade point average, and major choice. *Contemporary EconomicPolicy* 20, no. 4:415–28.

Yamada T, Kendix M, Yamada T. 1998. The impact of alcohol and marijuana consumption on high school graduation. *Health Economics* 7: 77–92.

Table 1. Summary statistics of non-genetic variables, Add Health data, Wave III

Variables	Mean	Min.	Max.
Dependent Variable			
High School Education	0.88	0	1
Variables of Interest			
BMI	22.46	12.52	46.32
Smoking, ever	0.18	0	1
Smoking, amount	1.92	0	90
Drugs, ever	0.51	0	1
Demographic Variables			
Age at wave 1 = 13	0.02	0	1
Age at wave 1 = 14	0.10	0	1
Age at wave 1 = 15	0.33	0	1
Age at wave 1 = 16	0.21	0	1
Age at wave 1 = 17	0.18	0	1
Age at wave 1 = 18	0.13	0	1
Age at wave 1 = 19	0.02	0	1
African-American	0.166	0	1
Hispanic	0.147	0	1
Married	0.003	0	1
Perceived Health Status			
Excellent or very good	0.672	0	1
Available at home			
Cigarettes	0.298	0	1
Alcohol	0.263	0	1
Illegal drugs	0.028	0	1
Guns	0.230	0	1
No. students in school	1115.2	26	3546
Regional Variables			
West	0.230	0	1
Midwest	0.322	0	1
South	0.347	0	1
N	2,027		

Table 2. Summary statistics of genetic variables, Add Health data, Wave III

	Mean
MAOA: Monoamine Oxidase A-uVNTR	
MAOA_VA_321	0.504
Allele A has 321 frequency	
MAOA_VA_351	0.460
Allele A has 351 frequency	
MAOA_VB_351	0.687
Allele B has 351 frequency	
M321_either	0.510
Either Allele A or B has 321 frequency	
Mother321	0.006
Allele A has other, B has 321 frequency	
DRD4: Dopamine D4 Receptor	
DRD4A_427	0.052
Allele A has 427 frequency	
D427other	0.051
Allele A has 427 frequency, B has other	
D427either	
Ether Allele A or B has 427 frequency	
DRD4a_379	0.161
Allele A has 379 frequency	
D379other	0.147
Allele A has 379 frequency, B has other	
D619other	0.005
Allele A has619 frequency, B has other	
D379379	0.013
Both alleles have 379 frequency	
DRD4B_619	0.352
Allele has 619 frequency	
D427427	0.001
Both alleles have 427 frequency	
D619either	0.358
Ether Allele A or B has 619 frequency	
DRD2: Dopamine D2 Receptor	
DRD2A_178	0.922
Allele A has 178 frequency	
RS13280604	
RS13280604b_1942	0.764
Allele B has 1942 frequency	
N	2027

Table 3. Specification tests of the instrumental variables

	Specification Tests			
	IV	Over-id	Exogeneity	
Endogenous Variable	**Strength**	***p*-value**	***p*-value**	**Conclusions**
BMI	$F = 9.01$ $p < .0001$	0.374	0.006	Good IVs, endogenous
Smoke, ever	$F = 10.09$ $p < .0001$	0.554	0.803	Good IVs, exogenous
Smoke, amount	$F = 10.99$ $p < .0001$	0.462	0.654	Good IVs, exogenous
Drugs, ever	$F = 17.12$ $p < .0001$	0.123	0.770	Good IVs, exogenous

$N = 2011$. The null hypothesis that the over-identifying instruments are excluded from the main equations was tested using an LM test. The null hypothesis that BMI is exogenous was tested using an NR^2 test.

Table 4. Two-stage least squares results to predict high school education in main equation

Variables	*BMI*	*Smoke*	*Drugs*	*ALL*
Constant	1.57 **	0.80 **	0.71 **	1.31 *
	(0.55)	(0.21)	(0.23)	(0.65)
BMI	−.029 **			−0.018
	(0.011)			(0.021)
Smoke		−0.016 **		−0.009
		(0.0031)		(0.012)
Drugs			0.07	−0.11
			(0.18)	(0.25)
N	2128	2081	2139	2027

Robust standard errors are in parentheses. *p*-value <0.01: **, <0.05 *. All regressions also include controls for demographics, self-reported health status, availability of risky things at home, and regional indicators