CESifo Working Papers
www.cesifo.org/wp

Examining the Education Gradient in Chronic Illness

Pinka Chatterji
Heesoo Joo
Kajal Lahiri

CESIFO WORKING PAPER NO. 3892
CATEGORY 3: SOCIAL PROTECTION
JULY 2012

CESifo *Working Paper No. 3892*

Examining the Education Gradient in Chronic Illness

Abstract

This study examines the education gradient in three chronic conditions – diabetes, hypertension, and high cholesterol. In the analysis, we take into account diagnosed as well as undiagnosed cases, and we use methods that account for the possibility that unmeasured factors exist that are correlated with education and drive both the likelihood of having the illness and the propensity to be diagnosed with illness if it exists. Data come from the National Health and Nutrition Examination Survey (NHANES) 1999-2008. Our findings show that education is not associated with diagnosed diabetes or hypertension, and it is positively associated with having been diagnosed with high cholesterol. However, when we consider both undiagnosed and diagnosed cases, there is a strong, negative association between education and having diabetes or hypertension. A small, positive association between education and high cholesterol persists, even when we include undiagnosed cases. When we account for the possibility of shared, unmeasured determinants of disease prevalence and disease diagnosis that are correlated with education, we find that for all three chronic conditions, education is negatively associated with having undiagnosed chronic disease.

JEL-Code: I190, I310.

Pinka Chatterji
Department of Economics
University at Albany: SUNY
USA - Albany, NY 12222
pchatterji@albany.edu

Heesoo Joo
Department of Economics
University at Albany: SUNY
USA - Albany, NY 12222
hjoo@albany.edu

Kajal Lahiri
Department of Economics
University at Albany: SUNY
USA - Albany, NY 12222
klahiri@albany.edu

February 6, 2012
This research was supported by the National Institute on Minority Health and Health Disparities, National Institutes of Health (grant number 1 P20 MD003373). The content is solely the responsibility of the authors and does not represent the official views of the National Institute on Minority Health and Health Disparities or the National Institutes of Health.

1. Introduction

Among adults in the US, there are important differences across education groups in the prevalence, morbidity, and mortality associated with chronic disease. Based on data from NHANES 1999-2008, the age-adjusted prevalence rates for diabetes, hypertension, and high cholesterol were 7.6 percent, 25.6 percent, and 25.9 percent respectively among adults with more than 12 years of education, and 14.6, 31.8, 27.0 percent respectively among adults with less than 12 years of education. Mortality differences by education group have been widening in recent years, and part of the explanation for these worsening disparities is education-related differences in chronic disease. Meara, Richards & Cutler (2008), for example, report that about 13 percent of the growth in education-related differences in mortality among US adults between 1990 and 2000 can be attributed to heart disease (Meara et al., 2008).

When estimating education-related differences in chronic disease, economists typically measure chronic disease using self-reported information. This approach may lead to biased estimates of education-related health disparities if individuals have unmeasured characteristics that are correlated with education, associated with disease prevalence, and associated with the propensity to be diagnosed given that a disease exists (Johnston, Propper & Shields, 2009). Chronic diseases are frequently asymptomatic in their early stages, and as a result education may play a particularly important role in increasing the likelihood that individuals with disease are diagnosed and aware of their condition. This empirical issue has important policy implications. The recently released Healthy People 2020 has a new topic area specifically addressing education and other social determinants of health. To measure progress in this domain, we need estimates of the education gradient that account for the fact that for many chronic illnesses, the less educated are both more likely to have disease as well as less likely to be diagnosed given that a disease exists. (http://healthypeople.gov/2020/).

This study examines the education gradient in three chronic diseases (diabetes, hypertension, and high cholesterol) using data from the National Health and Nutrition Examination Survey (NHANES) 1999-2008. Because NHANES includes both self-reports of medical diagnoses as well as medical assessments for these conditions, we are able to estimate education-related disparities in self-reported diagnosed disease as well as education-related disparities in total prevalence of disease, which includes both diagnosed and undiagnosed cases. We then estimate education-related disparities in chronic disease using selection models that account for the possibility that unmeasured factors exist which underlie both total prevalence of disease and self-reports of diagnosed illness among those who have disease. All models are estimated using semi nonparametric methods.

Our findings show that if we consider only diagnosed cases of disease, there is no association between education and disease in the case of diabetes and hypertension. Education is positively associated with being diagnosed with high cholesterol. However, when we account for undiagnosed cases, a different pattern emerges. Education has a strong, negative association with having diabetes and hypertension. A small but statistically significant, positive correlation between education and high cholesterol persists, even when undiagnosed cases are considered. After accounting for shared, unmeasured determinants of disease prevalence and disease diagnosis, we find that more educated individuals are more likely than those with less education to be diagnosed, given they have a chronic disease. Our results indicate that in order to generate credible estimates of the education/health gradient, it is critical to consider both diagnosed and undiagnosed cases of disease.

2. The education gradient in chronic disease – previous research

In the case of diabetes, there is some mixed evidence that education is negatively associated with diabetes prevalence (having diagnosed or undiagnosed disease) and positively associated with having diagnosed illness, given the disease exists. For example, Wilder et al. (2005), based on NHANES III do not find an association between education and the likelihood of having undiagnosed diabetes. Smith (2007), on the other hand, use data from NHANES II, III, and IV and report a mixed pattern of findings – college education (16+ years) is associated with a reduction in the total prevalence of diabetes in NHANES III, and having at least some college education (more than a high school degree) is associated with lower likelihood of being undiagnosed with an existing disease in NHANES IV, but otherwise education is not associated with diabetes prevalence and diagnosis. Neither of these studies account for the possibility that common, unmeasured factors drive both prevalence and diagnosis of disease, and these unmeasured factors may be correlated with education. If this is the case, these estimates of the education gradient in diabetes may be distorted.

Johnston, Propper & Shields (2009), in a study of the income gradient in hypertension based on the Health Survey for England, address this issue by estimating a censored probit model which accounts for the following: (1) the distribution of being diagnosed with hypertension is censored (because we do not observe what diagnosis status would have been for individuals who do not have hypertension if they actually had hypertension); and (2) this censoring may be driven by unmeasured factors that are correlated with education and also affect the likelihood of having hypertension. These authors find a strong income gradient exists in objective, but not in self-reported, measures of hypertension. Estimates from the censored probit model indicate that income and education are negatively associated with having undiagnosed hypertension among those with the disease. These findings indicate that (1) relying solely on self-reports of chronic health conditions; and (2) ignoring selection into disease prevalence and diagnosis can yield misleading estimates of the SES gradient in chronic disease.

Johnston et al. (2009) focus on a single chronic disease (hypertension) and use data from the UK. In the present study, we contribute to this line of research by estimating the education gradient in chronic disease in the U.S. using the NHANES 1999-2008 – data which include both self-reports of medical diagnoses and medical examination results for chronic illness – as well as methods that account for selection. We consider three of the most highly prevalent and costly chronic conditions in the US -- diabetes, hypertension, and high cholesterol. Prior studies based on US data have not accounted for the possibility that unmeasured factors correlated with SES, disease prevalence, and disease diagnosis may confound estimates of the education-health gradient.

3. Data, Definitions, and Sample Statistics

A. Data

This study uses data on adults aged 25 to 75 from NHANES 1999-2008. The NHANES is a nationally representative survey designed to collect detailed health and nutritional information from adults and children in the United States. In 1999-2008, NHANES interviewed 20,465 individuals aged 25 to 75. Among interviewed individuals, 19,492 individuals (95.2%) participated in the medical examination portion of the NHANES survey, which includes a blood draw and three readings of blood pressure.[1] NHANES respondents who participated in the medical exam were randomly assigned to a morning or an afternoon/evening medical exam. Of the 19,492 participants, 9,541 individuals (48.9%) were assigned to the morning session and the remaining 9,951 individuals (51.1%) participated in the afternoon/evening session. As a result, 8-h fasting plasma glucose test results, which we use to objectively measure diabetes (described in the next section), are only available for those examined during the morning session, while objective blood pressure and total cholesterol measurements are available for both morning and afternoon/evening session participants. Thus, 8,282 individuals have objective measures of diabetes, while 14,885 and 18,318 individuals have objective measures for hypertension and high cholesterol. Our final analytic samples, which are limited to respondents with non-missing information for all variables used in the analysis, include 6,879, 12,344, and 15,122 respondents for the analyses of diabetes, hypertension and high cholesterol, respectively.

B. Definitions

To define the total prevalence of diabetes, hypertension, and high cholesterol, we include both diagnosed and undiagnosed individuals. That is, we draw on both self-reported information on whether respondents have ever been diagnosed by a doctor with these conditions as well as

[1] We exclude from our hypertension analysis sample respondents who do not provide both a 2nd and a 3rd reading of blood pressure since we use the average of 2nd and 3rd reading as an objective measure of hypertension. We also exclude respondents who report having alcohol, coffee, or cigarettes in the 30 minutes before measuring blood pressure.

results from the NHANES blood draw conducted during the medical exam portion of the survey. Note that if we were to rely on self-reports of chronic conditions only, we would not capture those who actually have chronic conditions but are undiagnosed. If we were only to use the medical examination results, we would only capture individuals with uncontrolled disease and would miss individuals who have a chronic condition but are controlled by medication.

In the case of hypertension, the total prevalence includes respondents whose medical examination findings indicate systolic blood pressure over 140 mmHg or diastolic blood pressure over 90mmHg, as well as respondents who self-report that they are currently taking antihypertensive medication (Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure, 2003). Similarly, the total prevalence of high cholesterol includes respondents who self-report currently using cholesterol lowering medication, as well as respondents whose medical examination results show a total cholesterol level over 240 mg/*dl* (ATP III, 2001). The total prevalence of diabetes includes those who self-report a previous diagnosis of diabetes from a health professional, or whose medical examination results show a fasting plasma glucose (FPG) over 7.0mmol/*l* (126mg/*dl*) (American Diabetes Association, 2009).

We define being undiagnosed with an existing disease as the following – conditional on having the disease (based on the prevalence definition described above), the respondent responds negatively to the NHANES question "Have you ever been told by a doctor or other health professional that you had (diabetes, hypertension, blood cholesterol level that was high)?"[2] Note that there is potential for misreporting in either direction – we focus on the problem of undiagnosed individuals who truly have disease, but it is also possible that respondents self-

[2] For high cholesterol, the NHANES respondents are initially asked "Have you ever had your blood cholesterol checked?" Those respondents who report "no" to this question but have blood test results indicating high cholesterol are considered to be unaware of having high cholesterol.

report having an illness, and their medical examination does not indicate the existence of an illness. These cases may indeed be true cases of disease if individuals are effectively controlling disease through lifestyle changes and/or have less severe disease. However, these cases also may be false positives. Since we cannot separate these two groups in the data, we do not focus on this point in the paper.

C. Sample Statistics

Table 1 shows weighted sample characteristics by education group (low - <12 years of education; mid - 12 years of education; high - >12 years of education) for the diabetes, hypertension, and high cholesterol samples. There is a clear educational gradient in health, access to health care, and health behaviors. For example, in the sample used to examine diabetes outcomes, about 26 percent of those in the high education group (>12 years) report being in excellent overall health while only around 11 percent of those in the low education group (<12 years) report excellent self-assessed health. Rates of self-reported diabetes and hypertension (but not self-reported high cholesterol) are higher in the lower education groups vs. the high group. Less educated individuals are more likely than the more educated to have an overnight hospital stay in the past year, but they are less likely than the more educated to report having access to routine care (Table 1). In the diabetes analysis sample, about 36 percent of low educated individuals are current smokers vs. 17 percent in the high educated group.

The top panel of Table 2 shows the weighted chronic disease prevalence rates (which include diagnosed and undiagnosed cases, as described above) by education group. The high educated (>12 years) group has significantly lower prevalence of chronic disease compared to the low educated (<12 years) group, particularly in the case of diabetes. Only 7.6 percent of the high education group has diabetes while 16.0 percent of the low education group suffers from diabetes. The bottom panel of Table 2 shows weighted percentages of disease diagnosis, or the proportion of those with the disease who self-report being diagnosed with it, by education group. The only significant educational gradient in awareness occurs for high cholesterol -- 66.3 percent of those who have high cholesterol and are in the low education group have been diagnosed with high cholesterol, while 72.7 percent of those who have high cholesterol and are in high education group have been diagnoses with the disease. For diabetes, the lowest diagnosed group is the mid education group (12 years) and diagnosis rates between the high and low education group are not significantly different. For hypertension, diagnosis rates do not vary much by education group (76.4-78.0 percent).

We emphasize that the descriptive relationships between disease diagnosis and education that we observe in Table 2 may be confounded by the other factors, such as the severity of health conditions. If less educated people are more likely to be in later stages of chronic conditions than more educated people, this greater severity may make them more likely to be diagnosed with illness. This type of bias from unmeasured severity should be considered when estimating the education gradient in disease diagnosis. In fact, the data support the idea that lower educated people have more severe conditions, and that people with severe conditions are more likely to be diagnosed with and aware of having an existing chronic disease. Co-morbidity is more common among individuals in the low and mid education group vs. the high education group (Figure 1), and, among individuals with chronic disease, high educated people (>12 years) are more likely than lower educated people (<12 years, 12 years) to report being in excellent health (Figure 2).

Figure 3 similarly supports the idea that those who are in worse overall health are more likely to be diagnosed with and aware of an existing health condition than those who are in better overall health. Only about 55 percent of people who report "excellent" overall health and actually have a chronic disease (diagnosed or undiagnosed) report having been diagnoses with the disease, but over 80 percent of people who report “poor” overall health have been diagnosed with and are aware of having an existing chronic condition. These relationships may indicate that severity increases the probability of diagnosis, but the causal pathway also may run in the opposite direction if self-assessed health is updated by being diagnosed with a chronic condition.

4. Model

Initially, we estimate two sets of models in which we examine: (1) the education gradient in self-reported diagnosis of diabetes, high blood pressure, and high cholesterol; and (2) the education gradient in the total prevalence of each of these three diseases (diagnosed and undiagnosed). The dependent variable in these models is a binary indicator of either self-reported disease, or a binary indicator of whether the respondent either self-reports disease and/or has the disease based on the NHANES medical examination results. For estimation of these models, we use semi nonparametric univariate binary choice models as an alternative to standard probit models because in several cases, we rejected normality based on a log-likelihood ratio test. We estimate a basic and an extended specification of each model. The basic specification includes controls for education, poverty-income ratio, race/ethnicity, gender, marital status, age, age squared, survey year, and whether the respondent reports having relatives with diabetes and hypertension. The extended specification additionally includes controls for self-reported co-morbid chronic conditions, whether the respondent is overweight or obese, and smoking status. We interpret findings from the extended specification with some caution since the additional variables included are potentially endogenous – for example, individuals with diabetes may quit smoking or lose weight to better control their disease.

Next, we account for the possibility that common, unmeasured factors underlie both disease prevalence (the existence of disease, diagnosed or undiagnosed) and undiagnosed disease (respondent does not self-report being diagnosed with disease, given that disease exists). To do so, we estimate the semi-nonparametric bivariate binary-choice model with sample selection discussed by Gallant and Nychka (1987)[3]. This model has two advantages that apply to our problem. First, it allows for non-zero covariance, ρ, between the disease prevalence and undiagnosed equations, thus allowing for shared, unmeasured determinants of disease prevalence and undiagnosed disease. Second, the distribution function of the latent regressor errors has a less restricted form compared to that of a parametric approach, such as the widely-used bivariate probit model with sample selection. The semi-nonparametric approach assumes that error distributions are unknown and estimated by Hermite polynomial expansion (De Luca 2008). This appears to be important in our case, since inspection of the residuals from the model shows appreciable deviations from normality, mainly in the diabetes and high cholesterol models.

More formally, the basic structure of the model is:

$$Y_j^* = \beta_j{}'X_j + U_j \qquad j=1,2 \qquad (1)$$

$$Y_1 = 1 \qquad if\ Y_1^* \geq 0$$

$$Y_2 = 1 \qquad if\ Y_2^* \geq 0\ \ and\ \ Y_1^* \geq 0$$

where Y_j^* is unobservable and Y_j is an observable outcome vector of n_j individuals ($n_1 > n_2$). In our case, Y_1 is a binary indicator of disease prevalence (whether the respondent has disease, diagnosed or undiagnosed) and Y_2 is a binary indicator of whether the respondent self-reports having been diagnosed with the disease, given that the disease exists. Since our interest is in

[3] We use the STATA command snp2s written by De Luca (2008) for estimation.

understanding how the relationship between education and undiagnosed disease affects estimation of the education gradient in chronic disease, Y_2 is only of interest when the respondent actually has the disease, $\mathrm{Y}_1 = 1$. To sharpen the identification of the model, we include two variables in the disease prevalence model that are not included in the disease diagnosis equation. These variables are: whether the respondent has relatives with hypertension, and whether the respondent has relatives with diabetes. Presumably, these variables will capture genetic aspects of the disease, but they should not be directly related to disease diagnosis, holding other factors constant.

The joint log likelihood function for bivariate binary-choice model with sample selection is given by:

$$lnL = \sum_{i=1}^{n} Y_{i1}Y_{i2}ln\Pi_{i11}(\beta) + Y_{i1}(1 - Y_{i2})ln\Pi_{i10}(\beta) + (1 - Y_{i1})ln\Pi_{i0}(\beta_1) \qquad (2)$$

where $\beta = (\beta_1, \beta_2, \rho)$, and the probabilities underlying three possible combination of outcomes Y_1 and Y_2 are given by:

$$\Pi_{11}(\beta_1, \beta_2) = \Pr(Y_1 = 1, Y_2 = 1) = 1 - F_1(-\mu_1) - F_2(-\mu_2) + F(-\mu_1, -\mu_2)$$

$$\Pi_{10}(\beta_1, \beta_2) = \Pr(Y_1 = 1, Y_2 = 0) = F_2(-\mu_2) - F(-\mu_1, -\mu_2)$$

$$\Pi_0(\beta_1, \beta_2) = \Pr(Y_1 = 0) = F_1(-\mu_1)$$

where $\mu_j = \beta_j{}'\mathrm{X}_j$. F_j is an unknown marginal distribution function of the latent regression error U_j, j=1,2 and F is a unknown joint distribution function of (U_1, U_2) in semi-nonparametric approach while parametric approach assume Gaussian distribution functions for both F_j and F.

Since F_j and F is unknown, we need to estimate those functions. In the semi-nonparametric approach we use in this paper, an unknown joint density f of (u_1, u_2) is approximated by a Hermite polynomial expansion (Gallant & Nychka 1987). The discussion below is based on DeLuca (2008). The approximated joint density function f^* is given by:

$$f^*(u_1, u_2) = \frac{1}{\psi_R}\tau_R(u_1, u_2)^2\varphi(u_1)\varphi(u_2) \qquad (3)$$

In Equation 3, φ represents the standardized Gaussian density, $\tau_R(u_1, u_2) = \sum_{h=0}^{R_1}\sum_{k=0}^{R_2}\tau_{hk}u_1^h\, u_2^k$ is polynomial in u_1 and u_2 in order $R = (R_1, R_2)$, and ψ_R is the normalization factor: $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\tau_R(u_1, u_2)^2\varphi(u_1)\varphi(u_2)du_1du_2$ (De Luca 2008). To get the approximated marginal densities $f_1^*(u_1)$ and $f_2^*(u_2)$, we integrate $f^*(u_1, u_2)$ with respect to u_2 and u_1, respectively. Using approximated density functions f^*, f_1^*, and f_2^*, we get the approximated distribution function F^*, F_1^*, and F_2^*. We substitute F^*, F_1^*, and F_2^* into unknown F, $F_1$and F_2 in equation (2). Then, we estimate β_1, β_2 using pseudo-ML estimation.[4] Based on this model, we compute and report marginal effects – the effect of education on having undiagnosed chronic disease using Equation 4 below:

$$\text{E}(Y_{2i} = 1)=\text{Prob}(Y_{2i} = 1)=1-F_2(-\beta_2 X_2) \qquad (4)$$

5. Results

Tables 3 and 4 show results from semi nonparametric univariate binary choice models in which we examine the education gradient in chronic illness. Table 3 shows results from models in which the dependent variable is a binary indicator of whether or not the respondent self-reports having been diagnosed with diabetes (Panel 1), hypertension (Panel 2), or high

[4] De Luca(2008) provide a detailed explanation about the approximation.

cholesterol (Panel 3). Table 4 shows results from a similar set of models, but in these models the dependent variables measure the total prevalence of disease – that is, the dependent variables include both diagnosed and undiagnosed cases.

The results in Table 3 indicate that after controlling for other factors, having 12 years of education (relative to less than 12 years) is not associated with being diagnosed with diabetes or hypertension (Panels 1 and 2, Table 3). Having more than 12 years of education is negatively associated with having been diagnosed with diabetes and hypertension, but this effect is marginally statistically significant in the basic model and not statistically different from zero in the extended model, which includes additional controls for health status (Panels 1 and 2, Table 3). In the case of cholesterol, higher educated people are actually more likely to report having been diagnosed with high cholesterol than less educated people (Panel 3, Table 3). This effect is statistically significant in both the basic and extended models. There is also a strong, positive income gradient in self-reported diagnosis of high cholesterol (Panel 3, Table 3).

When we also consider undiagnosed cases, however, a different pattern of findings emerges (Table 4). There is now a strong post-secondary education gradient in both diabetes and hypertension – individuals with at least some college education are less likely to have diabetes and hypertension in both the basic and extended models (Table 4, Panels 1 and 2). There is a strong income gradient in diabetes, but not in hypertension. African-Americans are more likely to have diabetes and hypertension than non-Latino whites, while Latinos are more likely to have diabetes and less likely to have hypertension compared to non-Latino whites. In the case of high cholesterol, there is a mixed pattern of findings. In the basic model, there is no relationship between education and high cholesterol. When additional health controls are included in the extended model, however, education is positively related to having high cholesterol (although the estimated coefficient for post-secondary education is very close to zero). African-Americans and Latinos are less likely to have high cholesterol compared to non-Latino whites.

The findings in Tables 3 and 4 suggest that one cannot generate credible estimates of the education gradient in chronic disease by using self-reported disease information alone, since education may be related not only to the onset of a disease but also to the likelihood of being diagnosed with and correctly self-reporting an existing disease. In Table 5, we estimate the effect of education on the likelihood that an individual has undiagnosed disease (e.g., the respondent has diabetes based on the NHANES medical examination, but the respondent does not self-report having been diagnosed with diabetes). The estimates come from a semi-nonparametric bivariate binary-choice model with sample selection, which accounts for the fact that there may be correlated factors that drive both prevalence and diagnosis of disease, and are associated with education.

In the case of hypertension and high cholesterol, the estimated ρ is large and positive across all models, indicating that estimating the prevalence and undiagnosed equations jointly is appropriate. For these two conditions, unmeasured factors exist that are positively associated with disease prevalence and undiagnosed disease, given an individual actually has the disease. In the case of diabetes, however, the estimated ρ is negative in the basic model and small in magnitude and positive in the extended model. Possibly, this difference in the estimated ρ across the models may be because lifestyle factors are more critical in the case of hypertension and high cholesterol, while genetics play a larger role in the case of diabetes. For example, individuals with poor dietary habits may be more likely than those with better habits to develop hypertension and high cholesterol, as well as less likely to be diagnosed if they do indeed have the disease. This may be true to a much lesser extent for diabetes.

The findings in Table 5 show that after accounting for unmeasured, correlated determinants of disease prevalence and diagnosis: (1) having some post-secondary education reduces the likelihood of having diabetes and hypertension, but has no association with having high cholesterol; and (2) given that disease exists, those with post-secondary education are less likely to be undiagnosed for all three chronic diseases. The magnitude of the effect of college education on the probability of being undiagnosed is a reduction of 6 percentage points in the probability of having undiagnosed diabetes (an 23% reduction at the sample mean), a reduction of about 4 percentage points in the probability of having undiagnosed hypertension (a 18% reduction at the sample mean), and a reduction of 1 percentage point in the probability of having undiagnosed high cholesterol (an 3% reduction at the sample mean). In the case of high cholesterol, in addition to post-secondary education, high school completion is also negatively associated with being undiagnosed.

There is a strong income gradient in diabetes prevalence, but income is not related to the prevalence of hypertension and high cholesterol after accounting for other factors. Among those with disease, higher income individuals are less likely to be undiagnosed than lower income individuals in the case of high cholesterol; however, income is not related to diagnosis for hypertension and, surprisingly, higher income is actually associated with an increase in the probability than diabetic individuals are undiagnosed, after controlling for other factors.

6. Conclusions

As a whole, these findings indicate that relying on self-reported information to estimate the education-chronic disease gradient understates the gradient because less educated individuals are more likely than more educated individuals to have undiagnosed disease. Including undiagnosed cases, and allowing for shared, unmeasured factors that drive prevalence and diagnosis, reveals that education plays an important role not only in having disease but also in being diagnosed with an existing disease. Higher income individuals with high cholesterol are more likely to be diagnosed with it than lower income individuals, but there is no association between income and hypertension diagnosis and a counter-intuitive, negative association between income and diabetes diagnosis. Income, then, plays a somewhat different role than education in the prevalence and diagnosis of chronic disease.

Our findings for undiagnosed diabetes are similar to those of Smith (2007), who uses earlier NHANES data (1999-2002) and a standard probit model to estimate the effect of education on the probability of being undiagnosed, among those who have diabetes. He reports an 8 percentage point reduction associated with having at least some college education – we find a 6 percentage point reduction, using more recent data and empirical methods accounting for correlation between the unmeasured determinants of having diabetes and being undiagnosed given that diabetes exists. Like us, Smith (2007) also finds a counter-intuitive positive relationship between income and undiagnosed diabetes, although the estimated coefficient is marginally statistically significant.

Our findings may indicate that disease diagnosis is an important mechanism through which education ultimately affects health. However, to examine this idea further, longitudinal data are needed to examine whether compared to higher educated individuals, less educated people experience earlier onset of disease, later diagnosis of existing disease, or both. If both effects occur, it is important to understand which effect is stronger, and also the mechanisms at work. If the main issue is that the less educated experience earlier onset and higher prevalence, this may be due to factors such as lifestyle and environment. If the primary issue is that the less educated are less likely to be diagnosed, this may result from lack of accessible screening and primary care services. Understanding these linkages will have important implications for public policies targeted at reducing disparities in chronic disease.

References

American Diabetes Association. 2009. “Standards of medical care in diabetes.” *Diabetes Care* 32 (S1): S13-S61.Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL Jr.,

Jones DW, Materson BJ, Oparil S, Wright JT, Jr., Roccella EJ, the National High Blood Pressure Education Program Coordinating Committee. 2003. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure (JNC 7). *Hypertension* 42:1206-1252. Available online at http://hyper.ahajournals.org/cgi/content/full/42/5/1206.

De Luca, Giuseppe. 2008. “SNP and SML estimation of univariate and bivariate binary-choice models.” *The Stata Journal* 8 (2): 190-220.

Gallant, A. Ronald, and Douglas W. Nychka . 1987. “Semi-Nonparametric maximum likelihood estimation.” *Econometrica*, 55 (2): 363-390.

Johnston, David W., Carol Propper, Michael A. Shield. 2009. “Comparing subjective and objective measure of health: Evidence from hypertension for the income/health gradient.” *Journal of Health Economics* 28: 540-552.

Meara, Ellen R., Seth Richards, and David M. Cutler. 2008. “The gap gets bigger: changes in mortality and life expectancy, by education, 1981-2000.” *Health Affairs* 27(2): 350-360.

Smith, James P. 2007. “Nature and causes of trends in male diabetes prevalence, undiagnosed diabetes, and the socioeconomic status health gradient.” *Proceedings of the National Academy of Science* 104(33): 13225-13231.

Third Report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Washington, DC: National Institutes of Health. 2001.

Wilder, Ronald P., Sumit R. Majumdar, Scott W. Klarenbach, and Philip Jacobs. 2005. “Socio-economic status and undiagnosed diabetes.” *Diabetes Research and Clinical Practice*, 70: 26-30.

Table 1: Weighted sample characteristics by education

	(1) Diabetes Analysis Sample				(2) Hypertension Analysis Sample				(3) High Cholesterol Analysis Sample			
	Low	Mid	High	Total	Low	Mid	High	Total	Low	Mid	High	Total
Non-Latino white	47.5	76.4	78.2	72.3	48.7	75.7	79.3	73.1	47.9	75.5	78.6	72.5
African-American	16.2	10.3	9.3	10.8	16.1	9.7	8.4	10.1	16.1	9.9	8.7	10.3
Latino	31.8	9.2	6.7	11.8	30.9	10.5	7.1	12.0	31.2	10.4	7.4	12.2
Other race/ethnicity	4.5	4.1	5.8	5.1	4.4	4.1	5.2	4.8	4.9	4.3	5.4	5.0
Female	50.6	50.4	52.9	51.9	50.8	50.5	52.8	51.9	50.1	49.9	52.8	51.6
Married	57.8	64.9	68.7	65.8	58.3	64.3	67.7	65.3	57.2	63.9	68.0	65.1
Age (years)	47.6	47.0	45.6	46.3	48.2	47.3	45.6	46.5	48.0	47.2	45.6	46.4
Self-reported												
Diabetes	11.6	7.8	5.6	7.2	12.6	7.6	5.2	7.1	12.3	7.4	5.7	7.3
Hypertension	34.2	34.4	26.7	29.9	34.4	31.3	26.4	29.0	34.2	31.6	26.0	28.8
High cholesterol	27.4	32.7	31.1	30.8	28.6	31.2	29.8	29.9	28.4	31.7	30.5	30.4
Overweight	33.8	35.4	34.1	34.4	35.2	34.0	34.9	34.8	35.6	34.7	34.7	34.8
Obese	38.5	36.7	32.0	34.3	37.9	37.5	31.9	34.3	37.3	37.3	32.1	34.3
Ever smoker	24.2	25.2	27.2	26.2	23.0	25.4	26.7	25.7	23.0	25.1	26.5	25.5
Current smoker	36.1	31.0	17.4	24.1	35.5	29.8	16.8	23.2	36.6	30.9	17.3	24.0
Sample size	1,927	1,582	3,370	6,879	3,438	2,883	6,023	12,344	4,260	3,514	7,348	15,122

Note: All numbers are percentages except age.

Table 2: Prevalence and diagnosis of chronic conditions by education group (%)

	Education			Total
	Low	Mid	High	
Total Prevalence – (Diagnosed and Undiagnosed)				
Diabetes	15.9	11.8	7.5	10.1
	(13.6-18.3)	(9.9-13.7)	(6.6-8.4)	(9.1-11.0)
Hypertension	36.3	32.6	26.0	29.4
	(34.0-38.6)	(30.7-34.6)	(24.6-27.4)	(28.2-30.6)
High Cholesterol	29.2	30.9	26.1	27.8
	(27.2-31.2)	(29.1-32.8)	(24.6-27.6)	(26.8-28.9)
Diagnosed -				
Diabetes	72.9	66.5	75.2	72.0
	(67.8-78.0)	(59.3-73.8)	(69.3-81.1)	(68.3-75.8)
Hypertension	78.0	77.0	77.1	77.3
	(75.1-80.9)	(73.8-80.3)	(74.4-79.8)	(75.3-79.3)
High Cholesterol	66.0	69.7	72.4	70.5
	(62.2-69.7)	(66.0-73.5)	(69.5-75.3)	(68.5-72.5)

Notes: 95% CI is in parentheses. Statistics are weighted.

Table 3: Effect of education on probability of having chronic disease (self-reported, 1/0)

	(1) Diabetes				(2) Hypertension				(3) High Cholesterol			
	Basic		Extended		Basic		Extended		Basic		Extended	
	coeff	z	Coeff	z	coeff	z	coeff	z	coeff	Z	coeff	z
Education - Mid	-0.1009	-1.43	-0.1320	-0.71	0.0018	0.05	-0.0119	-0.30	0.0818	1.89	0.0725	2.00
High	-0.1260	-1.88	-0.1211	-0.68	-0.0657	-1.72	-0.0579	-1.50	0.1029	2.29	0.1179	3.17
Poverty income ratio	-0.0991	-4.21	-0.0871	-0.74	-0.0097	-0.98	-0.0027	-0.27	0.0360	2.73	0.0525	4.14
African-American	0.3589	4.32	0.2723	0.73	0.3246	6.33	0.2644	5.71	-0.1335	-2.66	-0.2023	-4.34
Latino	0.2531	3.51	0.2650	0.72	-0.1645	-3.88	-0.2140	-5.02	-0.1314	-2.64	-0.1180	-3.24
Other races	0.1536	1.13	0.1803	0.64	-0.0214	-0.28	0.0142	0.19	-0.0102	-0.16	-0.0054	-0.01
Female	-0.1168	-2.23	-0.1558	-0.71	0.0117	0.43	0.0128	0.46	-0.0513	-1.82	-0.0242	-0.98
Married	0.0380	0.71	0.0360	0.56	-0.0313	-1.06	-0.0561	-1.84	0.0042	0.15	-0.0054	-0.21
Age	0.1053	4.68	0.0636	0.71	0.0801	5.62	0.0457	3.18	0.1052	3.20	0.0532	6.75
Age^2	-0.0007	-3.57	-0.0004	-0.70	-0.0004	-3.22	-0.0001	-0.87	-0.0007	-2.97	-0.0003	-4.26
Relatives have:												
Diabetes	0.7270	5.95	0.5850	0.74	0.1891	5.39	0.0600	2.09	0.2257	3.36	0.1230	3.97
Hypertension	0.1936	3.05	0.1083	0.68	0.4548	7.50	0.4147	7.93	0.2014	3.21	0.1066	3.04
Self-reported												
Diabetes			-	-			0.4672	7.55			0.4611	4.46
Hypertension			0.3268	0.74			-	-			0.4739	5.00
High cholesterol			0.2971	0.74			0.4324	8.56			-	-
Overweight			0.1051	0.66			0.2520	5.23			0.1820	4.38
Obese			0.4518	0.73			0.6621	8.40			0.2322	4.66
Ever smoker			-0.0464	-0.57			0.0043	0.15			0.1177	3.34
Current smoker			0.0104	0.17			0.0140	0.38			0.0458	1.48
Sample size	6,879				12,344				15,122			

Notes: Table shows estimated coefficients and z-statistics from semi nonparametric binary choice model. Omitted categories: low education, non-Latino whites, male, not married, relatives do not have diabetes, relatives do not have hypertension, not self-reported diabetes, not self-reported hypertension, not self-reported high cholesterol, normal BMI, and non-smoker. Survey year dummy variables included in models but are not shown in table.

Table 4: Effect of education on probability of having chronic disease (self-reported + undiagnosed, 1/0)

	(1) Diabetes				(2) Hypertension				(3) High Cholesterol			
	Basic		Extended		Basic		Extended		Basic		Extended	
	coeff	z	coeff	z	coeff	z	coeff	z	coeff	z	coeff	z
Education - Mid	-0.0732	-1.26	-0.1076	-1.31	-0.0032	-0.09	-0.0147	-0.41	0.0611	0.91	0.0445	1.08
High	-0.1890	-3.22	-0.2081	-1.75	-0.1342	-2.66	-0.1384	-3.35	-0.0577	-0.94	0.0018	0.05
Poverty income ratio	-0.0726	-4.43	-0.0819	-1.89	-0.0011	-0.13	0.0034	0.39	-0.0044	-0.29	0.0060	0.63
African-American	0.2920	4.70	0.2655	1.86	0.3927	3.49	0.3289	4.97	-0.2698	-4.39	-0.2720	-2.95
Latino	0.2056	3.75	0.2705	1.87	-0.1071	-2.45	-0.1597	-3.70	-0.2137	-3.56	-0.1795	-2.91
Other races	0.1082	0.99	0.1737	1.16	0.0994	1.39	0.1344	1.96	0.1268	1.13	0.1349	1.69
Female	-0.2184	-4.28	-0.2767	-1.91	-0.0591	-1.69	-0.0710	-2.40	-0.1560	-3.38	0.0184	0.67
Married	-0.0093	-0.21	0.0052	0.11	-0.0236	-0.85	-0.0417	-1.50	0.0502	1.04	0.0605	1.68
Age	0.0808	7.58	0.0790	1.91	0.0629	4.66	0.0515	6.40	0.0471	9.06	-0.0057	-0.43
Age^2	-0.0005	-4.41	-0.0005	-1.76	-0.0002	-1.53	-0.0001	-1.13	0.0001	0.96	0.0003	1.47
Relatives												
Diabetes	0.6194	7.65	0.6291	2.01	0.1254	2.95	0.0266	1.06	0.3203	6.67	0.0469	1.57
Hypertension	0.1199	2.29	0.0340	0.62	0.3697	3.56	0.3328	5.23	0.2515	4.82	0.1006	2.26
Self-reported												
Diabetes			-	-			0.3799	4.46			0.5138	2.55
Hypertension			0.3532	1.96			-	-			0.3328	3.09
High cholesterol			0.2514	1.90			0.2098	4.46			-	-
Overweight			0.1569	1.55			0.1936	4.43			0.2918	3.11
Obese			0.5911	1.98			0.5537	5.60			0.2709	3.15
Ever smoker			0.0268	0.49			-0.0608	-1.95			0.0944	2.15
Current smoker			0.0486	0.75			-0.0678	-1.98			0.0630	1.72
Sample size	6,879				12,344				15,122			

Notes: Table shows estimated coefficients and z-statistics from semi nonparametric binary choice model. Omitted categories: low education, non-Latino whites, male, not married, relatives do not have diabetes, relatives do not have hypertension, not self-reported diabetes, not self-reported hypertension, not self-reported high cholesterol, normal BMI, and non-smoker. Survey year dummy variables included in models but are not shown in table.

Table 5: Effect of education on probability of having disease and probability of being undiagnosed, given disease exists

	(1) Diabetes				(2) Hypertension				(3) High Cholesterol			
	Basic		Extended		Basic		Extended		Basic		Extended	
	coeff	z	coeff	z	coeff	z	coeff	z	Coeff	z	coeff	z
Has disease (diagnosed or undiagnosed)												
Education - Mid	-0.0713	-1.22	-0.1034	-1.69	-0.0167	-0.33	-0.0254	-0.56	0.0375	0.56	0.0490	0.88
High	-0.1859	-3.00	-0.1919	-2.92	-0.1964	-3.56	-0.1786	-3.86	-0.0745	-1.19	-0.0088	-0.17
Poverty income ratio	-0.0739	-3.93	-0.0695	-3.95	0.0003	0.02	0.0052	0.47	-0.0052	-0.32	0.0074	0.55
Undiagnosed disease, given disease exists												
Education - Mid	0.1488	0.74	0.4632	1.47	-0.0472	-0.62	-0.0392	-0.56	-0.1274	-1.85	-0.1243	-2.30
High	-0.3799	-1.54	-0.7328	-2.52	-0.1420	-1.83	-0.1185	-1.61	-0.2060	-2.84	-0.1807	-3.34
Poverty income ratio	0.0517	1.07	0.2107	2.65	-0.0050	-0.26	-0.0127	-0.71	-0.0584	-3.24	-0.0594	-4.41
Rho	-0.1718		0.0187		0.5267		0.4409		0.2811		0.4810	
Sample size	6,879				12,344				15,122			

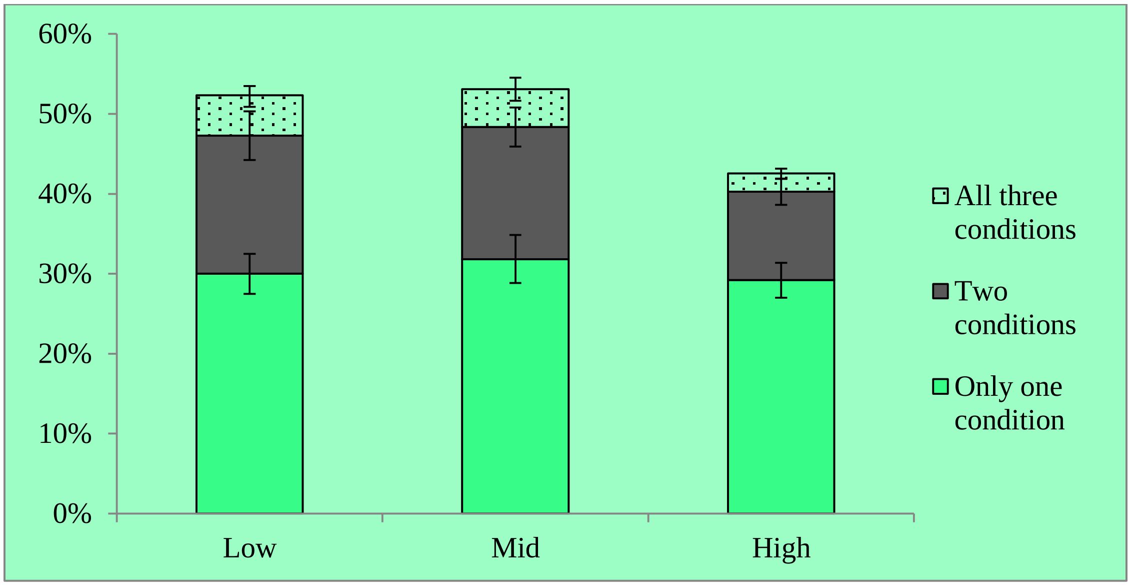
Notes: Table shows estimated coefficients and z-statistics from semi nonparametric bivariate binary choice model with sample selection. Omitted categories: low education, non-Latino whites, male, not married, relatives do not have diabetes, relatives do not have hypertension, not self-reported diabetes, not self-reported hypertension, not self-reported high cholesterol, normal BMI, and non-smoker. Race, gender, marital status, and survey year dummy variables included in basic and extended models but are not shown in table. Self-reported diabetes, hypertension, and high cholesterol, obesity, and smoking status variables included in extended models. Relatives who have diabetes and relatives who have hypertension variables are only included in "has disease" equation.

Table 6: Marginal effect of education on undiagnosed chronic disease

	(1) Diabetes		(2) Hypertension		(3) High Cholesterol	
	Basic	Extended	Basic	Extended	Basic	Extended
Education - Mid	0.0280	0.0395	-0.0277	-0.0126	-0.0138	-0.0085
High	-0.0715	-0.0624	-0.0834	-0.0380	-0.0223	-0.0124
Poverty income ratio	0.0097	0.0180	-0.0029	-0.0041	-0.0063	-0.0041
Sample size	918		4,192		4,444	

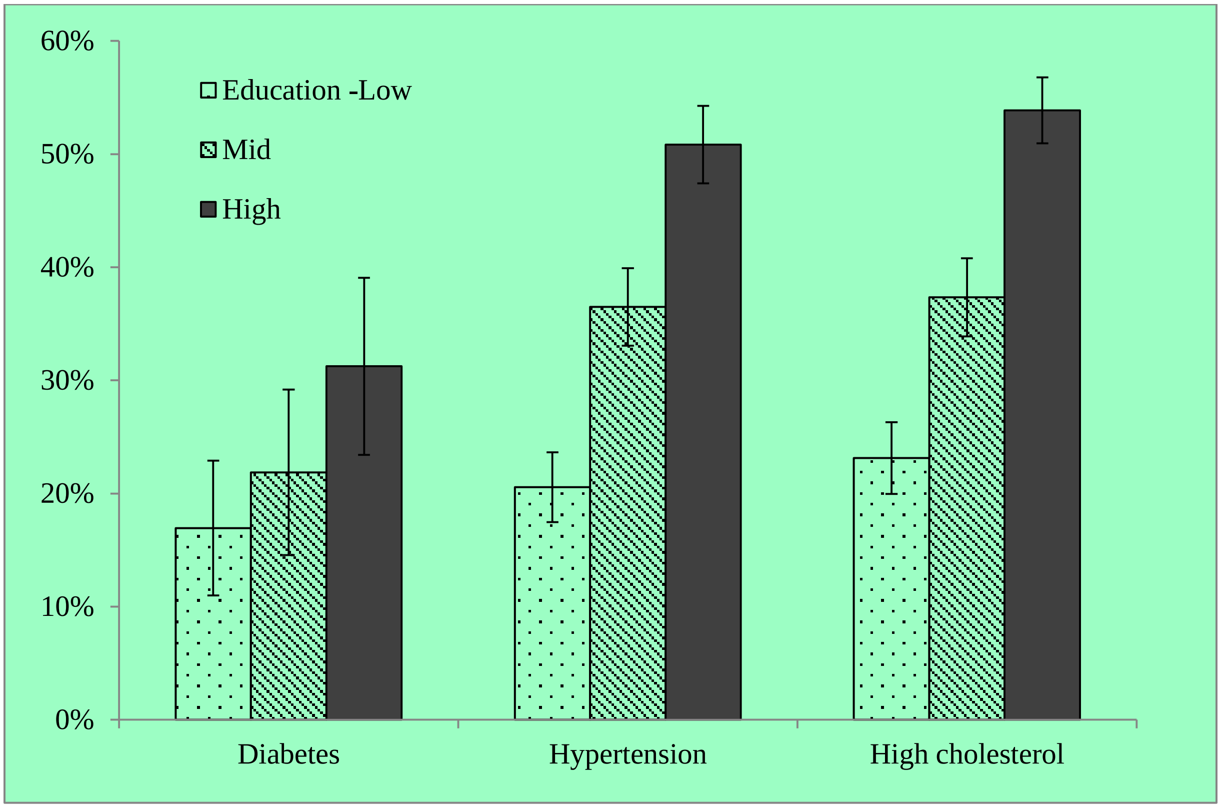
Notes: Table shows estimated marginal effects corresponding to selected coefficients from the semi nonparametric bivariate binary choice model with sample selection. Omitted categories: low education, non-Latino whites, male, not married, survey year ’99-’00, normal BMI and non-smoker.

Figure 1: Percentage of respondents who have diabetes, hypertension, and high cholesterol by number of chronic conditions and by education group, 1999-2008



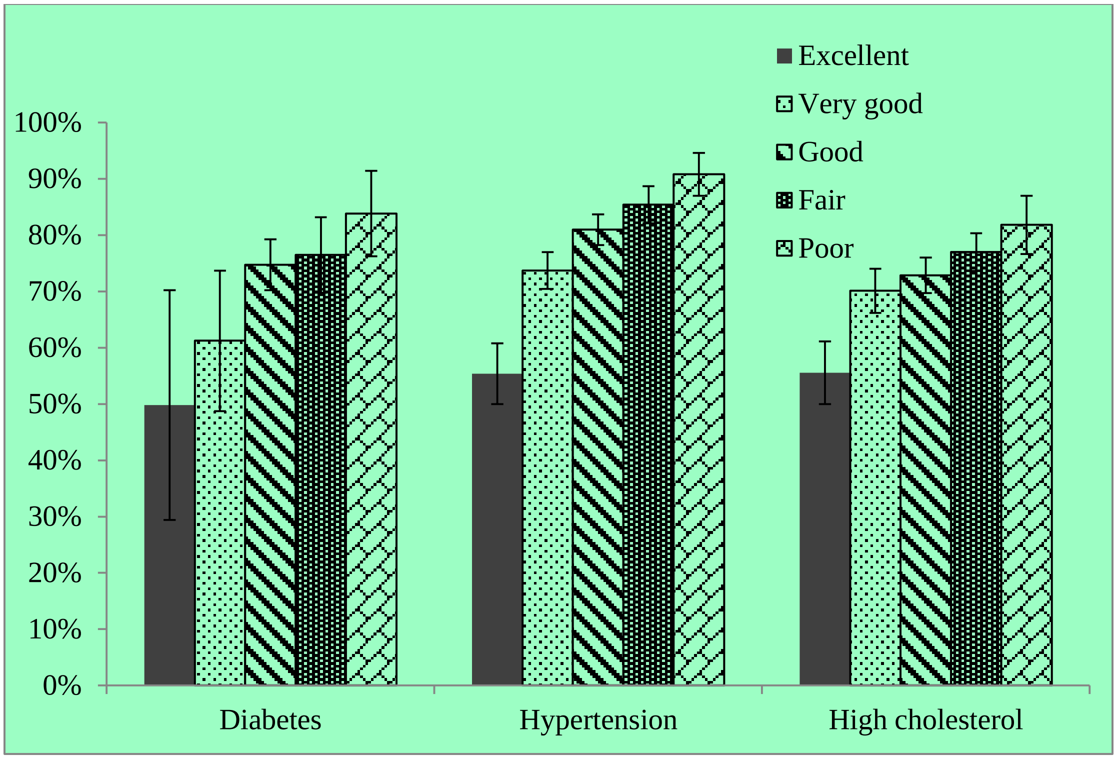
Notes: N= 5,381. Figure based on sample of respondents who have information on all three chronic diseases (diabetes, hypertension, high cholesterol). Statistics are weighted.

Figure 2: Self-assessed health (Excellent or Very Good) by education group among those with chronic disease, 1999-2008



Note: Each sample is limited to respondents who have the disease, diagnosed or undiagnosed.

Figure 3: Percentage diagnosed (among those with disease) by self-assessed health status



Note: Each sample is limited to respondents who have the disease, diagnosed or undiagnosed.