Racial, Ethnic and Educational Disparities in Pharmaceutical Expenditures

Devin Incerti

July 9, 2015

Abstract

This paper identifies differences in prescription drug utilization as a potential mechanism for the well known racial/ethnic and educational gradients in health. A two-part model predicts that, on average, blacks and Hispanics spend $350 and $560 less than whites respectively and that an additional 4 years of education increases prescription drug expenditures by $155. These documented disparities occur for two primary reasons: first, there are differences in the probability of being diagnosed with a disease; and second, there are gradients in expenditures conditional on diagnosis. Access related factors can only explain a small fraction of the black-white ($\approx 3\%$) and Hispanic-white ($\approx 14\%$) differentials, although a slightly larger fraction of the education gradient is due to differences in access ($\approx 34\%$).

1 Introduction

A nearly ubiquitous finding in the health and social science literatures is that there are racial, ethnic and socioeconomic disparities in health status. Link and Phelan (1995) have gone so far as to call social conditions "fundamental causes of disease." Policymakers have tended to focus on access to medical services and differences in the quality of care when addressing these disparities. And while medical care may not be the most important component of an individuals overall health, a focus on medical care is warranted.

Its importance is best illustrated by considering the drastic increase in life expectancy and overall health during the 20th and 21st centuries. Much of the improvements during the latter half of the 20th century coincided with technological innovations and improvements in the quality of medical care. These included new surgical techniques to treat heart attacks, technologies such as special ventilators and artificial surfactants for low birthweight infants and the use of serotonin reuptake inhibitors (SSRIs) to treat depression. Research studies suggest that these innovations are the primary drivers of health improvements in the United States since the 1950s (Cutler and McClellan 2001; Cutler, Rosen and Vijan 2006). In addition, new studies that take advantage of natural experiments have shown that additional medical spending reduces mortality (e.g. Almond et al. 2010; Doyle and Joseph 2011).

It is perhaps no surprise then that disparities in access to care and utilization have been an important research area in the health policy literature. Early studies focused on access measures—such as physician office visits—and documented significant racial and ethic disparities (e.g. Collins et al. 1999; Mayberry, Mili and Ofili 2000; Weinick, Zuvekas and Cohen 2000), while more recent work has progressed toward disparities in use (e.g. Blanco et al. 2007; Cook, McGuire and Miranda 2007; Gross et al. 2008; Martinez et al. 2008; Cook and Manning 2009). The emphasis of these studies is either global (all medical use) or on particular conditions (i.e. treatments for cancers).

In this paper, I study a particular form of medical care that is influenced greatly by

technological innovation but has received less attention from the disparities literature: prescription drugs. Like much of the recent literature, I focus on expenditures because it can capture differences in the intensity of care. I argue that disparities in the use of prescribed medicines should be thought of as a two-stage process that mimics an individual's demand problem. In the first stage, an individual chooses how much to invest in health–which depends on her unique environment, financial situation, preferences and biology. This choice, in turn, determines which (if any) medical conditions a provider has diagnosed her with. In the second stage, the individual has already been diagnosed with a medical condition and must decide—in what is really a joint decision with her provider and payer—whether to purchase prescribed medications to treat the disease. Disparities in prescription drug utilization can therefore exist because of differences in the probability of being diagnosed with a disease or because, conditional on having a diagnosis, individuals use fewer prescription drugs.

I define disparities to be racial, ethnic, or educational differences in care that are not due to heterogeneous preferences or clinical need.[1] This definition recognizes that spending differences should adjust for differences in health and age. Disparities can, however, be caused by access related factors like income or health insurance status.

The empirical exercises begin by first documenting large disparities in prescription drug use. Prescription drug expenditures are modeled using a two-part model (e.g. Duan et al. 1983) that accounts for the large number nonspenders in expenditure data and the right skewed nature of nonzero values. On average, blacks and Hispanics are predicted to spend $350 and $560 less than whites respectively. Similarly, an additional 4 years of education is predicted to increase prescription drug spending by $155. Access related factors can only explain a small percentage of the racial/ethnic disparities ($\approx$ 3% of the black-white differential and 14% of the Hispanic-white differential), while a slightly larger percentage of the education gradient is due to differences in access ($\approx$ 34%).

I then replicate the two stage decision process by first modeling the probability of being

[1]This is consistent with the Institute of Medicine (IOM) definition of disparities used in related articles (e.g. Cook, McGuire and Miranda 2007).

diagnosed with a disease and then modeling expenditures conditional on diagnosis. The first stage results suggest large racial/ethnic and educational gradients in diagnosis rates. For instance, on average, the black-white and Hispanic-white percentage differences in the probability of being diagnosed with a medical condition are 6% and 11% respectively while an additional 4 years of education increases the probability of diagnosis by 4%. These disparities are also consistent across a range of disease groups.

The second stage results show that nearly all of the expenditure differentials remain even after conditioning on diagnoses. I estimate that only 10 to 30 percent of the overall disparities in both the probability of positive expenditures and total expenditures conditional on positive expenditures can be accounted for by disease groups. Furthermore, the racial/ethnic disparities are consistent across three drug classes—anti-cholesterol medications, anti-diabetics and antidepressants—although there is only an education gradient for antidepressants. All told, the results suggest that there are large racial/ethnic and educational disparities in prescription drug expenditures and that these disparities are caused by both differences in diagnosis rates and differences in utilization conditional on diagnosis.

2 Theoretical Framework and Empirical Estimation

The market for pharmaceutical products differs from the markets for other goods and services in a number of important ways. Chief among them are that prescription drugs are only demanded in response to medical need and that they are prescribed by providers—rather than consumers—who neither consume nor pay for them. This means that an individual will only use a prescription drug if three important steps are fulfilled: first, she must receive a diagnosis from a provider; second, the provider must write a prescription for that diagnosis; and third, the she must fill the prescription. These three steps can be thought of as occurring within the context of a two-stage decision process, where in the the first stage, the individual much decide how much to invest in her health—which determines whether she is diagnosed with a medical condition—and in the second stage, she must make a decision about whether

to purchase a prescription for her diagnosed condition. More formally, expected expenditures can be decomposed (approximately) as,

$$E(y) \approx \sum_d E(y|c_d)p(c_d), \tag{1}$$

where c_d and y denote medical condition d and expenditures on prescribed medications respectively. This relationship is only approximate because the drugs used to treat different diseases may be the same.[2].

To stay consistent with the consumer's two stage decision process, I model $p(c_d)$ and $E(y|c_d)$ separately. In the first stage, the probability of being diagnosed with a medical condition depends on an individual's propensity to seek care that allows for the detection of the disease, which can depend on either the demand for curative care or preventive care (or both). In some cases, the disease will cause enough discomfort to cause the individual to seek curative care to receive relief from symptoms. In other cases, the disease will remain undetected unless the patient uses preventive services such as a physical examination or a medical screening test.[3] The probability of being diagnosed with a disease is modeled with a simple logistic regression,

$$p(c_{idt} = 1) = \Lambda\left(z_{it}^T \gamma\right), \tag{2}$$

where $c_{idt} = 1$ is an indicator variable coded as 1 if individual i has condition d in year t and 0 otherwise, $\Lambda(\cdot)$ is the inverse logit function, and z_{it} is a column vector containing the primary variables of interest (black race, Hispanic ethnicity and years of education) as well as relevant control variables.

In the second stage, the consumer decides whether to purchase prescribed medications

[2]See, for instance, the drug hierarchies used in the CMS risk adjustment models (e.g. Robst, Levy and Ingber 2007)

[3]This type of prevention, which reduces the severity of disease without affecting the probability that the disease occurs is called secondary prevention. Activities that reduce the probability of an illness occurring, such as exercise, are known as primary prevention.

after having been diagnosed with a medical condition. Unlike in stage one, demand is only for curative care since the condition has already been diagnosed. Modeling medical expenditures is not straightforward due to the large number of observed zero's and the heavily right skewed distribution of the remaining values (see Figure C.1 in the appendix). To address these issues, I model expenditure in two parts. First, the individual decides whether to use prescription drugs, and then, conditional on usage, decides how much to spend. In addition to providing a better fit to the data, the two-part approach is useful conceptually because it captures both the extensive margin (whether individuals choose to consume a drug) and the intensive margin (how intensely individuals consume the drug). This distinction is important because the choice of whether to use a drug is predicated on prior beliefs about the efficacy of medications while the intensity of consumption depends on posterior beliefs obtained through experience.[4].

Formally, the model can be written as,

$$D_{it}^* = x_{1it}^T \alpha + \epsilon_{1it}, \tag{3}$$

$$\ln y_{it} | D_{it}^* > 0 = x_{2it}^T \beta + \epsilon_{2it}, \tag{4}$$

where $D_{it} \equiv I(D_{it}^* > 0) = I(y_{it} > 0)$ is a latent variable that describes whether expenditures are positive or zero. An important question is whether the error terms, ϵ_{1it} and ϵ_{2it}, are correlated. If they are correlated then a type 2 Tobit model is appropriate; otherwise, a two-part model can be used. This paper focuses on the two-part model because it has a more natural interpretation with medical data. As argued by Duan et al. (1983), a selection model estimates the amount that all individuals (including non-spenders) would have spent if they were spenders. In contrast, the second part of the two-part model describes the conditional mean of expenditures given that expenditures are positive, which is meaningful

[4]Prescription drugs are experience drugs because utility can only be ascertained upon consumption and continued use depends on satisfactory experiences. Indeed, Dranove (2009) notes that health care may be "the quintessential experience good."

since zero expenditures are observed data rather than missing data as in the selection model. That said, the appendix shows that both approaches generate similar results but that the two-part model fits the data better.

If ϵ_{1it} and ϵ_{2it} are independent, then the two equations are separable. The first equation can be modeled using a standard regression model for binary data—this paper uses a logistic regression. The second equation can be estimated using a simple linear regression model. Importantly, y_{it} is log transformed because the regression coefficients are more likely to be additive on the log scale and multiplicative on the raw scale.[5] As a result, the estimated gradients in prescription drug expenditures are easily interpreted in percentage terms. A disadvantage of the transformation is that it is more difficult to estimate the impact of the coefficients on expenditures in levels because predictions for y depend of the distribution of ϵ_{2it} when y_{it} is log transformed. Section 4 and Appendix A examine this issue in more detail.

3 Data

This paper uses data from the Medical Expenditure Panel Survey (MEPS) from 1996 to 2012 for individuals age 25 and older. Each survey is a nationally representative sample of the U.S. population drawn from the National Health Interview Survey (NHIS). The MEPS is the most reliable source of nationally representative medical expenditure data available and is used to construct data for the National Health Accounts. The dataset is a rolling two-year panel with approximately 15,000 new respondents interviewed each year. Respondents are followed for two years and interviews are conducted three times each year.[6] I combine data from the Full-Year Consolidated Files, Prescribed Medicine Files and Medical Conditions Files to obtain detailed information on individual characteristics, medical conditions and prescription drug expenditures. The pharmaceutical data is especially reliable because the

[5] Another common approach is to model the second component using a generalized linear model (GLM). See Appendix A for a brief discussion.

[6] Respondents are actually interviewed five times over the two years but the third round of the survey overlaps two calendar years and is split into two distinct periods.

data is primarily based on pharmacy records rather than self reports.[7]

3.1 Variables

Summary statistics for the main variables are shown in Table 1. Since the MEPS does not use simple random sampling, the table compares summary statistics from the sample with population estimates using the survey weights provided by the MEPS. The differences in the sample and population means tend to be small, although there are significant differences in the distributions of the race, income and education variables because the MEPS oversamples certain "policy relevant" sub-groups such as racial minorities and low-income groups.

Table 1: Summary Statistics

	Sample Statistics				Population Estimates	
	Mean	SD	Min	Max	Mean	Median
Rx expenditures	959.412	4861.699	0	2,273,083	1005.175	155
Purchased a prescription	0.664	0.472	0	1	0.701	1
Rx medical condition	0.644	0.479	0	1	0.670	1
White	0.559	0.497	0	1	0.716	1
Other race	0.064	0.245	0	1	0.057	0
Black	0.156	0.363	0	1	0.110	0
Hispanic	0.221	0.415	0	1	0.118	0
Years of education	12.467	3.323	0	17	13.095	13
Health status excellent	0.221	0.415	0	1	0.240	0
Heath status very good	0.313	0.464	0	1	0.333	0
Health status good	0.302	0.459	0	1	0.286	0
Health status fair	0.120	0.325	0	1	0.103	0
Health status poor	0.044	0.204	0	1	0.038	0
Poor	0.150	0.357	0	1	0.101	0
Near poor	0.054	0.227	0	1	0.040	0
Low income	0.155	0.362	0	1	0.130	0
Middle income	0.306	0.461	0	1	0.310	0
High income	0.335	0.472	0	1	0.419	0
Insured	0.825	0.380	0	1	0.873	1
Age	48.751	15.854	25	90	49.351	47
Female	0.541	0.498	0	1	0.522	1

Notes: The sample consists of 330,580 individuals sampled by the MEPS from 1996 - 2012. Income categories are calculated by dividing family income by the applicable poverty line; the five categories are poor (less than 100%), near poor (100% to less than 125%), low income (125% to less than 200%), middle income (200% to less than 400$) and high income (greater than 400%). The population estimates are calculated using the survey weights provided by the MEPS. Rx expenditures are in 2012 dollars.

The primary dependent variable is prescription drug expenditures, which is normalized

[7] If individuals refuse to release their pharmacy records, then expenditures are based on self-reports that are adjusted for outliers and item non-response based on imputations from the pharmacy data.

to 2012 dollars using the CPI. As shown in the table, expenditures are heavily right skewed with a mean that is considerably higher than the median (also see Figure C.2). The two-part model seems well suited for modeling pharmaceutical expenditures since expenditures are zero for 34% of the observations. Maximum expenditures are over $2 million, which is more than 5 times as large as the next largest value. Although including this observation has little effect on the primary variables of interest it grossly inflates prediction errors so its is dropped from the analysis.[8]

The primary independent variables are the racial/ethnic variables and the years of education variable. The years of education variable ranges from 0 to 17; the average individual has obtained 12 years of schooling, which is approximately equivalent to a high school diploma. The race/ethnicity variables contain four mutually exclusive indicator variables for non-Hispanic whites, non-Hispanic blacks, Hispanics, and an "other" race group.

Disease groups included in the model are those that have proven to be the most important predictors of prescription drug expenditures. Specifically, I use the prescription drug condition categories used by the Center for Medicare & Medicaid Services (CMS) in its risk adjustment models for Medicare Part D, the prescription drug portion of Medicare (see Robst, Levy and Ingber 2007).[9] Each condition category consists of ICD-9 codes that are clinically homogeneous, similarly expensive to treat, and treated with similar drugs. I use the MEPS medical conditions files to create indicator variables for 70 condition categories based on the CMS model, whose construction is described in more detail in Appendix B).[10] Not surprisingly, the proportion of individuals with at least one Rx medical condition, 0.64, is nearly identical to the proportion of individuals who have purchased at least one prescription, 0.66.

[8]The RMSE in the k-fold cross-validation reported in Appendix A increases by a factor of nearly 4 when the observation is included in the validation set.

[9]Disease categories are only included in the final CMS model if they are deemed to be important predictors of prescription drug expenditures.

[10]Unlike the expenditure data, the medical conditions are based solely on self-reports and may therefore be measured with error due to misreporting. However, to the extent that purchasing prescribed medicines depends on knowledge of one's own medical conditions rather than actual diagnoses from providers, using self-reported medical conditions may be superior to actual diagnoses.

To get a sense of whether the condition categories are important predictors of pharmaceutical utilization, Table 2 shows three measures of utilization by whether an individual has at least one condition. As one would expect, those without a condition use far fewer prescribed medications than those with a condition. Those without a condition, do however, use some medications, which could be due to reporting error since the medical conditions are self-reported or because there are some ICD-9 codes that are not included in the 70 disease groups that sometimes require a small amount of medication.

Table 2: Prescription Drug Utilization by Disease Status

	Rx condition	
	No	Yes
Proportion using prescibed medications	0.287	0.873
Number of prescriptions filled	1.161	18.954
Rx expenditures	64.841	1454.869

The remaining variables were chosen to be consistent with my definition of disparities. Variables used to adjust for clinical need are dummy variables for self-reported health status (five categories from poor to excellent) and single year of age. Additional covariates are included to control for geographic region (west, midwest, south and northeast), marital status and calendar year. The clinical need variables and additional covariates are referred to as the basic controls. Access variables include dummy variables for income category (five categories from poor to high income) and insurance status (equal to 1 if insured).[11] I use a variable for health insurance rather than prescription drug insurance because the variables needed to construct a measure of prescription drug insurance contain a substantial number of missing observations.

The same basic controls and access variables are included in all models so in general, $z_{it} = x_{1it} = x_{2it}$. The one exception is when the design matrix for the expenditure regressions includes the full set of Rx disease group dummies. In practice, the Heckman 2-step estimator

[11]Income categories are calculated by dividing family income by the applicable poverty line; the five categories are poor (less than 100%), near poor (100% to less than 125%), low income (125% to less than 200%), middle income (200% to less than 400$) and high income (greater than 400%)

is problematic because there are no obvious variables that should be included in x_{2it} but not x_{1it}. Identification in the second stage regression comes solely from the nonlinearity of the inverse mills ratio which is approximately linear over a large range of its arguments (Puhani 2000).

3.2 Unadjusted Differences in Expenditures

Figure 1 shows prescription drug expenditures and the proportion of individuals who purchased at least one medication by year and race/ethnicity from 1996 to 2012. Although there are clear time trends in both expenditures (upward) and the proportion of individuals using (slightly downward), the racial gaps have remained fairly constant over time: whites have consistently used more prescription drugs than blacks and Hispanics.

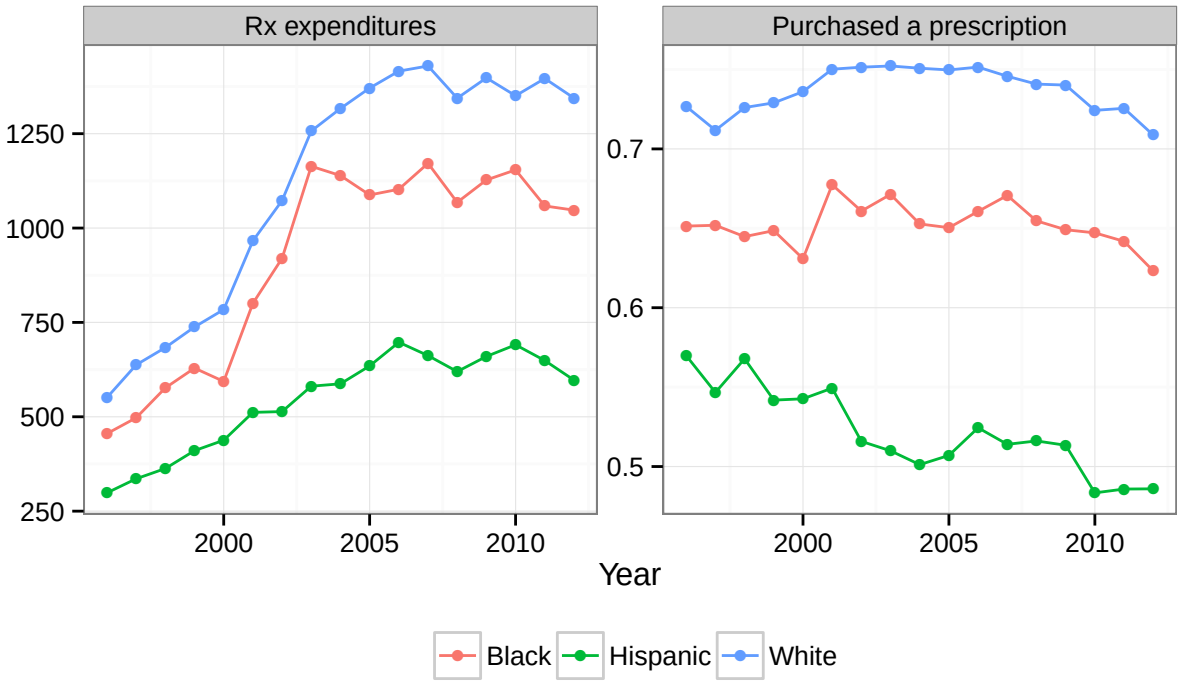

Figure 1: Prescription Drug Use By Year and Race/Ethnicity

Notes: Each point corresponds to mean drug use by race and year. The "other" race group is omitted.

Figure 2 repeats this figure for years of education. The right panel is consistent with the

race/ethnicity disparities as the less educated are less likely to have nonzero expenditures than the more educated. However, the opposite is true in the leftmost panel: there is actually a negative association between prescription drug spending and education. This finding is not entirely unexpected because better educated individuals tend to be in better health which should decrease demand for prescription drugs. Indeed, as shown in the next section, this result reverses after controlling for self-reported health status.

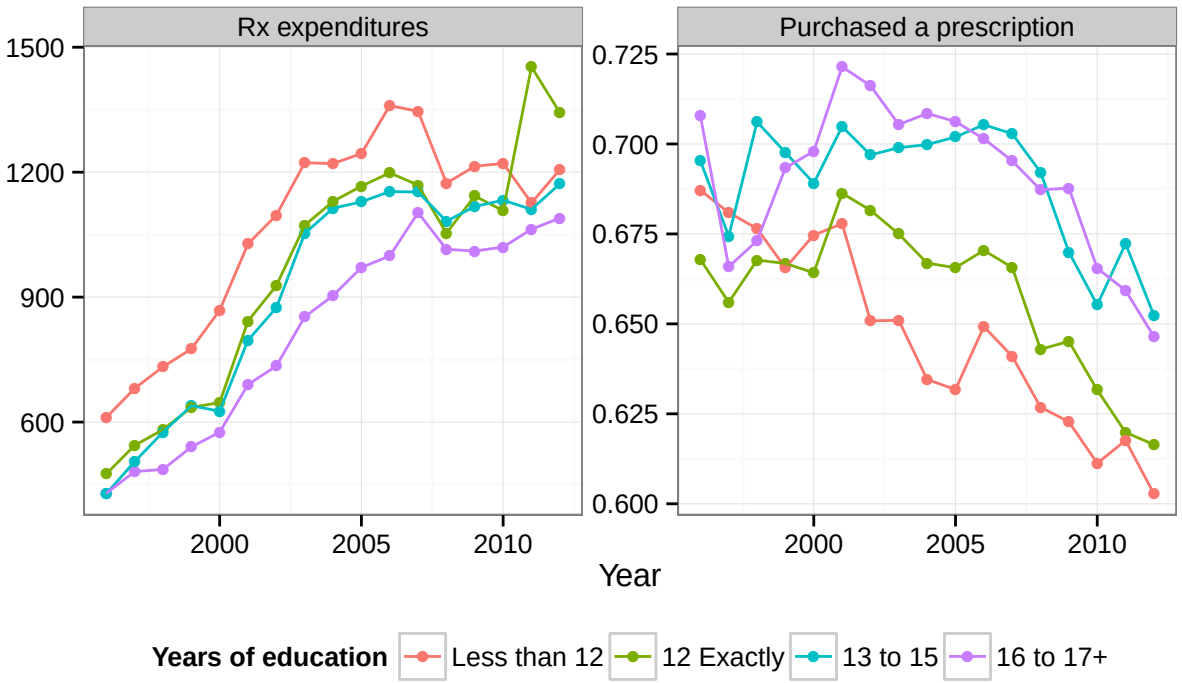

Figure 2: Prescription Drug Use By Year and Education

Notes: Each point corresponds to mean drug use by education category and year.

Table 3 reports unconditional results from the first stage of the decision process. In particular, it provides summary data on the Rx condition categories by race/ethnicity and education. Similar to the patterns in the expenditure data, there are clear racial and ethnic gradients in diagnoses but there is no evidence of an education gradient. Whites, on average, have more total medical conditions and are the most likely to have been diagnosed with at least one condition, while the Rx conditions are the least prevalent among Hispanics. There

is, if anything, a negative association between years of education and diagnoses, although the pattern is not entirely linear. The racial and ethnic differentials are somewhat surprising since one would expect racial and ethnic minorities—who have been documented in study after study to be in worse health—to have more medical conditions. As with the expenditure data, Section 4 shows that apparent negative relationship between education and the Rx conditions reverses after controlling for health status, which ultimately suggests that there is in fact an education gradient in diagnoses.

Table 3: Rx Conditions by Race/Ethnicity and Education

	Proportion with an Rx condition	Mean number of Rx conditions
Race/Ethnicity		
White	0.704	1.946
Hispanic	0.512	1.200
Black	0.643	1.701
Other	0.572	1.378
Years of education		
Less than 12	0.651	1.901
12 Exactly	0.643	1.712
13 to 15	0.656	1.720
16 to 17+	0.639	1.517

4 Results

This section reports disparities in prescription drug spending after adjusting for covariates. It begins by providing regression adjusted estimates of disparities without conditioning on disease groups and then examines the role that diagnoses play in those estimates. More precisely, it first provides an estimate of overall disparities and then breaks the disparities down into the first and second stages of the consumer decision problem. All analyses are based on the $314,462$ observations with complete data for both the basic controls and access related variables.

Regression estimates of overall disparities in expenditures are shown in Table 4. Estimates are based on the two-part model described in Section 2. The first three columns report

estimates from logistic regressions estimating the probability of positive expenditures. The final three columns report regression coefficients (and robust standard errors) from OLS regressions predicting the log of expenditures given that expenditures are positive. The table includes models with basic controls and models that add the access variables. The percent reduction in the coefficients after adding the access variables are computed for convenience.

Table 4: Regression Estimates of Disparities in Utilization

	$p(expenditures > 0)$			$\ln(expenditures) \mid expenditures > 0$		
	Basic controls	Adding access controls		Basic controls	Adding access controls	
	Coefficient	Coefficient	Reduction in coefficient	Coefficient	Coefficient	Reduction in coefficient
Black	−0.398	−0.372	7%	−0.310	-0.308	1%
	(0.012)	(0.013)		(0.010)	(0.010)	
Hispanic	−0.715	−0.543	24%	−0.528	-0.466	12%
	(0.011)	(0.012)		(0.011)	(0.011)	
Years of education	0.070	0.042	39%	0.026	0.018	30%
	(0.001)	(0.002)		(0.001)	(0.001)	
Mean of dv	0.670	0.670		1431.97	1431.97	
Observations	314,462	314,462		210,746	210,746	

Notes: The 1st-3rd and 4th-6th columns report results from logistic and OLS regressions respectively. Standard errors are in parentheses. Basic controls are indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year. Additional access controls are indicator variables for five income categories and health insurance status.

The table provides the means of the dependent variables to help with the interpretation of the coefficients. The marginal effect of variable j in the logistic regression is approximately $\alpha_j \Lambda(x_{1it}^T \alpha)[1 - \Lambda(x_{1it}^T \alpha)] < \alpha_j/4$ since the inverse logit function is steepest at $\Lambda(0) = 1/2$. At the sample average, the impact of a coefficient is approximately $\alpha_j(0.67)(1 - 0.67) \approx 0.22\alpha_j$. For indicator variables, it is preferable to calculate an incremental effect—the change in the probability of positive expenditures when a variable is changed from 0 to 1— but the marginal effects are still a reasonable approximation.[12] For instance, at the sample average, the marginal effect of being Hispanic on the probability of positive expenditures in the first column is −0.158 and the incremental effect is −0.172. Both estimates suggest large effects:

[12]Mathematically, the incremental effect for variable j is $\Lambda(x_{1it,-j}^T \alpha_{-j} + \alpha_j) - \Lambda(x_{1it,-j}^T \alpha_{-j})$ where $-j$ refers to all variables except variable j.

the model predicts that an Hispanic individual would have around a 52% chance of using prescription drugs when, all else equal, a white individual would have a 67% probability of usage.

The estimated disparities are consistent across both parts of the two-part model: blacks, Hispanics and the less educated are less likely to use prescribed medications and spend less on them when they do use them. The largest disparities are between Hispanics and whites in both components of the two-part model. The second component predicts that Hispanics spend $100 \cdot |\exp(-0.53) - 1| \approx 41\%$ less than whites without accounting for differences in access to care. Disparities for blacks and the less educated are significant as well but not quite as large. For example, when the model includes basic controls, an additional 4 years of education is predicted to increase the probability of using prescription drugs by 6% at the sample average and to increase spending conditional on nonzero spending by approximately 10%. The education gradient in nonzero expenditures is consistent with previous work which has shown that, unadjusted, those with less education spend more on medical care than more educated individuals but that this result reverses after controlling for health status (Bhattacharya and Lakdawalla 2006).[13]

Table C.1 in Appendix C looks at whether these results change if ϵ_1 and ϵ_2 are allowed to be correlated. Regression coefficients are reported for three different estimators: the two-part model with a probit regression in the first component, a Heckman 2-step selection model and a maximum likelihood type 2 Tobit model.[14][15] The coefficients between all methods are similar and none of the results are qualitatively different.

To assess the full two-part model, I examine the impact of race/ethnicity and education on predicted expenditures. Since the second part of the model is in logs, expected expenditures are $E(y|x) = \Lambda(x_1\alpha)\exp(x_2\beta)E(\exp(\epsilon_2)|x)$. The third term depends on both the distribution

[13] Bhattacharya and Lakdawalla (2006) control for insurance status as well.

[14] The Heckman 2-step and maximum likelihood models are often referred to as limited information maximum likelihood (LIML) and full information maximum likelihood (FIML) methods respectively.

[15] A probit specification is used in the first component of the two-part model to maintain consistency with the selection models.

of the errors and whether the errors are heteroscedastic. Unfortunately, neither a normality assumption or a constant variance assumption is innocuous. I consequently use a Duan (1983) smearing factor that varies by age to estimate the distribution of the errors.[16] Appendix A shows that predictions with this smearing factor are much more accurate than predictions that assume that the errors are normally distributed or assume a constant smearing factor.

Figure 3 uses the smearing factor to calculate predicted expenditures and average marginal/incremental effects (AME's/AIE's) for the three primary variables of interest in models with basic controls. Panel (a) shows how varying an "average" individual's education or race/ethnicity would change his or her predicted expenditures. This "average" individual is initially assumed to be white with 12 years of schooling. Given these baseline characteristics, the probability of using prescription drugs, expenditures conditional on use, and the smearing factor are then set to sample averages.[17] A white individual with 12 years of education is consequently predicted to have expenditures equal to average expenditures in the estimation sample, or $960.

The figure shows that the racial/ethnic and educational gradients in utilization for this "average" individual are economically significant. A white individual with a college education (16 years of schooling) is predicted to spend $198 more than if he or she had a high school education (12 years of schooling). The effects are even larger by race/ethnicity: blacks and Hispanics with 12 years of schooling are predicted to spend spend $210 and $428 less than their white counterparts respectively. It is worth noting that the education gradient is slighter steeper for whites than for blacks and Hispanics because the covariates have multiplicative effects on non-transformed expenditures and predicted expenditures are higher for whites.

Panel (b) calculates AME's for education and AIE's for race/ethnicity. Calculations are made with both the basic controls and after adding the access variables. The AME is the marginal impact of education on (raw) expenditures averaged across all individuals in the

[16] Duan's smearing factor is $\hat{s} = \frac{1}{N}\exp(\ln y - x_2\hat{\beta})$ where N is the number of observations in the data.

[17] More specifically, I set $\Lambda(x_1\alpha) = \overline{D}$, $exp(x_2\beta)s = (\overline{y}|y > 0)$ and solve for $x_1\alpha$ and $x_2\beta$ when the smearing factor, s, is equal to its mean. Predictions are then made by changing $x_1\alpha$ and $x_2\beta$ by varying race, ethnicity and education.

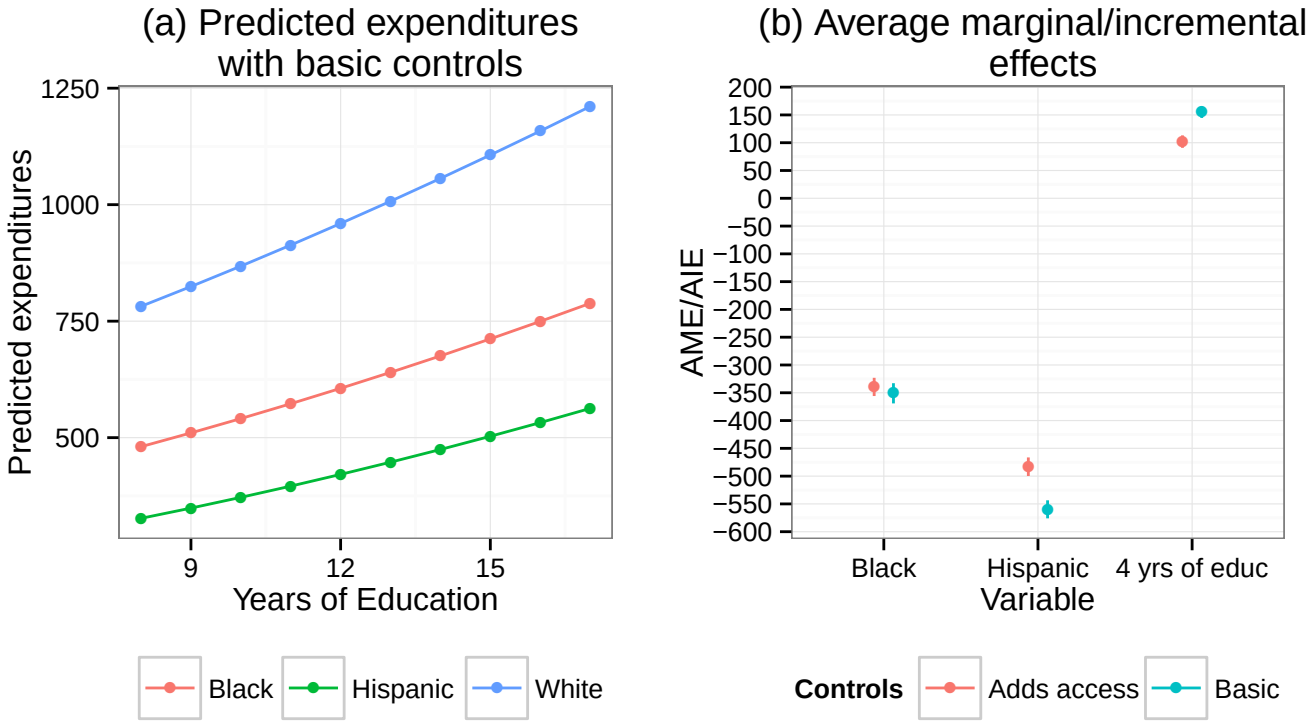

Figure 3: Predicted Disparities in Expenditures

Notes: (a) plots mean predicted expenditures by education and race/ethnicity for an "average" individual. (b) plots average marginal effects and bootstrapped 95% confidence intervals for each of the primary variables of interest. The AME for education is the original estimate multiplied by 4. Predicted expenditures and AME/AIEs are calculated using the two-part model and a Duan Smearing estimator that varies by single-year of age.

data. An AIE is the effect of switching an indicator variable from 0 to 1 on expenditures, again averaged over survey respondents. Both calculations are made using a smearing factor that varies by age. Standard errors are calculated using a 1,100 boostrap replications in order to account for uncertainty in both the coefficients and the transformation of expenditures from the log to the raw scale.

The AME/AIE's are consistent with the predicted expenditures in panel (a). The coefficients are estimated very precisely since the sample size is so large. Blacks and Hispanics are predicted to spend, on average, $350 and $560 less than whites. An additional 4 years of education has a significant impact as well as it is predicted to increase expenditures by $155.

Only a relatively small fraction of the racial/ethnic disparities in expenditures—3% of the black-white differences and 14% of Hispanic-white differences—are associated with access to care. Access factors explain a larger percentage of the education gradient ($\approx 34\%$) but a substantial fraction of the gradient still remains unaccounted for.

4.1 Disparities in Diagnoses

The outcome of a consumer's first stage decision problem is whether she has been diagnosed with medical condition. Table 5 provides regression estimates of (1) the probability of being diagnosed with at least one of the CMS disease groups and (2) the number of disease groups that an individual is diagnosed with. The binary variable is modeled with a logistic regression and the count variable is modeled using a negative binomial regression.[18]

The table shows that the racial and ethnic disparities from the raw data remain after controlling for covariates, but that an education gradient emerges as well. There are racial and educational gradients in both the probability of having a condition and the number of total conditions. The largest disparities are again between whites and Hispanics and the smallest are between education levels.

To interpret the coefficients, note that the marginal effect of covariate j in the negative binomial regression is $\kappa_j \exp(z_{it}^T \kappa)$ and the incremental effect of a one unit change in a variable is $\exp(z_{it}^T \kappa)[\exp(\kappa_j) - 1]$ where κ is the vector of covariates. The economic importance of the covariates can then be quickly assessed by setting $\exp(z_{it}^T \kappa)$ and $\Lambda(z_{it}^T \gamma)$ equal to sample averages of the dependent variables in the negative binomial and logistic regressions respectively and calculating incremental and marginal effects for the appropriate model. For instance, at sample averages (using the basic controls), the probability that an Hispanic individual is diagnosed with a medical condition is 15% lower than the probability that a similar white individual would have been diagnosed. Likewise, the negative binomial regression predicts that an Hispanic individual is diagnosed with -0.51 fewer medical conditions that a white

[18] A regression based test of overdispersion, as suggested by Cameron and Trivedi (1990), indicated the presence of significant overdispersion.

Table 5: Regression Estimates of Disparities in Diagnoses

	p(# of conditions > 0)			# of conditions		
	Basic controls	Adding access controls		Basic controls	Adding access controls	
	Coefficient	Coefficient	Reduction in coefficient	Coefficient	Coefficient	Reduction in coefficient
Black	−0.357	−0.343	4%	−0.219	-0.221	-1%
	(0.013)	(0.013)		(0.005)	(0.005)	
Hispanic	−0.613	−0.494	19%	−0.351	-0.295	16%
	(0.012)	(0.012)		(0.005)	(0.005)	
Years of education	0.055	0.038	32%	0.018	0.014	23%
	(0.001)	(0.002)		(0.001)	(0.001)	
Mean of dv	0.647	0.647		1.713	1.713	
Observations	314,462	314,462		314,462	314,462	

Notes: The 1st-3rd and 4th-6th columns report results from logistic and negative binomial regressions respectively. Standard errors are in parentheses. Basic controls are indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year. Additional access controls are indicator variables for five income categories and insurance status.

individual. The education gradient is smaller but not insignificant: an additional 4 years of education is associated with a 5% increase in the probability of having a medical condition and a 0.12 more medical conditions. AME's and AIE's are very similar to the effects at sample averages: averaged over the population, the black-white and Hispanic-white differences in the probability of having a medical condition are 6% and 11% respectively, and 4 additional years of education raises the probability of having a medical condition by 4%.

The effects of the access variables on the gradients in diagnoses are consistent with their effects on the gradients in expenditures. The access variables cannot account for almost any of the differences in diagnoses between blacks and whites but can account for a small fraction of the differences between Hispanics and whites. A larger fraction of the education gradient is associated with access factors, but most the education gradient remains unexplained even after controlling for income and health insurance status.

I next turn to more disaggregated evidence to examine whether there are specific disease groups that are driving these results. Table 6 lists the 20 most common diseases and the proportion with each disease by race/ethnicity. Whites are more likely to have been diagnosed with nearly all conditions than blacks and Hispanics. There are however some

important exceptions such as diabetes and hypertension (for blacks but not Hispanics). Not surprisingly, the most commonly dispensed prescriptions in the United States as reported by IMS Health—which include antihypertensives, antidepressants, lipid regulators and antidiabetics—are used to treat the most prevalent disease groups listed in Table 6.

Table 6: Prevalence of Most Common Rx Disease Groups

	Proportion with disease			
Disease group	All	White	Black	Hispanic
Other Musculoskeletal and Connective Tissue Disorders	0.283	0.317	0.283	0.211
Hypertension	0.238	0.241	0.338	0.168
Disorders of Lipid Metabolism	0.146	0.167	0.134	0.100
Diabetes without Complication	0.096	0.082	0.137	0.102
Mild Depression	0.086	0.105	0.063	0.066
Allergic Rhinitis	0.077	0.089	0.056	0.059
Asthma and COPD	0.060	0.067	0.066	0.040
Thyroid Disorders	0.059	0.078	0.032	0.037
Anxiety Disorders	0.056	0.072	0.037	0.036
Headache (Including Migraine)	0.055	0.058	0.049	0.052
Esophagael Disorders	0.051	0.066	0.051	0.021
Coronary Artery Disease	0.035	0.042	0.035	0.020
Acute Bronchitis and Congenital Lung/Respiratory Anomaly	0.034	0.045	0.022	0.018
Disorders of the Vertebrae and Spinal Discs	0.029	0.039	0.023	0.013
Bone/Joint/Muscle Infections	0.028	0.037	0.013	0.017
Other Diseases of Upper Respiratory Tract	0.023	0.025	0.023	0.020
Peptic Ulcer and Gastrointestinal Hemorrhage	0.022	0.021	0.021	0.031
Urinary Obstruction and Retention	0.022	0.025	0.019	0.017
Menopausal Disorders	0.021	0.027	0.013	0.011
Other Cancers and Tumors	0.021	0.026	0.021	0.010

To see whether these disparities hold up after regression adjustment, I use logistic regression models to predict the probability of having each of the 70 Rx condition categories. Models are estimated using both the basic controls and after adding the access variables. Regression coefficients and 95 percent confidence intervals for the 20 most common diseases (those listed in Table 6) are shown in Figure 4. The coefficients on the black and Hispanic indicator variables are consistently negative and statistically distinguishable from 0. As with the unadjusted data, the two primary exceptions are diabetes and hypertension. The effects on each disease are relatively small, as the coefficients tend to be less than 0.05. Nonetheless, when summed across multiple diseases these small differences create significant disparities

in the aggregate probability of being diagnosed with a disease.

The coefficient on the education variable is multiplied by 4 so that its economic importance can be more reasonably compared with the race and ethnicity variables. The coefficient tends to be small and in most cases (although still positive) close to 0. This is consistent with the results shown in columns 1 and 2 of Table 5, which showed that additional years of education were associated with small, but positive, increases in the probability of having at least one medical condition.

As expected, the access variables tend to push the estimated coefficients toward zero. The effects are small although they are once again more pronounced for the education and Hispanic ethnicity variables than for the black race variable.

4.2 Disparities in Conditional Expenditures

The second stage of the consumer decision problem occurs after an individual is diagnosed with a medical condition. To examine whether there are disparities at this stage I add dummy variables for each of the 70 disease groups to the two-part expenditure model. The idea is to see if disparities remain even after conditioning on the disease groups that are the best predictors of prescription drug expenditures. The results are reported in Table 7.

To facilitate comparison, I re-report the regression coefficients from Table 4, which are from models that did not condition on the disease groups. The table displays the reduction in two gradients—the full gradient (i.e. models with basic controls) and the gradient unexplained by access (i.e. models with access controls)—after controlling for diagnoses. The size of the coefficients after controlling for the disease groups are measures of gradients in the consumer's second stage decision problem.

The coefficients in both models with and without access controls tend to decline by around 10 to 30 percent. The decline is due to the positive correlation between diagnoses and spending and the negative correlation between diagnoses and disadvantaged groups. Even so, the coefficients remain highly statistically and economically significant. To illustrate, in

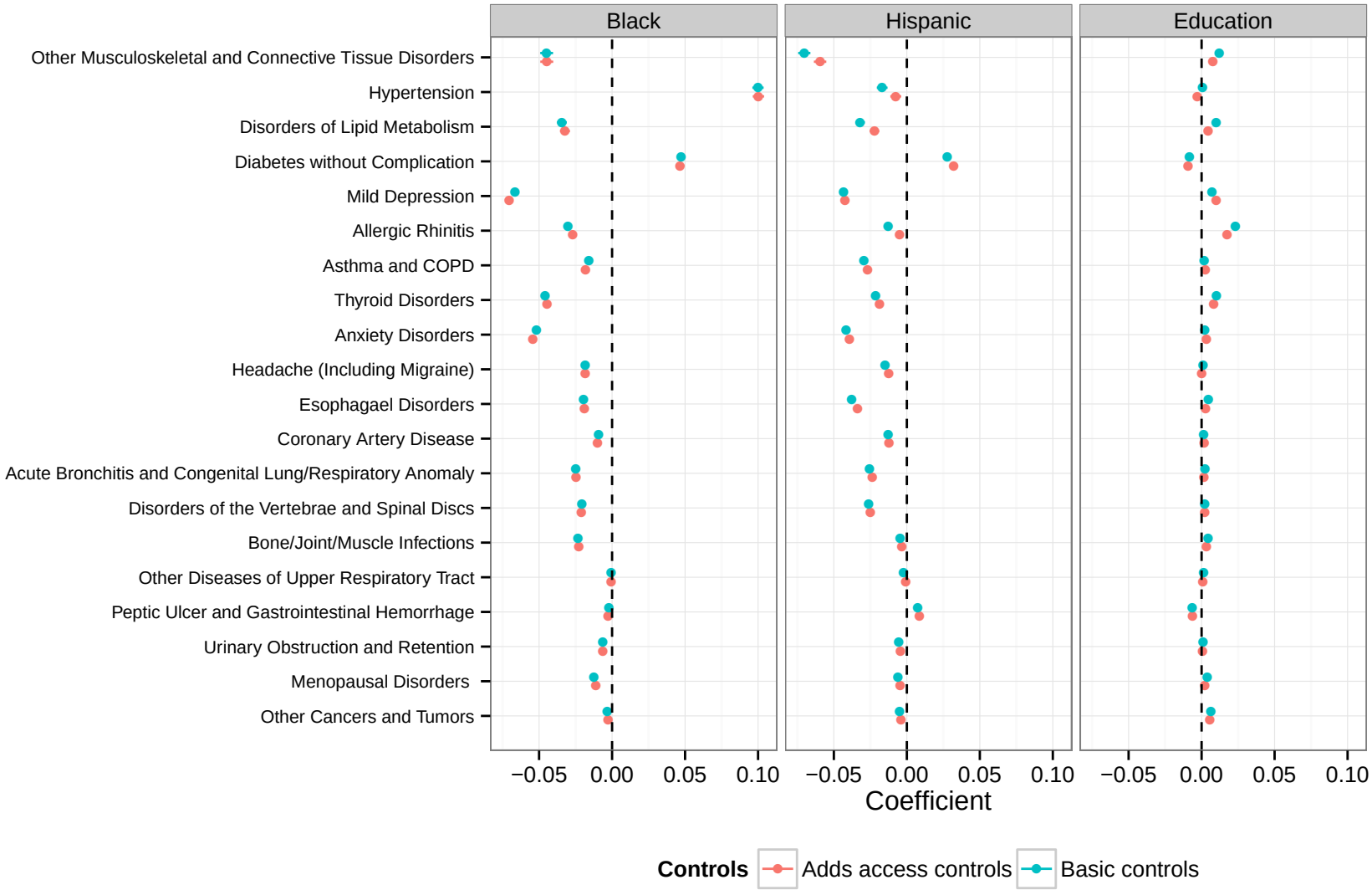

Figure 4: Regression Estimates of Disparities in Specific Diagnoses

Notes: The figure plots coefficients and 95% confidence intervals from logistic regression models. Coefficients are for the 20 most common disease groups. The reported coefficient for the education variable is the original estimate multiplied by 4. Basic controls are indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year. Additional access controls are indicator variables for five income categories and insurance status.

Table 7: Regression Estimates of Disparities in Utilization Conditional on Rx Condition Categories

	Basic controls			Adding access controls		
	No Rx disease groups	Adds Rx disease groups		No Rx disease groups	Adds Rx disease groups	
	Coefficient	Coefficient	Reduction in coefficient	Coefficient	Coefficient	Reduction in coefficient
Panel A.	$p(expenditures > 0)$					
Black	−0.398	−0.333	16%	−0.372	−0.308	17%
	(0.012)	(0.015)		(0.013)	(0.015)	
Hispanic	−0.715	−0.520	27%	−0.543	−0.378	30%
	(0.011)	(0.014)		(0.012)	(0.014)	
Years of education	0.070	0.056	20%	0.042	0.031	27%
	(0.001)	(0.002)		(0.002)	(0.002)	
Mean of dv	0.670	0.670		0.670	0.670	
Observations	314,462	314,462		314,462	314,462	
Panel B.	$ln(expenditures)\|expenditures > 0$					
Black	−0.310	−0.241	22%	−0.308	−0.228	26%
	(0.010)	(0.009)		(0.010)	(0.009)	
Hispanic	−0.528	−0.378	28%	−0.466	−0.331	29%
	(0.011)	(0.009)		(0.011)	(0.009)	
Years of education	0.026	0.023	12%	0.018	0.014	23%
	(0.001)	(0.001)		(0.001)	(0.001)	
Mean of dv	1431.97	1431.97		1431.97	1431.97	
Observations	210,746	210,746		210,746	210,746	

Notes: Models in Panel A and Panel B are estimated using logistic regression and OLS respectively. The first three columns report regression results from models that include the basic controls (indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year). The final three columns report regression coefficients from models that add the access controls (indicator variables for five income quintiles and health insurance status) to the basic controls. Models with Rx disease groups include 70 dummy variables based on the CMS risk adjustment model for Medicare Part D. Standard errors are in parentheses.

models with basic controls (and when incremental effects are evaluated at sample averages), being Hispanic is still predicted to lower the probability of positive expenditures by 12% after conditioning on disease groups. Moreover, Hispanics are predicted to spend 31% less than whites in the second component of the two-part model even after conditioning on diagnoses. This shows that the gradients in prescription drug expenditures are due to disparities in the likelihood of receiving proper diagnoses for diseases *and* from disparities in the use of prescription drugs after diagnosis.

Next, I look at whether these "conditional" disparities exist for a few of the most common therapeutic drug classes: anti-diabetics, antidepressants and anti-cholesterol medication. For each therapeutic class, I limit the sample to those diagnosed with a relevant disease.[19] The results are displayed in Figure 5. For reference, the proportion of survey respondents in the anti-diabetic, antidepressant and anti-cholesterol samples with positive expenditures are 0.83, 0.6 and 0.78 respectively.

The racial and ethnic disparities for each drug class are generally consistent with those in Table 7: blacks and Hispanics are less likely to have purchased prescribed medications for each drug class (except antidiabetics) and spend less when they do. The education gradient, on the other hand, is not: education only has a consistent positive effect on expenditures for anti-depressant drugs. The coefficients move toward zero after accounting for access but the size of the movement is generally small.

Estimates of the gradients in expenditures from the full two-part model are shown in Figure 6. The figure shows mean predicted expenditures for an "average" individual using the same procedure used for panel (a) of Figure 3. Mean expenditures for each drug class are fairly large—$806, $382, $564 for the anti-diabetic, antidepressant and anti-cholesterol samples respectively. The slope of education line is only significantly positive for antidepressants but expenditures at each level of education are considerably lower for blacks and Hispanics

[19]The anti-diabetic sample consists of individuals with diabetes with or without complications; the antidepressant sample contains individuals with mild depression, episodic mood disorders, personality disorders, or anxiety disorders; and individuals in the anti-cholesterol sample are those with disorders of lipid metabolism.

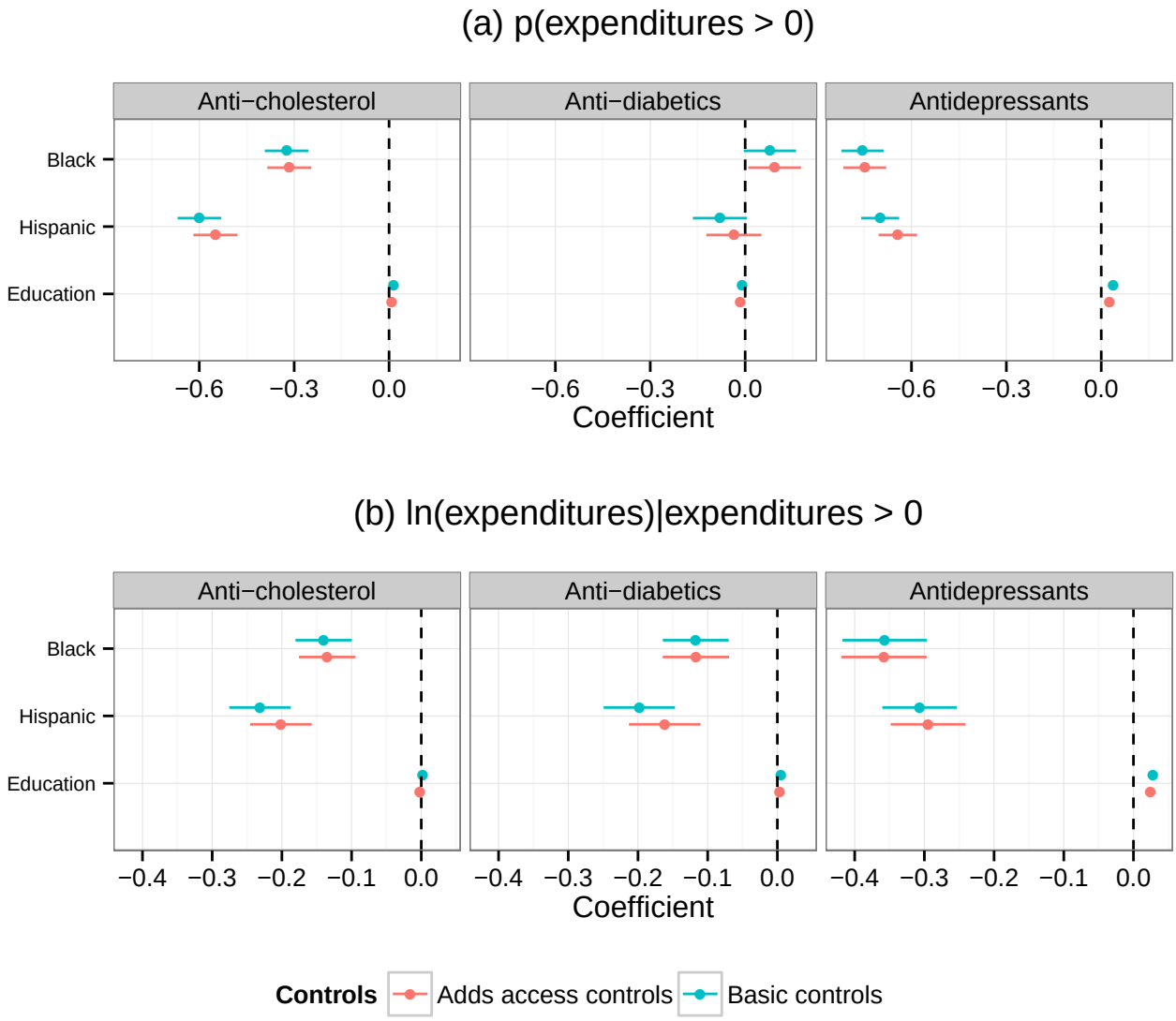

Figure 5: Regression Estimates of Disparities in Utilization for Common Drug Classes

Notes: Panel A and B are estimated with logistic and OLS regressions respectively. Basic controls are indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year. Access controls are indicator variables for five income categories and health insurance status.

than whites. Differences in expenditures between whites and racial minorities tend to be on the order of $100 to $200. An "average" white individuals diagnosed with depression with a college level of education (16 years of schooling) is predicted to spend $70 more on antidepressants than an "average" white individuals with a high school education (12 years

of schooling).

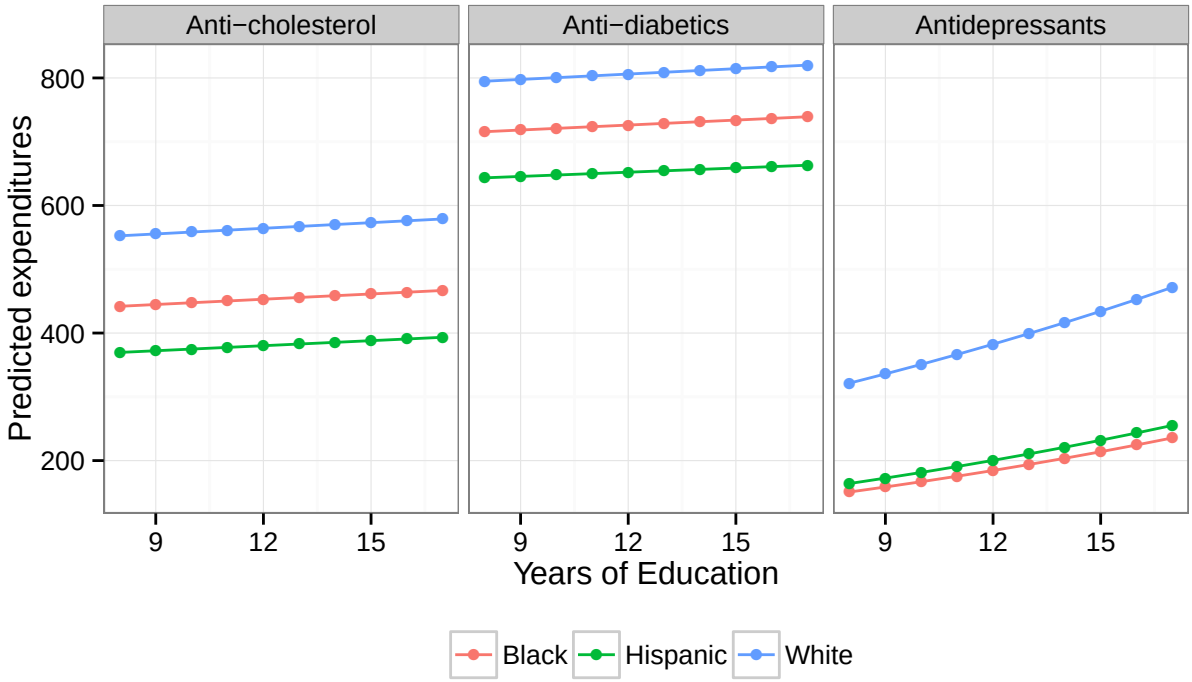

Figure 6: Predicted Mean Expenditures by Education, Race/Ethnicity and Drug Class

Notes: Predicted expenditures are for an "average" individual. Predictions for each drug class are calculated using the two-part model and a Duan smearing estimator that varies by single year of age.

5 Discussion

5.1 Models for Specific Drug Classes

One drawback of this study is that it examines prescription drug expenditures on fairly aggregate levels in order to document larger patterns in the data. As a result, it cannot control for all access-related variables (like distance from a provider) or clinical variables. One way forward is to focus on the utilization of drugs used to treat specific diseases and incorporate detailed clinical data if possible.

A more narrow focus would allow researchers to better understand the documented dis-

parities in diagnoses. It would be useful to separate the gradients in diagnoses into (1) gradients in actual health conditions and (2) gradients in diagnoses conditional on those health conditions. The second gradient is the true gradient in diagnoses and is likely underestimated in this paper because of well documented racial/ethnic and educational disparities in actual health conditions.

Analyses that focus on the treatment of specific diseases could also utilize discrete-choice models to analyze highly disaggregated consumption data. This approach would create a more uniform sample and allow researchers to detail heterogeneity in choices for drug types within a therapeutic class. For example, while this paper examines expenditures aggregated across diabetes drugs, a more disaggregated analysis could look at which specific drugs (biguanides, sulfonylureas, etc) individuals use to treat diabetes. Repeated purchases for individuals would also allow for the use of hierarchical models to estimate the importance of individual heterogeneity.

5.2 Impacts on Health

The importance of the disparities documented in this paper depend on their impact on health, which is uncertain, but should be significant for a number of reasons. First, the disparities might have been caused by ineffective adherence to treatment protocols, which would be consistent with evidence from Goldman and Smith (2002) showing that more educated HIV and diabetes patients are more likely to adhere to therapy and less likely to switch to treatments that worsen health.[20] Second, if there are disparities in the timing of diagnoses (in addition to those at a given point in time), then racial/ethnic minorities and the less educated might begin taking prescribed medications at later stages of the disease process. This would likely increase disease severity and increase mortality rates, especially since research (Thomas et al. 1997) has shown that there are racial disparities in the onset of disease. Third, disadvantaged groups might adopt new medications more slowly, which

[20]There were also some racial differences in adherence, but they were less pronounced than the educational differences.

is consistent with evidence showing that mortality rates for diseases treated with innovative technologies have declined more for the highly educated (Glied and Lleras-Muney 2008; Kinsey et al. 2008). There is also some evidence that these disparities are due to differences in learning (Lleras-Muney and Lichtenberg 2002) rather than knowledge, so educational differences could be more pronounced than racial or ethnic differences.

5.3 Policy Implications

Provided that one is confident that the disparities found in this paper are not due to omitted variable bias and are actually detrimental to health, it is worth considering policy interventions that could reduce disparities. Some important distinctions for policy are whether the differences in prescription drug expenditures are due to differences in understanding about the benefits of treatment or differences in preferences for the consumption of prescription drugs versus the consumption of other goods and services. If these differences are simply due to differences in preferences, then policy interventions are not needed; but if differences are due to differences in understanding, then policy interventions are warranted.

Current research suggests that differences in understanding may be the primary culprit. For instance, Whittle et al. (1997) has shown that race is not a significant predictor of the willingness to undergo revascularization procedures after controlling for familiarity with the procedures. A plausible way to improve familiarity is to improve communication between providers and patients, which pertains to another line of research suggesting that there are obstacles to effective communication between white physicians and black patients (Cooper-Patrick et al. 1999; Johnson et al. 2004). Another potential mechanism to improve understanding is by increasing health knowledge, which has been shown to be associated with disparities (Cutler and Lleras-Muney 2010).

These are not the only possible explanations though. For example, another reasonable source of the observed disparities is that racial minorities have had negative experiences with health care in the past. Mistrust in the medical system is particularly plausible for African

Americans, who were severely misreated in the infamous Tuskegee Syphilis Study, in which researchers studied the natural progression of untreated syphilis in poor and mainly illiterate black men, but did not tell participants that they had syphilis or treat them for it despite known effective treatments.

More research is needed on each of these mechanisms so that policymakers and health professionals can optimize their efforts to help reduce disparities. To ensure that the policymakers target the correct mechanisms, researchers should analyze how these mechanism vary with different medical treatments (i.e. prescription drugs vs surgery) and diseases (i.e. antidiabetics vs. anti-cholesterol medication). There is also a need for evaluations of specific interventions aimed at reducing racial and ethnic differences in understanding to determine (1) whether they are effective and (2) which interventions are most effective.

6 Conclusion

This paper provides empirical evidence showing that there are racial/ethnic and educational disparities in prescription drug expenditures. The documented disparities occur for two primary reasons: first, there are differences in the probability of being diagnosed with a disease; and second, there are gradients in expenditures conditional on diagnosis. The racial/ethnic disparities exist both in the raw data and after controlling for variables proxying for clinical need. The education gradients exist after controlling for health variables, but not unadjusted. Access related factors account for almost none of the black-white differences in expenditures ($\approx 3\%$), a small fraction of the Hispanic-white differences ($\approx 14\%$), and a larger fraction of the education gradient ($\approx 34\%$). The disparities—especially those between racial and ethnic groups—are consistent across disease groups and the most common drug classes. Overall, a two-part model predicts that, on average, blacks and Hispanics spend \$350 and \$560 less than whites respectively and that an additional 4 years of education increases prescription drug expenditures by \$155.

Appendices

A Predicting Positive Rx Expenditures

Health care expenditures in the second component of two-part model's are often log transformed. This is done for a number of reasons. First, predictions in non transformed models can be negative. Second, inferences are often sensitive to outlines. Third, the model is unlikely to be linear on the raw scale. Fourth, prediction intervals are more difficult to generate with heavily right skewed data.

Log transformed models are not without their own drawbacks though. The main problem is that predictions in levels are very sensitive to assumptions about the error term. If the error is not homoscedastic or normally distributed, then predictions can be very inaccurate. Table 8 illustrates this point. The table shows predicted mean expenditures and R^2 (measured as the square of the correlation between observed and predicted expenditures) for a number of different models. Each model includes the basic controls, the race/ethnicity indicator variables and the years of education variable.

Table 8: In-Sample Predictions of Nonzero Rx Expenditures

Model	Predicted mean expenditures	R^2
OLS on y	1431.97	0.09
OLS on ln(y) normal homoskedastic	2101.77	0.07
OLS on ln(y) Duan homoskedastic	1823.74	0.07
OLS on ln(y) Duan heteroskedastic	1559.78	0.08
Selection 2-Step on ln(y)	813.11	0.05
Selection MLE on ln(y)	2007.74	0.07

Notes: y refers to Rx expenditures. All models are estimated conditional on $y > 0$. R^2 is the squre of the correlation between y and $\hat{y}$ where $\hat{y}$ is predicted form the model. The selection 2-step assumes a normally distributed error in the second step. The selection MLE assumes a bivariate normal distribution for the error terms. All models include the basic control variables discussed in the text.

The reference model is a simple OLS model on positive expenditures in levels. Predicted mean expenditures in this model are by definition equal to observed mean expenditures. The next three models are based on log transformed regression models. The first model assumes that the error term is normally distributed with a constant variance, σ^2, so that $E(y|x_2, y >$

0) = $\exp(x_2\beta + \sigma^2/2)$. The normal approximation is not unreasonable (see Figure C.3), but predictions are still quite inaccurate. Mean expenditures are overpredicted by over \$600. Using the Duan smearing factor described in the text to allow for non-normality slightly improves predictions but mean expenditures are still overestimated. The largest improvement comes from a Duan smearing factor that varies by age: mean predictions only differ by around \$100 from observed mean expenditures and the R^2 is higher than the R^2 of any of the other models except for the reference OLS model. In-sample predictions can be made more accurate by allowing the variance to vary by additional groups, but this increases the risk of overfitting. Predictions from the selection models are made assuming that the errors are normally distributed and homoscedastic.[21]. Predictions from the selection model evaluated using maximum likelihood are similar to the predictions from the two-part model with normally distributed errors but the Heckman 2-step model is very inaccurate.

To ensure that the heteroscedastic Duan smearing estimator isn't overfitting the data, I used repeated 5-fold cross validation to test the out-of-sample performance of OLS models on the nontransformed data. 5-fold cross validation partitions the data into 5 folds. Each round of cross validation estimates the model on 4 folds (the training data) and evaluates its performance on the remaining fold (the validation data). Each of the 5 folds is used as validation data one time so there are 5 total rounds. This procedure was repeated 10 times which yields a total of 50 out-of-sample tests of the model.

The out-of-sample prediction results mirror the in-sample results. The Duan smearing estimator with variances that vary by age has the lowest root-mean-square-error (RMSE).[22] Moreover, predictions with a Duan smearing estimator with constant errors continue to outperform predictions from a model that assuming that the errors are normally distributed.

A number of recent studies have used generalized linear models (GLM's) to model health care expenditure data (e.g. Blough, Madden and Hornbrook 1999; Manning, Basu and Mul-

[21] See, for instance, the formulas for the conditional and unconditional means in chapter 16 from Cameron et al. (2009)

[22] The RMSE is $E\left[(y-\hat{y})^2\right]$ where y and $\hat{y}$ are observed and predicted expenditures respectively.

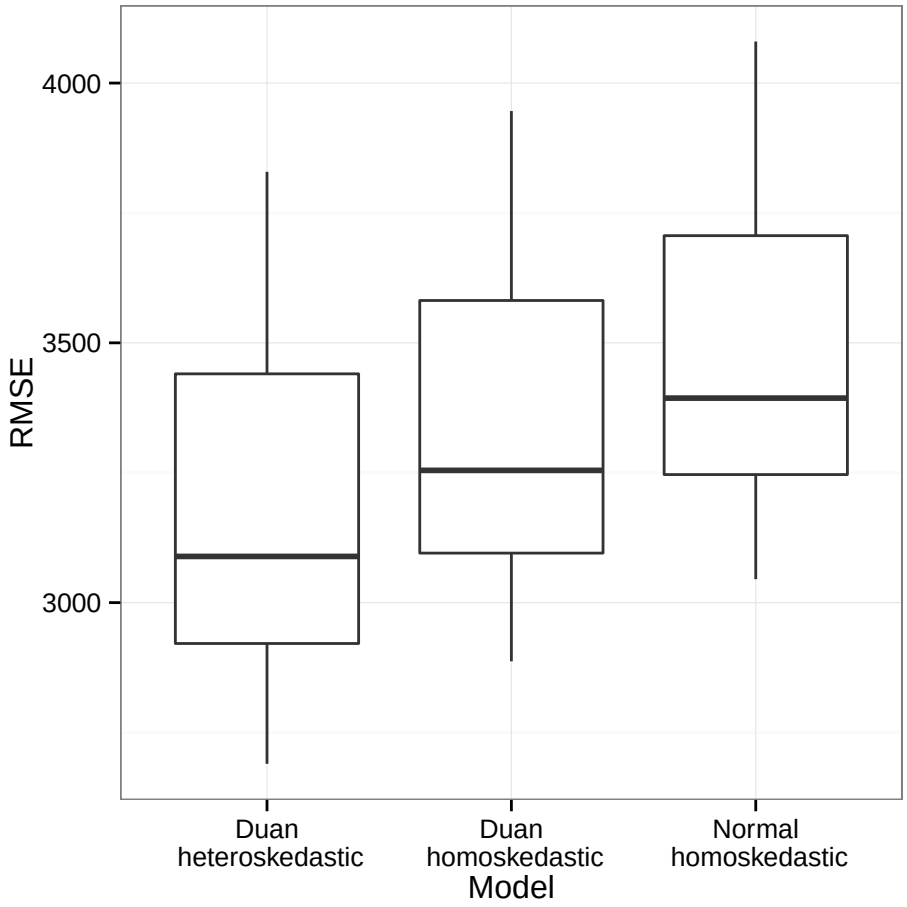

Figure 7: Results of Repeated 5-Fold Cross Validation

Notes: The figure is a box-and-whisker plot of the root-mean-square-error of 5-fold cross validation repeated 10 times. The RMSE is calculated for each model using 50 validation data-sets (5-fold cross validation and 10 repeats). All models include the basic control variables discussed in the text.

lahy 2005). A GLM with a gamma distribution and a log link function fit the data well when only basic controls were used. Predictions were generally similar to the OLS model with a heteroscedastic Duan smearing factor. The gamma GLM assumes that the variance is proportional to the mean squared which is reasonable in some model specifications. However, after controlling for the 70 Rx condition categories the gamma model was no longer suitable. The errors becomes increasingly complicated as more clinical variables are added to the model, which is clearly an area for future research.

B Constructing the RX Disease Categories

The disease categories in this paper are based on the disease categories used by the CMS to risk adjust health plans participating in Medicare Part D. Each disease group consists of clinically and financially related five-digit ICD-9 codes that are treated with similar drugs. The CMS model does not include diagnostic categories when "the diagnoses were vague/nonspecific, discretionary in medical treatment or coding, not significant predictors of drug use, or transitory or not admitting of definitive treatment." The disease groups in the final CMS model are called hierarchical condition categories (RxHCCs) because related disease groups are clustered into hierarchies and ranked by severity (if an individual has multiple conditions within a hierarchy, then only the most severe disease group is included in the CMS model).

The disease classifications are taken from the MEPS Medical Condition Files. Unfortunately, only 3-digit ICD-9 codes are available in the public use data set for confidentially reasons, so I am unable to recreate the RxHCCs exactly. I consequently created disease groups in the following manner. First, I considered including all RxHCCs from the 2010 (last year of the MEPS data used in this paper) and all RxHCCs from 2014 (most recent CMS model) that were not in the 2010 model. Next, I examined the 5-digit ICD-9 codes within each RxHCC and determined whether it was sensible to classify them based on 3-digit ICD-9 codes.[23] In some cases the RxHCCs match the MEPS clinical classification codes (CCCs), which aggregate 5-digit ICD-9 codes into clinically meaningful categories, quite closely. In these cases I use the CCCs to approximate the disease groups. For example, there is both a MEPS CCC and a RxHCC for congestive heart failure. The advantage of this approach is the CCCs are based on 5-digit, rather than 3-digit, ICD-9 codes. Finally, I considered adding additional CCCs that did not overlap with RxHCCs. These additional CCCs were only included it they were statistically significant in both components of the

[23] I considered it sensible to collapse 5-digit ICD-9 codes into 3-digit ones if a disease code contained the majority of the 5-digit codes within each 3-digit code.

two-part model using basic controls.

Unlike the CMS model, I did not cluster the disease groups into hierarchies. I chose to add additional disease groups in the model even while their might have been significant overlap in treatments because the purpose of this paper is estimating racial/ethnic and educational gradients rather than prediction. The rational is that conditioning on additional diseases produces a more conservative estimate of the gradients in expenditures conditional on diagnosis.

C Additional Tables and Figures

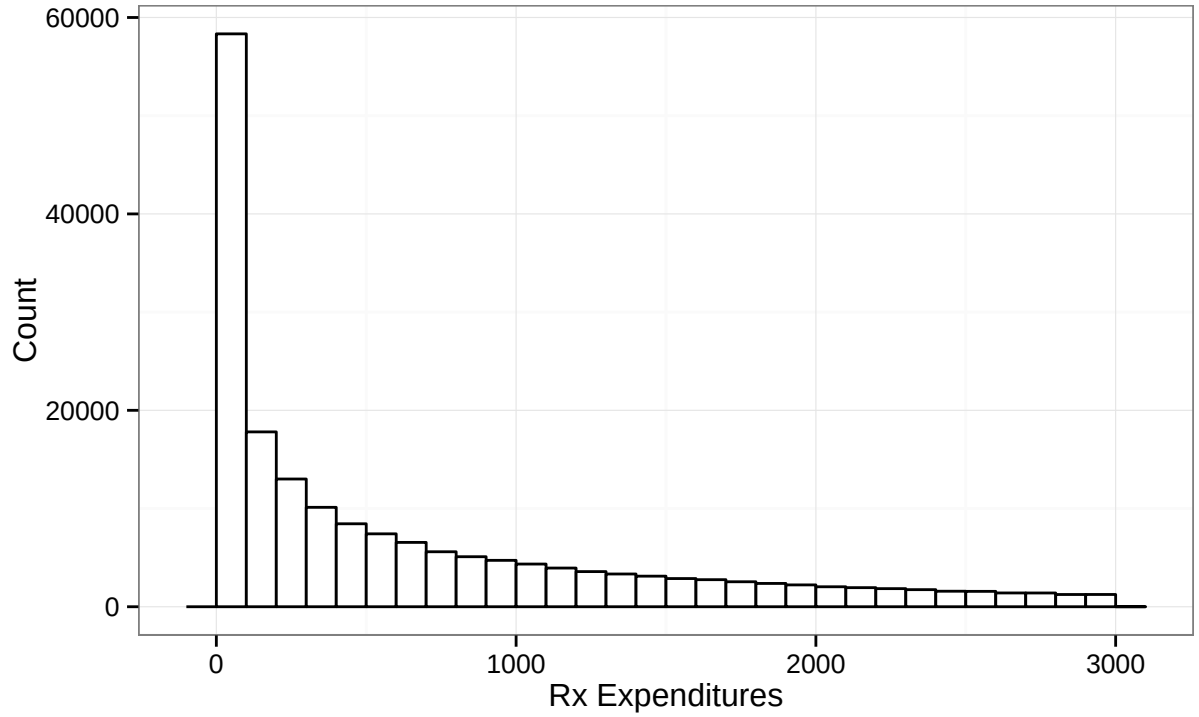

Figure C.1: Histogram of Prescription Drug Expenditures Conditional on having an Rx Medical Condition

Notes: The distribution is truncated at $3000 for graphical purposes, although some expenditures are considerably higher.

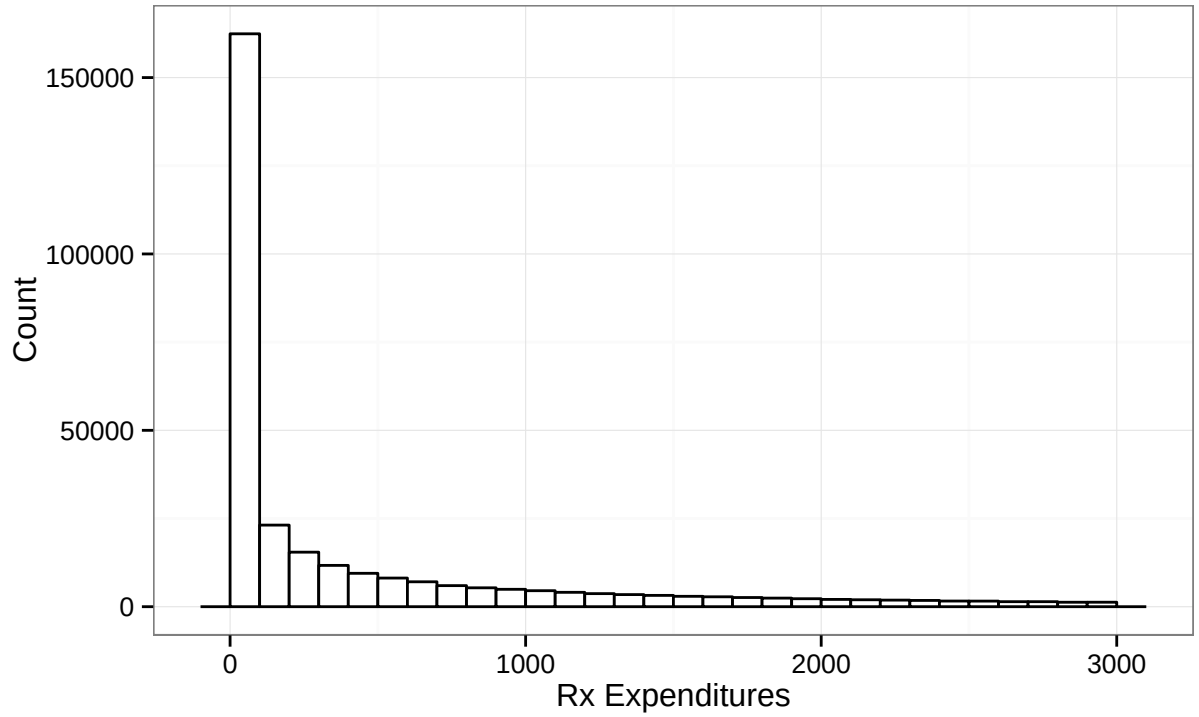

Figure C.2: Histogram of Prescription Drug Expenditures

Notes: The distribution is truncated at $3000 for graphical purposes, although some expenditures are considerably higher.

Table C.1: Regression Coefficients from Two-Part and Selection Models

	Two-part		Selection two-step		Selection MLE	
	Part 1	Part 2	Part 1	Part 2	Part 1	Part 2
Black	−0.235	−0.310	−0.235	−0.184	−0.236	−0.287
	(0.007)	(0.010)	(0.007)	(0.012)	(0.007)	(0.011)
Hispanic	−0.428	−0.528	−0.428	−0.270	−0.427	−0.481
	(0.007)	(0.010)	(0.007)	(0.016)	(0.007)	(0.012)
Years of education	0.041	0.026	0.041	0.005	0.041	0.022
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)

Notes: The two-part model replicates Table 4 using a probit regression for the first component rather than a logistic regression. The selection models use the same explanatory variables as the two-part model but allow the error terms to be correlated. Part 1 of the two-part and the two-step selection models are identical probit models so the coefficients are the same. Reported estimates are regression coefficients.

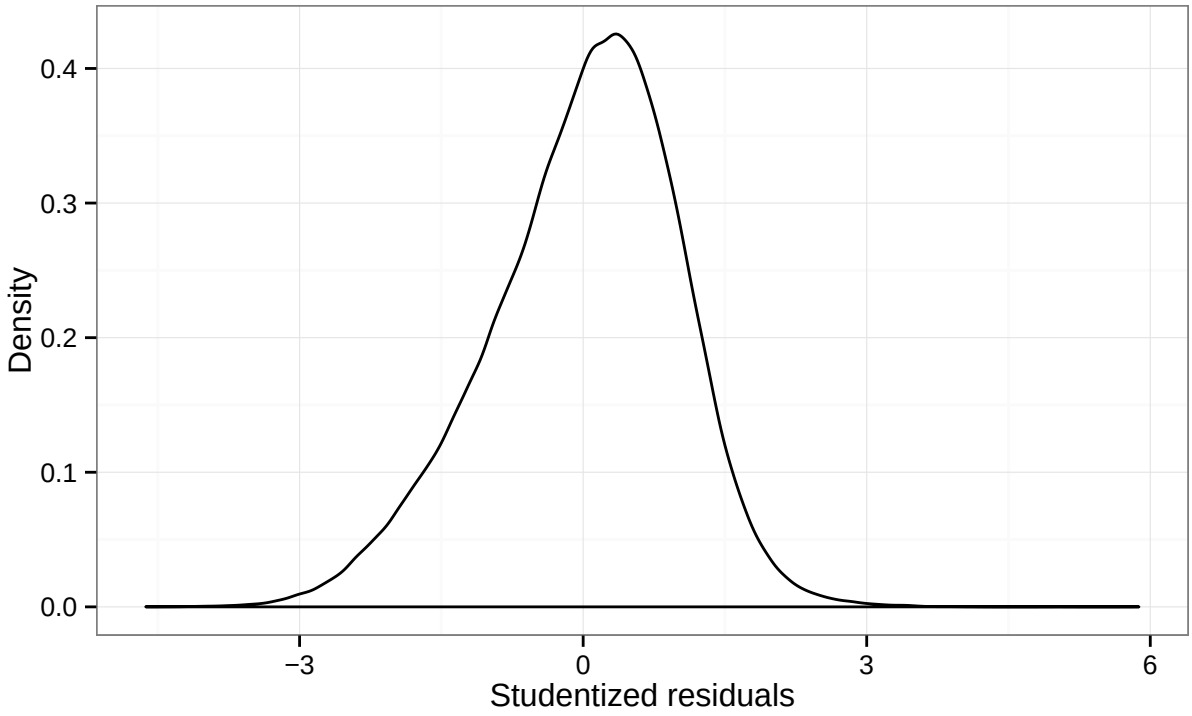

Figure C.3: Density of Studentized Residual for Second Component of Two-Part Model for Continuous Component of Two-Part Model

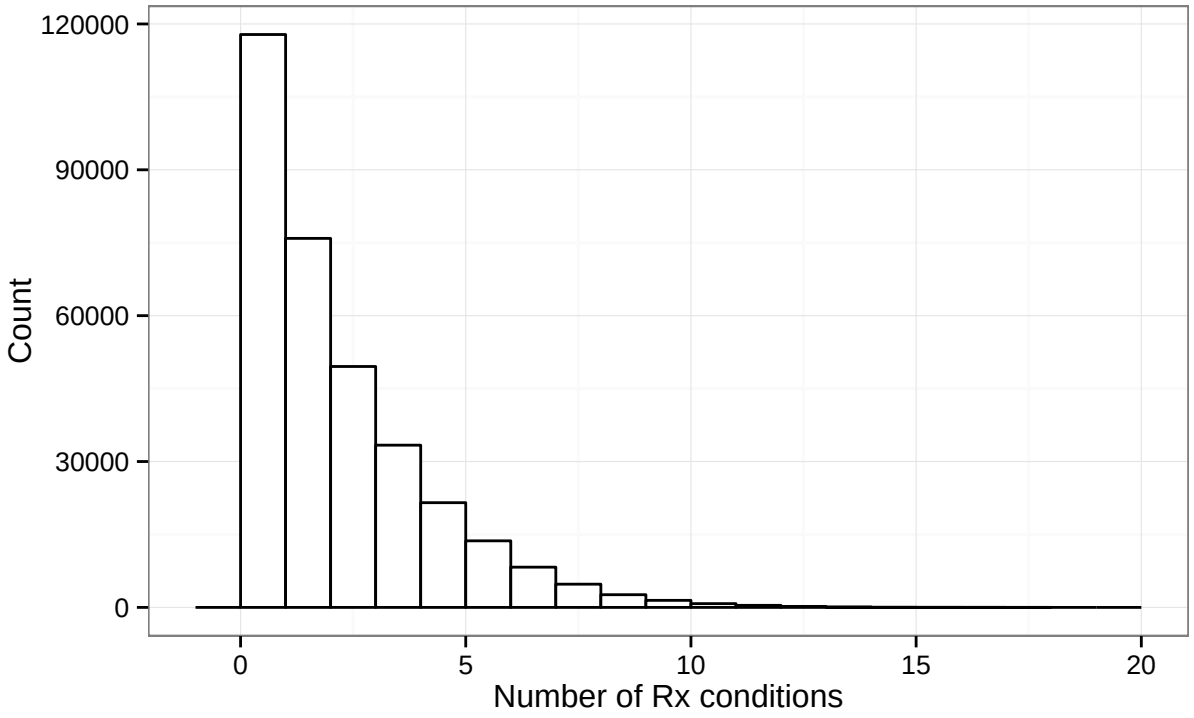

Figure C.4: Histogram of the Number of Rx Conditions

References

Almond, Douglas, Joseph J Doyle, Amanda E Kowalski and Heidi Williams. 2010. "Estimating marginal returns to medical care: Evidence from at-risk newborns." *The quarterly journal of economics* 125(2):591–634.

Bhattacharya, Jay and Darius Lakdawalla. 2006. "Does Medicare benefit the poor?" *Journal of Public Economics* 90(1):277–292.

Blanco, Carlos, Sapana R Patel, Linxu Liu, Huiping Jiang, Roberto Lewis-Fernández, Andrew B Schmidt, Michael R Liebowitz and Mark Olfson. 2007. "National trends in ethnic disparities in mental health care." *Medical Care* 45(11):1012–1019.

Blough, David K, Carolyn W Madden and Mark C Hornbrook. 1999. "Modeling risk using generalized linear models." *Journal of health economics* 18(2):153–171.

Cameron, A Colin, K Pravin, Ulrich Kohler, Frauke Kreuter and J Scott Long. 2009. "Microeconometrics Using Stata." *Atlanta* p. 376.

Cameron, A Colin and Pravin K Trivedi. 1990. "Regression-based tests for overdispersion in the Poisson model." *Journal of Econometrics* 46(3):347–364.

Collins, Karen Scott, Allyson G Hall, Charlotte Neuhaus and Commonwealth Fund. 1999. "US minority health: A chartbook.".

Cook, Benjamin Lê and Willard G Manning. 2009. "Measuring racial/ethnic disparities across the distribution of health care expenditures." *Health services research* 44(5p1):1603–1621.

Cook, Benjamin, Thomas McGuire and Jeanne Miranda. 2007. "Measuring trends in mental health care disparities, 2000–2004." *Psychiatric Services* 58(12):1533–1540.

Cooper-Patrick, Lisa, Joseph J Gallo, Junius J Gonzales, Hong Thi Vu, Neil R Powe, Christine Nelson and Daniel E Ford. 1999. "Race, gender, and partnership in the patient-physician relationship." *Jama* 282(6):583–589.

Cutler, David M and Adriana Lleras-Muney. 2010. "Understanding differences in health behaviors by education." *Journal of health economics* 29(1):1–28.

Cutler, D.M., A.B. Rosen and S. Vijan. 2006. "The value of medical spending in the United States, 1960–2000." *New England Journal of Medicine* 355(9):920–927.

Cutler, D.M. and M. McClellan. 2001. "Is technological change in medicine worth it?" *Health affairs* 20(5):11–29.

Doyle, J. and J. Joseph. 2011. "Returns to Local-Area Health Care Spending: Evidence from Health Shocks to Patients Far From Home." *American Economic Journal: Applied Economics* 3(3):221–243.

Dranove, David. 2009. *The economic evolution of American health care: from Marcus Welby to managed care*. Princeton University Press.

Duan, Naihua. 1983. "Smearing estimate: a nonparametric retransformation method." *Journal of the American Statistical Association* 78(383):605–610.

Duan, Naihua, Willard G Manning, Carl N Morris and Joseph P Newhouse. 1983. "A comparison of alternative models for the demand for medical care." *Journal of business & economic statistics* 1(2):115–126.

Glied, Sherry and Adriana Lleras-Muney. 2008. "Technological innovation and inequality in health." *Demography* 45(3):741–761.

Goldman, Dana P and James P Smith. 2002. "Can patient self-management help explain the SES health gradient?" *Proceedings of the National Academy of Sciences* 99(16):10929–10934.

Gross, Cary P, Benjamin D Smith, Elizabeth Wolf and Martin Andersen. 2008. "Racial disparities in cancer therapy." *Cancer* 112(4):900–908.

Johnson, Rachel L, Debra Roter, Neil R Powe and Lisa A Cooper. 2004. "Patient race/ethnicity and quality of patient-physician communication during medical visits." *American journal of public health* 94(12):2084–2090.

Kinsey, Tracy, Ahmedin Jemal, Jonathan Liff, Elizabeth Ward and Michael Thun. 2008. "Secular trends in mortality from common cancers in the United States by educational attainment, 1993–2001." *Journal of the National Cancer Institute* 100(14):1003–1012.

Link, B.G. and J. Phelan. 1995. "Social conditions as fundamental causes of disease." *Journal of health and social behavior* pp. 80–94.

Lleras-Muney, Adriana and Frank R Lichtenberg. 2002. The effect of education on medical technology adoption: are the more educated more likely to use new drugs. Technical report National Bureau of Economic Research.

Manning, Willard G, Anirban Basu and John Mullahy. 2005. "Generalized modeling approaches to risk adjustment of skewed outcomes data." *Journal of health economics* 24(3):465–488.

Martinez, Steve R, Anthony S Robbins, Frederick J Meyers, Richard J Bold, Vijay P Khatri and James E Goodnight. 2008. "Racial and ethnic differences in treatment and survival among adults with primary extremity soft-tissue sarcoma." *Cancer* 112(5):1162–1168.

Mayberry, Robert M, Fatima Mili and Elizabeth Ofili. 2000. "Racial and ethnic differences in access to medical care." *Medical Care Research and Review* 57(4 suppl):108–145.

Puhani, Patrick. 2000. "The Heckman correction for sample selection and its critique." *Journal of economic surveys* 14(1):53–68.

Robst, John, Jesse M Levy and Melvin J Ingber. 2007. "Diagnosis-based risk adjustment for medicare prescription drug plan payments." *Health Care Financing Review* 28(4):15.

Thomas, John, D Johniene Thomas, Thomas Pearson, Michael Klag and Lucy Mead. 1997. "Cardiovascular disease in African American and white physicians: The Meharry cohort and Meharry-Hopkins cohort studies." *Journal of health care for the poor and underserved* 8(3):270–283.

Weinick, Robin M, Samuel H Zuvekas and Joel W Cohen. 2000. "Racial and ethnic differences in access to and use of health care services, 1977 to 1996." *Medical Care Research and Review* 57(suppl 1):36–54.

Whittle, Jeff, Joseph Conigliaro, CB Good and Monica Joswiak. 1997. "Do patient preferences contribute to racial differences in cardiovascular procedure use?" *Journal of general internal medicine* 12(5):267–273.