Essays in Health Care Economics: Structural Approaches to Measuring Moral Hazard and Adverse Selection

A Thesis Submitted to the Faculty of the Graduate School of the

University of Minnesota

by

Christina L. Marsh

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Advisors: Patrick Bajari and Robert Town

July 2010

Acknowledgements

Particular thanks to Patrick Bajari and Robert Town for their significant support and encouragement. Also, I thank Amil Petrin, Thomas Holmes for their time and thoughtful comments. I would also like to thank Kyoo il Kim, Minjung Park, Connan Snider, Thomas Youle, Kevin Wiseman, Stefania Marcassa, Ioanna Grypari, Justin Barnette, Andrea Szabo, Konstantin Golayev, Meg Henkel, all members of the Applied Micro and Health Economics seminars.

Much gratitude also to the support staff of the University of Minnesota Department of Economics and to the Dissertation Support Group headed by Cynthia Fuller.

To my father who inspired it, and my mother who empowered it.

Abstract

The classic issues of moral hazard and adverse selection as they appear in health care are addressed in this dissertation using new tools of analysis.

In the first essay, I construct a new estimator and estimates to measure the price response of patients in health insurance. These estimates allow us to measure the magnitude of moral hazard. Recent health care initiatives attempt to stem rising costs by increasing patients' cost sharing. These initiatives include high deductible plans, the Medicare Drug Plan "doughnut hole," and Health Savings Accounts (HSAs). The success of such initiatives depends on how health expenditures change as patients' reimbursement decreases. Estimating this elasticity is complicated by selection bias, as high expenditure patients can self-select into high reimbursement plans. Additionally, nonlinear reimbursement is prevalent in U.S. insurance contracts and the aforementioned initiatives. Nonlinearities introduce bias when using previous estimation methods by simultaneously determining expenditure and reimbursement rate. This paper develops an elasticity estimation method that controls for selection bias by taking advantage of nonlinear reimbursement rates. Discontinuous reimbursement rates induced by a nonlinearity are used to isolate patients' expenditure choices. Using detailed claims-level data of employer-sponsored health insurance, I find a tight range of elasticities between -0.25 and -0.33 in the range of average U.S. spending. I then use these estimates in a policy experiment measuring moral hazard and calculate the resulting welfare effects. This paper's estimation method may be used on many policies with nonlinear reimbursement which previous tools could not address.

In the second essay, I present joint work with Patrick Bajari, Han Hong, and Ahmed Khwaja wherein we construct an estimator that can address both moral hazard and adverse selection simultaneously. Theoretical models predict asymmetric information in health insurance markets may generate inefficient outcomes due to adverse selection and moral hazard. However, previous empirical research has found it difficult to disentangle adverse selection from moral hazard in health care. We empirically study this question by a using unique data set with confidential information from a large self-insured employer to estimate a structural model of the demand for health insurance and medical care. We propose a two-step semiparametric estimation strategy that builds on the work on identification and estimation of auction models. We find significant evidence of moral hazard and adverse selection.

Contents

List of Tables

List of Figures

Chapter 1

Estimating Health Care Elasticities Using Nonlinear Reimbursement

1.1 Introduction

Costs of health care have been rising rapidly in the U.S. for the past decade. Real per capita health care spending grew an average of 2.2 percentage points faster than GDP since 1999 (Chernew, Hirth, and Cutler (2009)). Overconsumption caused by generous insurance plans is often cited as a com-

ponent of this rise.[1] That is, patients may consume to a point where their marginal benefit is less than the marginal cost of providing care. When patients pay only a fraction of the cost of their care, this disconnect contributes to moral hazard. Recent health care initiatives are responding by pushing more costs onto patients, for example the Medicare Drug Plan's "doughnut hole" region of full out-of-pocket expenditure, Health Savings Accounts (HSAs), and high-deductible plans with increased expenditure thresholds to receive full coverage.[2] To inform this debate, this paper asks the question, how do patients change their health expenditures when they have to pay more of the total cost?

The answer to this question determines how effectively health reforms may affect demand. Both public and private policy makers use elasticities to forecast the cost impacts of making local changes to patients' out-of-pocket cost. This paper introduces an elasticity estimator on the previously problematic regions of nonlinear reimbursement schedules to answer a question that has been limited by reliance on experimental data, and updates a measure of elasticity that is over 20 years old. Using these estimates, this paper addresses a second goal of measuring the magnitude of moral hazard associated with low out-of-pocket costs.

Nonlinear reimbursement schedules are prevalent in the U.S. health in-

[1] Wall Street Journal (June 12, 2007)

[2] For example, General Electric this year reduced the plan choices for all 75,000 of its salaried employees to three high-deductible plans. New York Times (October 17, 2009)

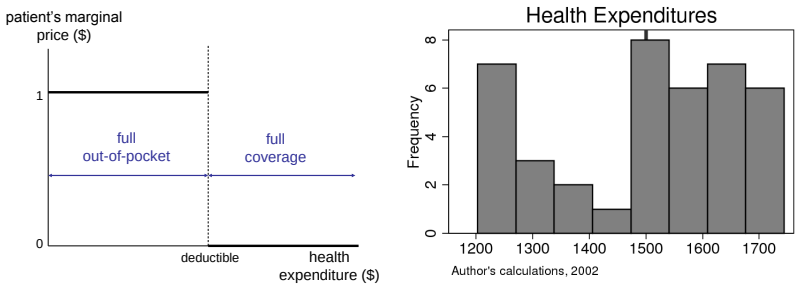
surance system, as well as in the aforementioned health initiatives. The most common example is a deductible, a threshold with full out-of-pocket cost to the patient before reaching the deductible but full reimbursement afterwards. Elasticity estimates are complicated by the presence of nonlinearities for several reasons. First, observational data is subject to selection bias because patients with greater illness severity will have correspondingly higher health expenditures. This higher level of expenditures in turn determines the patient's price. Additionally, the determinant of selection bias, patient characteristics which determine illness severity, is unobserved by the insurance provider and the econometrician. Error due to omitted patient characteristics will be correlated with observed characteristics. Secondly, expenditure and price are simultaneously determined because the level of expenditure determines the patient's position relative to the threshold, and thus the price a patient faces. OLS approaches to estimating elasticity will be biased in this case. Finally, existing estimates in the presence of nonlinearities assume patients face only one price over all expenditures, but, in practice, expenditures are based on two different marginal prices. Choosing a single price introduces systemic measurement error in price. The sign of the estimate's bias due to measurement error in price depends on the proportion of a patient's spending in each region. When all patients are aggregated, the attenuation bias is indeterminant. To control for these complications, this paper takes advantage of the discontinuous change in marginal prices

at the nonlinearity.

The fundamentals of the estimation framework are displayed in Figure 1.1 for the example of a deductible. The horizontal axis is an individual patient's health expenditures and the vertical axis displays the marginal price associated with a particular level of health expenditure. Before hitting the deductible, the patient must pay full out-of-pocket cost for each unit of health care. However, as soon as expenditures reach the deductible, the patient faces a marginal cost of zero because expenditures are fully covered by the insurance plan. Marginal price jumps discontinuously from one to zero at the nonlinearity.

However, consider the omitted patient characteristics in this same window. The characteristics of patients whose expenditures approach the discontinuity from the left should be very similar to the characteristics of patients whose expenditures approach the discontinuity from the right. Immediately before and after the deductible, patients will be similar but will face different marginal prices. To calculate an elasticity of health expenditure, we want to isolate the effect of marginal price while controlling for other omitted characteristics. The sharp change in marginal price around the deductible will be used to identify the effect of price changes.

Given that patients face a discontinuous change in their marginal price, Panel 1.1b displays the resulting change in observed expenditure behavior for a window surrounding the deductible. The histogram displays the dis-



(a) MC at Deductible (b) Expenditures At Deductible

Figure 1.1: Fundamentals of the Estimator

tribution of patients along the horizonal axis of annual health expenditure. The deductible threshold is represented by the vertical line at $1500. This vertical line is the point where patient cost changes from full out-of-pocket cost to free care. Notice that patient expenditure increases significantly at the deductible threshold. At this threshold, patients just above and just below share similar omitted characteristics – similar enough to induce almost the exact same level of expenditure. However, marginal prices were markedly different, which induces a dramatic change in observed aggregated patient behavior. To capture elasticities from the different pricing regions, I estimate the relationship between omitted patient characteristics and observed health expenditures in the each price region. Each pricing region has a different response in health expenditure to a marginal increase in omitted characteristics of illness severity. The difference between the two responses will identify the part of the response that is due solely to the price difference.

This change in patient behavior captures the elasticity.

To formalize the estimation of this change in behavior, I present a general model of health expenditure choice to capture a common prediction of previous approaches – that health expenditures are strictly increasing omitted patient characteristics, notably illness severity. As a result, the underlying distribution of omitted patient characteristics can be identified using the percentiles of observed health expenditures. From this step, I estimate an empirical inverse cdf of health expenditures across the window surrounding the nonlinearity. This empirical inverse cdf maps the levels of expenditure to the distribution of omitted patient characteristics. I then measure, on each side of the nonlinearity, the marginal increase in choice of health expenditure for a marginal increase in illness severity. As patients approach the nonlinearity, the only difference in the choice environment is the marginal price, thus the difference between each side's expenditure response to increasing illness severity isolates the portion the total response that is attributed solely to price. To convert this price response into an elasticity measure, I then multiply the price response by the price and expenditure levels at the nonlinearity where it was estimated.

I then apply this estimator to a unique claims-level dataset for a large employer. The plan offered by this employer includes two separate nonlinearities in the reimbursement schedule, and both of these nonlinearities' thresholds change from year to year. This feature of the dataset lends two

advantages to my estimates not usually available in estimation around a discontinuity. First, multiple estimation points allow robustness checks on the estimation method. Secondly, although the use of nonlinearities produces highly consistent estimates at the threshold of estimation, a disadvantage is that estimates only apply to a small window around the nonlinearity. However, the ability to estimate different threshold levels expands the region of expenditures to which this paper's estimates apply. The resulting estimates are representative of the most common category of insurance in the U.S., employer-sponsored insurance, and apply to a range into which over 50 percent of U.S. expenditures fall each year.

For local changes in patient reimbursement, I find an elasticity of approximately -0.26 in windows around $600 of health expenditures. This elasticity is comparable to the existing "gold standard" of -0.22 from the RAND Health Insurance Experiment (HIE), a large government experiment which addressed simultaneity bias by randomizing patients into insurance plans with different reimbursement levels. For a higher range of expenditures near $1,500, I find elasticity to be slightly more inelastic, at -0.8. A more inelastic response as expenditures increase was also a feature of the RAND estimates. Both estimates are lower than Kowalski (2009) and Eichner (1998), which are subject to attenuation bias due to price mismeasurement.

To conclude, I use these estimates to conduct counterfactual policy ex-

periments. The first policy experiment explores the welfare implications of decreasing the change between the marginal prices on each side of the non-linearity. This policy counterfactual is motivated by popular concern over the magnitude of the Medicare "doughnut hole" coverage gap. I use the elasticity estimates to construct individual patient demand and then perform traditional welfare calculations for the loss in consumer surplus and the results for insurers' costs. The second counterfactual measures the relative magnitude of inefficiency in health care consumption caused by full reimbursement of health expenditures. I compare two regimes, one with full reimbursement for all health expenditures and the other with positive out-of-pocket costs for patients. I then estimate the compensating transfers that would be necessary to make patients indifferent between full reimbursement and positive out-of-pocket costs. The results show that the compensating transfers are less than the cost savings in health care that result from decreasing patient reimbursement levels.

This paper makes three major contributions. First, I present an estimation method that exploits a prevalent feature of insurance contracts, non-linear reimbursement, that previously caused bias in estimates. Secondly, I use this method to find an elasticity for health expenditure in employer-sponsored insurance plans, an important estimate for policy makers. Finally, using this elasticity, I calculate the welfare effects of policy changes which increase cost sharing for patients and calculate a measure of moral hazard.

The rest of the paper proceeds as follows: Section 2 details the prevalence of nonlinearities in U.S. health insurance plans. Section 3 sets up a general model of health care to describe why traditional regression techniques cannot be used. Section 4 lays out the estimation framework, Section 5 describes the data, and Section 6 discusses the elasticity results. Section 7 describes the policy experiment and the results, and Section 8 describes the moral hazard estimation and results. Section 9 concludes.

1.2 Nonlinearities in Health Insurance

Nonlinear pricing is a common feature of many U.S. health insurance plans. This occurs when the price a patient pays varies as he increases his health expenditures. One common example of a nonlinearity is a deductible. A deductible usually features higher cost sharing with the patient before a threshold, with lower cost sharing after expenditures surpass the deductible. This structure is commonly found in private fee-for-service plans, in out-of-network expenditures for Preferred Provider Organization (PPO) plans, and some Health Maintenance Organization (HMO) plans. PPO plans covered 60 percent of workers in 2006, and 69 percent of these workers with single coverage faced a general plan deductible. For the other major U.S. plan categories in the same year, 32 percent of covered workers in a Point-of-Services (POS) plan faced deductibles and 12 percent of covered workers in

HMOs faced a deductible.[3] Outside of a general plan deductible, many plans contain a specific deductible in areas such as hospital services, outpatient procedures, or pharmaceuticals. Deductibles come into force well within average U.S. health expenditures. For plans with a general deductible, the average deductible for single coverage is $473 for PPOs, $352 for HMOs, $553 for POSs, and $1,715 for High Deductible Plans. (Kaiser Family Foundation (2006))

In the public sector, Medicare Part A, Medicare Part B, many plans in Medicare Advantage, and the Drug Plan (Part D) all contain deductibles. For 2010, the largest category of Medicare, Part A (Hospitalization Insurance), has a $1,100 deductible for hospitals stays in each benefit period. Medicare Part B (Medical Insurance) has a $155 deductible yearly for all covered services. Medicare Advantage, sometimes referred to as Part C, combines the coverage of Parts A and B by individually contracting with private insurance providers. The structure of these Medicare Advantage plan varies by provider, but the presence of deductibles mirrors the employer-based plans described above. Finally, in the Medicare Drug Plan, Part D, many plans features a "doughnut hole" area of nonlinear reimbursement where coverage of pharmaceutical expenditures drops from positive reimbursement, down to zero coverage, and then back up to full coverage. (Centers for Medicare and Medicaid Services (2009)) This "doughnut hole" is

[3]In 2006, PPOs covered 60 percent of workers, HMOs 20 percent, POS 13 percent, and High Deductible Plans covered 4 percent, conventional plans 3 percent. (**?**)

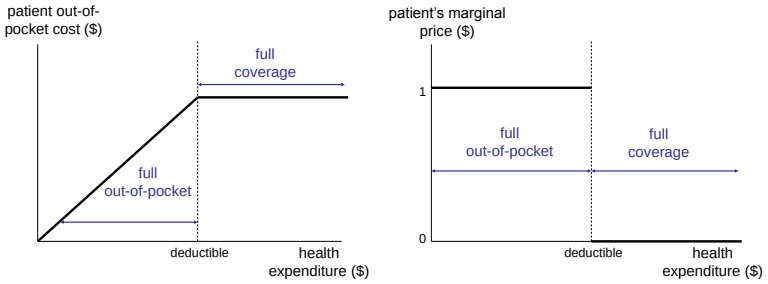
cause for popular concern, because approximately 25 percent of enrolled seniors' drug purchases surpassed the first nonlinearity in 2006, and only 5 percent of enrollees' purchases surpassed the second nonlinearity to gain the benefit of "catastrophic" coverage.[4]

Another common nonlinearity is an employer-funded Health Savings Account (HSA), where an employer puts funds into a patient's account that may only be used for health care purchases. These plans were created by the Medicare Modernization Act 2003, and have been suggested as a base plan for federal coverage, as well as a way of forcing patients to be more informed consumers of their own health care. Currently, 8 million to 10 million Americans are enrolled in high-deductible plans.[5]

In my estimation, I will be considering two nonlinearities, a deductible and a Health Savings Account. Figure 1.2 displays the relationship between a patient's out-of-pocket cost and his total expenditure in the presence of a deductible, a change from full out-of-pocket cost to full coverage. The relationship will be similar for other nonlinearities changing from any low reimbursement level to a higher reimbursement. In Figure 1.2a, the vertical axis is the total cost to the patient of his health care expenditures. This horizontal axis is the total cost of health care which is billed to the insurance. Before the deductible, the patient must pay the full amount of his health expenditures, so his out-of-pocket cost increases at the same rate as his

[4]US News and World Reports (February 3, 2009)
[5]Wall Street Journal (June 12, 2007))

total expenditures. After total health expenditures reach the level of the deductible, all subsequent expenditures are fully covered.

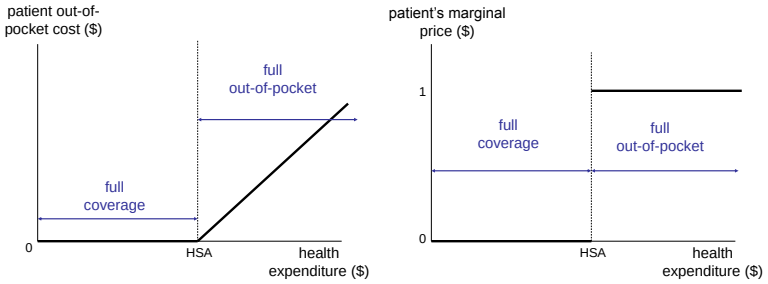


(a) Deductible in levels (b) Deductible in marginal cost

Figure 1.2: Nonlinear Pricing Schedule, Deductible

The patient's marginal cost changes discontinuously from 1 to 0. Figure 1.2b shows the marginal cost structure generated for the patient by the deductible presented in Figure 1.2a. The horizontal axis is again the total health expenditures in dollars made by the patient. On the vertical axis is the marginal price to the patient of an additional dollar of health expenditure.

A Health Savings Account (HSA) is the reverse nonlinearity scenario, as shown in Figure 1.3. In this case, a patient has full coverage for all expenditure that falls below the level of the HSA, with a corresponding marginal cost of zero per additional dollar of care. After reaching the full amount of the HSA, any health expenditure beyond the HSA level is full out-of-pocket spending, with a marginal cost of one.



(a) Health Savings Account in levels (b) Health Savings Account in marginal cost

Figure 1.3: Nonlinear Pricing Schedule, Health Savings Account

1.3 Nonlinearities to Elasticity Estimation

In this section, I set up a general model of a patient's choice of health expenditure using standard assumptions of previous work. I then examine the predictions of this model and demonstrate that any OLS approach would be biased in the presence of nonlinear reimbursement schedules. The model's predictions will be used in Section 4 for an estimator which addresses the difficulties presented here.

1.3.1 General Model of Health Expenditure

This section lays out a model of a patient's choice of health expenditure within an insurance plan. The model presented here is similar to the framework in Huang and Rosett (1973), and generates the same reduced form predictions as the approaches in Kowalski (2009), Manning, Newhouse, Duan, Keeler, and Leibowitz (1987), and Newhouse, Phelps, and Marquis (1980).

In what follows, the patient is choosing his annual expenditure in dollars after having already chosen his insurance plan. This framework could be modeled alternatively as a joint decision of a patient with his doctor, where the goal of the optimization is to maximize the health of the patient. Without loss of generality, here I model only the patient.

To place this model a context appropriate to the data that will be used, consider the example of a patient attending a physician visit, which is a typical scenario of the data used here, as well as of the expenditure levels in the reforms motivating the policy experiments. The patient may benefit from follow-up visits to the physician, with a cost for each visit. As the cost to the patient of each visit changes, the patient will also adjust the frequency with which he sees the doctor. This response is a combination of the severity of the patient's illness and the cost to the patient of visiting the physician.

A patient has utility over his health expenditure, h, factors which affect his illness severity θ, and composite good consumption, c:

$$U(h, \theta, c)$$

The illness severity θ is a random variable, with a cdf F_θ. Higher θ are more severe.

The utility function satisfies the following Utility Conditions:

$$U(h,\theta,c) = u(h,\theta)+c \quad (1.3.1)$$

$$\text{For any given } \theta,\ \exists\ \tilde{h}\ \text{ such that } \frac{\partial u(\tilde{h},\theta)}{\partial \tilde{h}} = 0 \quad (1.3.2)$$

$$\frac{\partial^2 u(h,\theta)}{\partial^2 h} < 0 \quad (1.3.3)$$

$$\frac{\partial u(h,\theta)}{\partial \theta} < 0 \quad (1.3.4)$$

$$\frac{\partial^2 u(h,\theta)}{\partial h \partial \theta} > 0 \quad (1.3.5)$$

Condition 1 is quasilinearity of composite good consumption in the utility function. Estimates will be only on expenditures of \$500 - \$1,500, so I will be assuming no income effects. This assumption is also founded on previous estimates of income elasticities which found income effects close to zero (Phelps (1992)). Condition 2 sets an expenditure point for each level of θ where marginal utility crosses zero. This simply captures the inconvenience cost of doctor visits. Previous work has already shown the importance of time prices, such as travel time, waiting time, and treatment time (Janssen (1992)). This condition also allows that, for patients with a small θ such as a mild cold, imposing high levels of health care, such as full hospitalization, will decrease marginal utility. Condition 3 states that health expenditures exhibit decreasing marginal returns to utility. Condition 4 shows that an increase in illness decreases utility. Finally, Condition 5 states that there are

complementarities in health expenditure and the severity of illness. As the severity of the shock increases, the marginal utility of health expenditure increases.

The patient's budget constraint balances the costs of health expenditures and the composite good consumption with his income. Denote the patient's budget constraint as:

$$c + h - r(h) \leq y$$

Annual income for each patient is denoted as y. The insurance plan reimburses the patient an amount $r(h)$, according to a reimbursement schedule.

The insurance plan's reimbursement schedule, $r(h)$, is nonlinear at a certain level of expenditure, denoted as $\bar{h}$. Consider the following reimbursement schedule, typical of a deductible, where a patient pays full out-of-pocket costs until reaching the deductible, then has no further out-of-pocket costs for any additional units of health expenditure. The reimbursement schedule for an example deductible, $\bar{h}$, is as follows:

$$r(h) = \begin{cases} 0 & \text{if } h \leq \bar{h} \\ h - \bar{h} & \text{if } h > \bar{h} \end{cases}$$

The corresponding marginal reimbursement rates, r', are:

$$r' = \begin{cases} 0 & \text{if } h \leq \bar{h} \\ 1 & \text{if } h > \bar{h} \end{cases}$$

The patient's effective marginal cost for a unit of health expenditure is $MC = 1 - r'$ per unit. Thus, denote the marginal cost schedule as:

$$MC = 1 - r' = \begin{cases} 1 & \text{if } h \leq \bar{h} \\ 0 & \text{if } h > \bar{h} \end{cases}$$

The FOC form the optimal decision rule for choice of health expenditures, h^*, on each r' segment, given the level of θ:

$$\begin{aligned} \frac{\partial u(h^*, \theta)}{\partial h^*} &= 1 - r' \\ MU^*_\theta &= MC \end{aligned} \tag{1.3.6}$$

Figure 1.4 shows a patient's optimization problem with sample marginal utility curves that satisfy the Utility Conditions previously listed. The marginal utility curves are combined with the marginal cost structure described above. Notice the first two curves demonstrate that a patient with greater illness severity θ', will have higher marginal utility for the same level of h, compared to a patient with lower illness severity, θ.

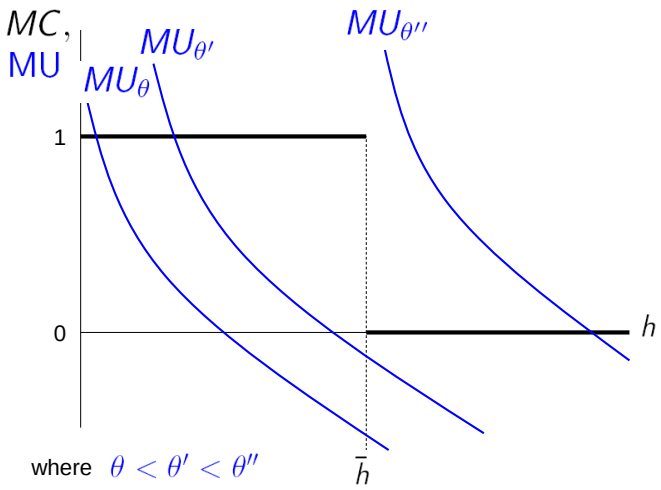


Figure 1.4: Example of Patient's Optimization

There are two general key points to notice in Figure 1.4.

1. Optimal h^* depends on both θ and r'.

2. Optimal h^* is strictly increasing in θ.

The first point states that the exogenous variables that the patient faces are his factors which affect illness severity θ and the reimbursement schedule $r(h)$, which is established by the insurance plan. Secondly, optimal health expenditure, h^*, is strictly increasing in θ. This relationship reflects the balance between Utility Condition 5, the complementarities between h and θ, and Utility Condition 2, decreasing marginal utility of health expenditure.

1.3.2 Invalidity of an OLS Approach

The general model above established the patient's optimal decision rule in an environment with a nonlinear reimbursement schedule. The optimal decision rule depends on marginal price r' and a patient's illness severity θ. However, the model also made clear, first, that r' is based on a nonlinear schedule $r(h)$, and thus is determined by h. Secondly, illness severity θ is a key determinant of the choice of h^*, yet it is unobserved by both the insurer and the econometrician. As a result, any OLS approach to estimating price response will be biased.

To see the bias that would arise from OLS estimation, consider a typical reduced form representation of the patient's optimal decision rule:

$$h_i = a_1 + a_2 r_i' + a_3 \theta_i \tag{1.3.7}$$

where the dependent variable h_i is the patient's choice of health expenditure, and the dependent variables which come from the model are r_i', the marginal price for h_i, and θ_i, factors influencing the patient's illness severity. The corresponding coefficients are $a_1 - a_3$.

Now consider the reduced form optimal decision rule in Equation 1.3.7 above as an estimation equation. Taking Equation 1.3.7 to the data, h_i would be annual health expenditures calculated through insurance claims and r_i' would be the marginal price a patient faced at for the last dollar of

annual expenditure, based on the nonlinear reimbursement schedule set by the insurance plan. The θ_i term is not explicitly observed, so the estimation equation can only include a vector of patient characteristics, X_i, such as age, gender, diagnoses, to approximate some portion of θ. The estimated decision rule is:

$$\hat{h_i} = \hat{a_1} + \hat{a_2}\hat{r_i'} + \hat{a_3}X_i + \varepsilon \tag{1.3.8}$$

The first source of bias in Equation 1.3.8 is that $\hat{h_i}$ and $\hat{r_i'}$ are simultaneously determined. With nonlinear reimbursement, the level of expenditure, h_i, determines the segment of the reimbursement schedule $r(h)$, and thus the r_i' that the patient faces. Additionally, both h_i and r_i' are determined by the underlying unobserved θ_i. Higher levels of illness severity will induce a higher h_i and its corresponding r'.

The second source of bias in the estimation equation 1.3.8 is that the vector of observed patient characteristics will be correlated with the error term, so $E[\varepsilon|X] \neq 0$. The observable characteristics, X, are correlated with unobserved θ. Take for example the age of a patient. An 80-year old patient's expenditures will most likely reflect a higher illness severity than a younger patient. In this case, different elasticity responses between a 20-year old patient and an 80-year old patient are not reflecting the variable of age, but instead the underlying variable of θ. Because illness severity θ_i is

not explicitly observable, the error term ε includes some portion of θ_i that is related to the X value.

A common approach to controlling for unobserved heterogeneity is fixed effects regression. Fixed effects in a panel data on individual patients that would allow the means of the θ_i distribution to be individual-specific. Unobserved illness severity will vary across time, thus fixed effects assumptions are not satisfied – individual effects cannot be differenced out. Without further information on individual θ_i distributions, fixed effects cannot fully remove the systemic correlation with the error term. An additional problem with a panel approach is attrition and selection problems caused by relying on variation across periods. One approach taken in previous literature has been to avoid comparing an individual decision rule and compare average expenditures of demographically similar populations using t-tests of means. A previous year is used as a control for a year with an exogenous price change (Cherkin, Grothaus, and Wagner (1989), Scitovsky and Snyder (1972), Scitovsky and McCall (1977)). A problem with this approach is that the distribution of θ may have changed between the two years that are being compared. For example, comparing demand for physician services must take into account confounding factors across time such as differences in flu seasons or the availability of new treatments or drugs. More importantly, comparing populations over time is necessarily more susceptible to the influence of moral hazard, because the most price sensitive patients drop out of

the sample as prices increase. Most plans allow patients to enroll or disenroll once a year. This paper's estimation method compares behaviors within a year, so relies on within-period variation, thus avoiding the intertemporal problem of exit and entry into the insurance plan. A practical advantage of this approach is that it can be used in cross-sectional data.

Finally, a nonlinear reimbursement schedule presents a patient with different marginal prices along the range of his expenditures. Choosing the appropriate marginal price to associate with a patient's final expenditures can introduce price mismeasurement. This leads to attenuation bias in the estimation. The magnitude of this bias can be large, as Kowalski (2009) finds that estimates can change by a factor of 10. The estimation in this paper instead compares marginal decisions between each marginal price region.

1.4 Estimation Method

The previous section laid out the fundamentals of a patient's optimal choice rule and the resulting biases that would result from an OLS regression estimation framework of the choice of health expenditures on patient incentives of price. Here I present an estimator of price response in health expenditures that captures the predictions of the general model, and is more general than a regression estimator.[6]

[6]The approach that follows here is nonparametric. The Appendix details a flexible parametric form of the model that could also be used to construct the predicted optimal decision equation and estimation equations.

This estimator is based on two conditions. First, patients' prices change discontinuously at the point $\bar{h}$. Second, patients' omitted characteristics which determine θ are continuous across the discontinuity of $\bar{h}$. As the level of expenditures approaches $\bar{h}$ from each side, the individual omitted characteristics, θ, will approach the same value, and the only difference on each side is the patient's marginal cost, $1 - r'$. By using observations just above and just below the discontinuity point, the estimation method controls for unobserved heterogeneity and then compares two similar populations who face a markedly different marginal price. This estimator is similar to a Regression Discontinuity design approach.[7]

Recall the patient faces two exogenous components – the marginal reimbursement rate, r_i', and his illness severity θ_i. While r_i' is observable, the second exogenous variable, θ_i is unobserved. Thus, we must construct an estimator of this unobserved heterogeneity. From the two conditions that emerge from the patient's optimal decision rule, write the general choice of health expenditure as:

$$h_i = G(r_i', \theta_i)$$

where $G : \{0, 1\} \times \Theta \to \mathbb{R}$ such that $\Theta \subset \mathbb{R}$ is support of θ_i, and $\{0, 1\}$ is the support of r_i'. G is strictly increasing in θ_i. This section develops an estimator for F_θ in terms of the distribution of the observed variable, h.

[7]For a more in-depth discussion of RD design, see van der Klaauw (2008) or Chapter 25 Cameron and Trivedi (2005).

In a neighborhood of $\bar{h}$, the decision rule above can be written up to a linear approximation as:

$$h_i = \alpha\theta_i + \delta\theta_i r_i' \tag{1.4.1}$$

where α, and δ are coefficients.[8] Because the position of the expenditure choice relative to $\bar{h}$ determines r_i', rewrite the marginal price in Equation 1.4.2 with an indicator function for the side of the discontinuity:

$$h_i = \alpha\theta_i + \delta\theta_i 1\{h_i > \bar{h}\} \tag{1.4.2}$$

It remains to construct a measure of the omitted patient characteristics, θ. Consider the percentile function, denoted as $p()$. This function is bijective, surjective, and monotone increasing. Because the function G representing the patient's optimal decision rule is strictly increasing in θ, we know that the qth percentile of F_θ, the distribution of θs, corresponds to the qth percentile of the distribution of h, health expenditures. Thus, the distribution of the underlying health shocks can be identified as $\theta = p(h)$. In this way, the distribution F_θ is observable to the econometrician.

The difference between this approach and RD design is that, here, the patient's omitted characteristics are the forcing variable which determines a patient's marginal price. Because the choice of h is strictly monotone in the

[8]These coefficients can also be built out of structural parameters, as detailed in the parametric form of the general utility model detailed the Appendix

omitted characteristics θ, the patient's level of θ is what forces him to the left or to the right of the discontinuity. Patient behavior is identified using the exogenous assignment of the marginal reimbursement rate on either side of the discontinuity. Because the choice of h is strictly increasing in θ the optimal decision rule can be rewritten with the indicator function $1\{h_i > \bar{h}\}$ alternately in terms of the θ_i term:

$$h_i = \alpha\theta_i + \delta\theta_i 1\{\theta_i > \bar{\theta}\} \tag{1.4.3}$$

where $\bar{\theta}$ is the level of the omitted illness severity at the discontinuity. That is, $\bar{\theta}$ is the level of θ that generates the expenditure level of $\bar{h}$.

Equation 1.4.3 can be used to estimate $\widehat{E}[h|\theta]$ using a local linear regression. The presence of the indicator function leads to two estimation equations, one on each side of the discontinuity. Let h_i denote the health expenditure choice of the ith patient.

If $\theta_i < \bar{\theta}$, the left-hand side of the discontinuity, then use:

$$\min_{a_L, b_L} \sum_{\theta_i < \bar{\theta}} \omega(\theta_i) \left(h_i - a_L(\theta) - b_L(\theta)(\theta_i - \theta_0)\right)^2$$

Analogously, if $\theta_i > \bar{\theta}$, the right-hand side of the discontinuity, then use:

$$\min_{a_R, b_R} \sum_{\theta_i > \bar{\theta}} \omega(\theta_i) \left(h_i - a_L(\theta) - b_R(\theta)(\theta_i - \theta_0)\right)^2 \tag{1.4.4}$$

where $\omega(\theta^{(n)}) = K\left(\frac{\theta^{(n)}-\theta_0}{k}\right)$ is a kernel estimator with bandwidth k, and each θ_0 is in a series of points within the range of estimation.

The resulting equations for choice of h will be as follows:

$$h_{iL} = \quad a_L(\bar{\theta}) + b_L(\bar{\theta})(\theta_{iL} - \bar{\theta}) \quad \text{for } \theta_i < \bar{\theta} \tag{1.4.5}$$

$$h_{iR} = \quad a_R(\bar{\theta}) + b_R(\bar{\theta})(\theta_{iR} - \bar{\theta}) \quad \text{for } \theta_i < \bar{\theta} \tag{1.4.6}$$

Approaching $\bar{\theta}$ at equal distances on each side, the only difference between the two equations is due to the part of the h expenditure choice that is due to the different marginal price on each side. Recall that in Equation 1.4.1, the coefficient of interest is δ, the change in the choice of h due to the change in the marginal reimbursement rate, r'. Thus, the response in expenditures due to differences in marginal reimbursement rate will be the difference between the two equations for h_{iL} and h_{iR} for the same marginal increase in θ. Take a θ pair distance K away from each other on each side of the discontinuity.

$$
\begin{aligned}
h_{jL} - h_{kL} &= \left[a_L(\bar{\theta}) + b_L(\bar{\theta})(\theta_{jL} - \bar{\theta})\right] \\
&\quad - \left[a_L(\bar{\theta}) + b_L(\bar{\theta})(\theta_{kL} - \bar{\theta})\right] \\
&= b_L(\bar{\theta})(\theta_{jL} - \theta_{kL}) \\
&= b_L(\bar{\theta})K
\end{aligned}
$$

$$
\begin{aligned}
h_{jR} - h_{kR} &= \left[a_R(\bar{\theta}) + b_R(\bar{\theta})(\theta_{jR} - \bar{\theta})\right] \\
&\quad - \left[a_R(\bar{\theta}) + b_R(\bar{\theta})(\theta_{kR} - \bar{\theta})\right] \\
&= b_R(\bar{\theta})(\theta_{jR} - \theta_{kR}) \\
&= b_R(\bar{\theta})K
\end{aligned}
$$

The difference in patient behavior on both sides due to a marginal increase in θ is captured by the slope coefficients $b_L(\theta)$ and $b_R(\theta)$. Then, the δ estimator for patient response to a change in reimbursement rate is the difference in the slopes $(b_L(\theta) - b_R(\theta))$ at the discontinuity:

$$\delta = \lim_{\theta \downarrow \bar{\theta}} b_L(\theta) - \lim_{\theta \uparrow \bar{\theta}} b_R(\theta)$$

To calculate the elasticity of the choice of health expenditure, the δ term

gives the change in health expenditure due to changes in reimbursement, which is then normalized by the percentile of omitted characteristics, the reimbursement rate, and the level of expenditure at the discontinuity. Thus, the final formula for local elasticity, η, at $\bar{h}$ is:

$$\begin{aligned} \eta &= \delta\,\theta\,\frac{r'}{\bar{h}} \\ \eta &= \left(\lim_{\theta\downarrow\bar{\theta}} b_L(\theta) - \lim_{\theta\uparrow\bar{\theta}} b_R(\theta)\right)\,\bar{\theta}\,\frac{1}{\bar{h}} \end{aligned}$$

1.5 Data

The dataset is claims-level for a large self-insured employer with several locations. Because the employer is self-insured, all individual claims are reported for each of the three years, 2002, 2004, and 2005.[9] An advantage of a self-insured employer is that income information is also available. Each claim entry contains all information necessary for classifying the services received and to remit payments. Thus, each claim contains information on the costs incurred by the patient and the amount covered by the employer, as well as information on the treatment facility, procedure codes, and diagnoses.

In particular, I study the Consumer-Driven Health Plan (CDHP) option available to enrollees, which is a high-deductible plan. This plan contains two nonlinearities. The first nonlinearity results from an employer-funded

[9]The data for 2003 contains inconsistent assignment of enrollees into plan choices and could not be used in estimation.

Health Savings Account (HSA), where the employer deposits funds that can be used to purchase health care from the first dollar spent until the HSA is exhausted. The second nonlinearity is a deductible.

The threshold levels of both the HSA and the deductible change from year to year. This variation in nonlinearity thresholds will aid in identification of patient response to the nonlinear reimbursement schedules. Changing levels of the nonlinearities aids the estimation in two ways. First, changing discontinuity points lend a robustness check to the estimation method if the elasticity estimates at each year remain similar despite the changing threshold level. Second, the estimator has strong validity for observations just at the discontinuity, but validity becomes more limited for observations far from the discontinuity. However, with similar estimates at different discontinuity points, the estimate's validity can be placed onto an expanded range of expenditures. The evolution of the plan nonlinearities during the years 2002, 2004, and 2005 are displayed in Table 1.1.

Table 1.1: Insurance plan structure

	First discontinuity	Second discontinuity
Year	HSA	Deductible
2002	\$500	\$1,250
2004	\$750	\$1,500
2005	\$600	\$1,500

Table 1.2 and Table 1.3 report plan summary statistics for patients who were enrolled under single, as opposed to family, coverage for the entire 12-

month benefit period. Table 1.2 reports the yearly means and medians of total expenditure, employer cost, the yearly means of the amount of HSA used, and the amount of deductible fulfilled. Average total expenditure for all three years was $7,387. However, health expenditure distributions tend to be skewed, so the median total expenditure is lower in all years. The lower expenditure levels of employer cost compared to total expenditures reflects positive out-of-pocket costs to patients. Patient expenditure variables in Table 1.2 are the amount of the deductible fulfilled and the amount of the HSA used. An average patient spent $644 toward his deductible. The average amount used of the HSA was $319.

Table 1.3 reports patient-level characteristics for single-coverage, full-year enrollees in the insurance plan. Plan enrollment in this category grew from 165 enrollees in 2002 to 349 in 2005. The average age over all three years is 48, and the average salary for the enrollees is $55,934. The plan enrolled 72 percent women. This average salary is in the lower range of common definitions of middle class, between 200 and 400 percent of the federal poverty level. Currently, this range is between $44,000 and $88,000. This income range would be eligible for increased insurance coverage and sliding subsidies in the health reform bill recently passed in Congress. The price response of this population is thus of timely importance.

Tables 1.4, 1.5, 1.6 and 1.7 present details of the data sample in the window of estimation around the first nonlinearity. These estimation windows

Table 1.2: Expenditures in Full Sample

Year	**Total Expenditure**		**Employer Cost**		**Deductible used**	**HSA used**
	mean	median	mean	median	mean	mean
2002	$5,251	$1,492	$4,852	$990	$540	$257
2004	$7,130	$2,024	$6,560	$1,551	$665	$377
2005	$8,647	$2,288	$7,824	$1,899	$672	$292
Total	$7,387	$2,016	$6,746	$1,537	$644	$319

Includes only single coverage, full year enrollment.

Table 1.3: Demographics in full Sample

Year	**Enrollees**	**% Female**	**Age**	**Salary**
2002	165	71	46	$58,783
2004	341	72	48	$51,224
2005	349	72	48	$58,965
Total	855	72	48	$55,935

Includes only single coverage, full year enrollment.

Reported age and salary are means.

do not include the entire sample, only a small region around the nonlinearity, as described in the Estimation Method section. Specifically, these tables are within a $400 window on each side of the discontinuity.

Table 1.4 displays the types of facilities where patients received medical services. The facility types are ranked by the sum of the total expenditures attributed to the facility type within the estimation window. The most common facility is a physician's office, at nearly 70 percent of expenditures in all years. Other important facilities include independent labs, OB/GYN offices, and hospital outpatient facilities.

The type of health care professionals delivering services are described in Table 1.5. The most prevalent health professional is a Medical Doctor. Remaining categories, at 10 percent or less, are clinicians, optometrists, lab technicians, and chiropractors.

The types of services used by patients in the HSA sample are detailed in Table 1.6. Corresponding to physicians offices as the most common facility type, the most common type of service is physician care. Other services of the six most common types are routine physicals, lab/pathology services, X-ray diagnostics and vision services.

The estimation method takes advantage of increasing similarity in unobserved characteristics as expenditures approach the nonlinearity. The omitted patient characteristics of interest in the estimator are factors which contribute to unobserved illness severity. To verify that the differences be-

tween sides are not being influenced by other factors, such as more expensive doctors, Table 1.7 compares service zipcodes on each side of the discontinuity. In 2004, the top three zipcodes were identical on each side of the nonlinearity. In 2005, two of the top three zipcodes were the same. Concerning these two differing zipcodes, the pre-HSA zipcode borders the first ranked zipcode within a large metro center, and the post-HSA zipcode is for business addresses within the first ranked zipcode. Zipcode information is not available for 2002.

Table 1.4: Facility Type in HSA Estimation Sample

Rank	**Place of Service**	**Total Expend.**
	2002	
1	Physicians Office	$11,675
2	Independent Lab	$2,480
3	Hospital Outpatient	$1,663
4	Clinic	$518
	Total	$17,179
	2004	
1	Physicians Office	$27,347
2	OB/GYN Office	$2,780
3	Hospital Outpatient	$2,392
4	Independent Lab	$1,908
	Total	$40,173
	2005	
1	Physicians Office	$27,115
2	OB/GYN Office	$1,769
3	Chiropractor	$1,521
4	Independent Lab	$1,374
	Total	$36,832

Includes only single coverage, full year enrollment.

Tables 1.8, 1.9, 1.10, and 1.11 display the corresponding summary statis-

Table 1.5: Provider Type in HSA Estimation Sample

	Percentage of Total Expenditures	
Category	**2004**	**2005**
Medical Doctor	73.0	59.8
Clinic	8.0	10.3
Doctor of Optometry	5.1	4.9
Independent Lab	3.8	2.2
Chiropractor	1.3	4.9
Other types	8.8	17.9
Total	100	100

Includes only single coverage, full year enrollment.

2002 information not available.

Table 1.6: Service type in HSA Estimation Sample

	Total expenditures		
Category	**2002**	**2004**	**2005**
Physician Care	$3,759	$8,149	$8,610
Lab/Pathology	$3,964	$7,820	$5,752
Vision	$2,542	$5,044	$4,812
Routine Physical	$1,727	$5,001	$2,982
X-ray Diagnostic	$1,099	$4,474	$5,679
Psychotherapy	$2,267	$2,185	$1,524
Other Services	$1,821	$7,500	$7,473
Total	$17,179	$40,173	$36,832

Includes only single coverage, full year enrollment.

Table 1.7: Comparison Pre-HSA and Post-HSA within Estimation Window

	Pre-HSA	**Post-HSA**	**Pre-HSA**	**Post-HSA**
	2004		**2005**	
Expenditure	xx480	xx480	xx440	xx480
Locations	xx455	xx455	xx404	xx485
(Modified zipcode)	xx440	xx440	xx480	xx440

Expenditure location is ranked by total annual expenditures.

NA: Zipcode not available for 2002.

tics for the estimation window around second nonlinearity in the insurance plan, a deductible. These tables are for within a $300 window on each side of the discontinuity.

The most common facility types are displayed in Table 1.8. Again the most common facilities are physicians' offices. Another important facility compared to the HSA window is independent labs, however hospital outpatient spending is now prevalent in each year, as well as a larger amount of clinic use.

Table 1.9 lists the most important providers at the deductible estimation window. Medical doctors are the largest percentage of expenditures, similar to the HSA window. However, clinics and hospitals feature more prominently, and expenditures are spread into a larger range of providers, as evidenced by a large amount of expenditure outside of the top 5 provider types.

The most common types of services are shown in Table 1.10. Physician care, lab/pathology, and vision services are among the top 6 services, in common with the HSA window. The deductible estimation window has a higher incidence of psychotherapy, surgery, and physical therapy services.

Finally, several common zipcodes appear in Table 1.11 on each side of the deductible. Again the zipcodes xx480 and xx440 are different business addresses in the same metropolitan center, so the most common zipcodes in both years before and after the nonlinearity are from the same geographical

area.

Table 1.8: Facility Type in Deductible Estimation Sample

Rank	Place of Service	Total Expend.
	2002	
1	Physicians Office	$18,139
2	Independent Lab	$1,938
3	Hospital Outpatient	$1,700
4	Pharmacy	$784
	Total	$24,931
	2004	
1	Physicians Office	$32,713
2	Hospital Outpatient	$6,024
3	Independent Lab	$5,748
4	Clinic	$3,897
	Total	$60,111
	2005	
1	Physicians Office	$28,460
2	Independent Lab	$4,928
3	Hospital Outpatient	$3,087
4	Clinic	$2,257
	Total	$51,713

Includes only single coverage, full year enrollment.

1.6 Elasticity Estimation Results

I estimate elasticities for all patients over their yearly expenditures, for each of the three years. Yearly estimates are necessary because the level of the nonlinearities, and thus the $\bar{\theta}$ changes every year. The results for the first nonlinearity, the HSA, are displayed in Table 1.12 and the results for the second nonlinearity, the deductible, are displayed in Table 1.13.

Standard errors were calculated using the asymptotic distribution prop-

Table 1.9: Provider Type in Deductible Estimation Sample

	Percentage of Total Expenditures	
Category	**2004**	**2005**
Medical Doctor	42.3	53.3
Clinic	12.6	14.7
General Hospital	3.1	12.0
Chiropractor	1.4	2.5
Doctor of Optometry	0.6	2.3
Other types	40.0	15.3
Total	100	100

Includes only single coverage, full year enrollment.
2002 information not available.

Table 1.10: Service type in Deductible Estimation Sample

	Total expenditures		
Category	**2002**	**2004**	**2005**
Lab/Pathology	$5,550	$10,191	$10,161
X-ray Diagnostic	$1,909	$8,598	$7,895
Psychotherapy	$4,035	$4,351	$2,991
Vision	$2,019	$3,082	$3,748
Surgery	$1,788	$2,746	$2,960
Physical Therapy	$808	$2,075	$4,084
Other Services	$8,822	$29,068	$19,874
Total	$24,931	$60,111	$51,713

Includes only single coverage, full year enrollment.

Table 1.11: Comparison Pre-Deductible and Post-Deductible within Estimation Window

	Pre-Deductible	**Post-Deductible**
	2004	
Expenditure Locations	xx480	xx440
(Modified zipcode)	xx440	xx403
	xx812	xx435
	2005	
Expenditure Locations	xx440	xx440
(Modified zipcode)	xx480	xx485
	xx486	xx480

Expenditure location is ranked by total annual expenditures.

NA: Zipcode not available for 2002.

erties developed in Bajari, Hong, Park, and Town (2010). In this method, the difference between the true and estimated value of the limit of the health care expenditure h values from above and below can be shown to converge to an independent exponential variable with hazard rates of $f^{-}(h_L)$ from below and $f^{+}(h_R)$ from above. The local linear regression slopes b_L and b_R can be written as the inverse of these hazard rates. Next, the hazard rates are estimated from the data with a one-sided kernel density. The difference between the estimated and true value of the difference in the slopes, $(b_L - b_R)$, converges asymptotically to a normal distribution with variance calculated using $f^{-}(h_L)$, $f^{+}(h_R)$, and properties of the chosen kernel.

The heading of Table 1.12 reports the elasticity formula from Equation 1.4.7 for the HSA nonlinearity. The HSA is an area that goes from zero out-of-pocket cost to full out-of-pocket cost, the reverse of the deductible.

Therefore the difference between the slope coefficients is $(b_L - b_R)$. Because the HSA marginal prices change in an opposite way from the deductible, the differences between the coefficient is simply the reverse order of slope coefficients from that detailed in the Estimation Section. The first column reports the year and the second column reports the health expenditure level of the discontinuity, $\bar{h}$. As stated above, these estimates are for the first discontinuity, so the $\bar{h}$ reported is the amount of the HSA for that year. The third and fourth column report the slope coefficients b_R and b_L at the limit from the local linear regressions on each side of the nonlinearity. Column 5 reports the difference in the slope coefficients – the coefficient δ from the estimation section. This is the treatment effect of changing reimbursement rates. Column 6 reports the value of $\bar{\theta}$ that corresponds to the $\bar{h}$ in that year. The final column reports elasticity estimates. Recall that the coefficient on the difference in behavior between the two pricing regions is the difference in slopes as the illness severity approaches the $\bar{\theta}$ at the discontinuity. Therefore, the value of the elasticity is reported over a small range of θ on either side of the discontinuity.

The elasticity estimates around the discontinuity range between -0.25 in 2005 to -0.26 in 2004, to -0.33 in 2002. All are in the low inelastic range. The estimates are similar across all the years. This is especially noteworthy because the discontinuity value of $\bar{h}$ varies across years, so observations in the window of the local linear regression are not in the same level of expenditure

Table 1.12: Elasticity Estimates at HSA

$$\text{Elasticity} = \eta = (b_R - b_L)\,\bar{\theta}\,\frac{1}{\bar{h}}$$

Year	N	$(b_R - b_L)$	$\bar{\theta}$	$\bar{h}$	η
2002	39	-2.66	62	500	**-0.33**
					(0.15)
2004	55	-3.74	51	750	**-0.26**
					(0.07)
2005	61	-3.30	46	600	**-0.25**
					(0.12)

Standard errors in parentheses.

in each year. The discontinuity expenditure level changes in Column 2 from \$500 to \$750 to \$600. This provides evidence that the estimates are robust to changes in the threshold level.

The elasticity estimates were performed on the first discontinuity in my data. Specifically, this nonlinear pricing region is preferable for estimation because the range of expenditures at this nonlinearity has the least income effects in marginal decisions. Because the discontinuity points were all less than \$1,000 and all the enrollees were employed, this region is the most reasonable to assume that income does not influence decisions.

The estimates above are similar to several other robustness specifications. Local linear regressions using different window sizes in the ranges of $[100, 400]$ produced similar elasticity estimates. Results were also similar with different bandwidth choices, using both an optimal bandwidth rule for a Gaussian kernel and by inspection.

Table 1.13 reports the elasticity estimates at the second nonlinearity, the deductible. The deductible is a case where MC shifts from full out-of-pocket cost, $MC = 1$, to full coverage, $MC = 0$, so the difference in the slope coefficients from the Estimation Section will be $(b_L - b_R)$. The elasticities at this nonlinearity are more inelastic. This is likely due to the higher level of expenditure at this nonlinearity. In fact, it is twice as large as the HSA levels.

Table 1.13: Elasticity Estimates at Deductible

$$\text{Elasticity} = \eta = (b_L - b_R)\,\bar{\theta}\,\frac{1}{\bar{h}}$$

Year	N	$(b_L - b_R)$	$\bar{\theta}$	$\bar{h}$	η
2002	22	-1.49	73	1250	**-0.09**
					(0.17)
2004	40	-2.64	46	1500	**-0.08**
					(0.05)
2005	34	-2.69	48	1500	**-0.09**
					(0.06)

Standard errors in parentheses.

1.6.1 Discussion of Results

These estimates are in line with the best existing estimates, those from the RAND government-funded health insurance experiment in the 1970s. These estimates place consumers squarely in the inelastic region of demand, which is in line with common sense about health care as a necessary good. The benefit of this paper's elasticities is that the elasticities are calculated

from non-experimental insurance data, and thus the same method may be applied to other non-experimental datasets to provide more precise estimates in cases that do not exactly replicate the setting of the RAND experiment.

The inelastic property of these elasticity estimates casts doubt on the efficacy of health reform initiatives that seek to control the costs of care through increasing patients' share in the cost of their care. The amount of overconsumption is likely to be small currently, and decreasing reimbursement rates may have only a small effect on patients' propensity to seek care. These estimates also suggest that the magnitude of moral hazard in this framework may be low because patients do not change behavior by a large magnitude given a change in reimbursement. That said, any consideration of moral hazard should also take into account the total cost of providing the care to the patient in relation to changes in his choice of expenditure. However, because this estimation method calculates patient response over a small window around the threshold, the difference in costs between observed expenditures is likely to be minimal, because it is bounded by the size of the window.

To place these results in context, we must also discuss the study population and the generalizability of these results to a larger population. The advantage of RD design over other quasi-experimental techniques is a high level of internal validity within the population. The most immediate population is patients enrolled in this health insurance plan in a given year,

who had health expenditures near the nonlinearity. The external validity to populations outside the sample requires further discussion.

An advantage of my data and approach is that the threshold value of the nonlinear reimbursement rates changes from year to year. Because I find a tight range of estimates across the different years of the study, and correspondingly different health expenditure levels, this supports these elasticities being more generally applied to the entire lower range of expenditures of the plan. Over 50 percent of the U.S. insured population had annual health expenditures that fall in this range, according to the Agency for Healthcare Research and Quality. In 2002, the majority of insured expenditure fell below $644. The expenditure levels of the nonlinearities above are within this range – $500, $750, and $600. The U.S. health care expenditure is skewed, with most expenditures in the lower ranges, and a very long tail in the high expenditure region. Thus, the expenditure levels represented by these elasticity estimates capture the activity of the most populous part of the U.S. health care expenditure distribution.

Secondly, these estimates are over a population of patients enrolled in employer-sponsored health insurance. Most Americans are insured through employer-based insurance, at over 63.5 percent of the non-elderly population in 2003. The population in this study are part of a large population of employees for the firm that provided the data. As previously detailed, the firm is large, with several locations, so the pool of employees represent a

sample window of U.S. patients with employer-sponsored plans. To consider the general applicability of this representative window, two primary points are important to consider. The population of this study is richer than the average within the large firm, thus elasticity estimates could be lower than a more general population of employer-based enrollees if the price of health care is a lower proportion of their total income. However, as discussed previously, the expenditure ranges considered here should not have large income effects for an employed population. Additionally, the patients in the study population are, on average, healthier than the firm's population as a whole. If the severity of health shocks is lower than a more general privately-insured population, then these elasticity estimates may show patients to be more responsive to reimbursement rate if elasticity increases with severity of a health shock.

Given these caveats, we can apply these estimates to the lower portion of health expenditures for patients in employer-sponsored plans. The nature of the RD design estimation above is to examine small changes around the nonlinearity, so these estimates are indicative of the response in health expenditures to local changes in the cost-sharing imposed on patients.

1.7 Policy Experiment

1.7.1 Setup

Given the above estimates of the local elasticity of the demand for health care, we can now consider the potential welfare implications of recent health initiatives. The drop in coverage used for estimation in this paper forms the basis of two important policy measures introduced recently as an attempt to control the increase in health care expenditures. The first policy is Consumer Driven Health Plans (CDHPs), which came into force in 2003 when federal legislation increased tax benefits for using these plans. Proponents of CDHPs claim that the plan forces patients to be more aware of their health care purchases, but critics claim that patients will avoid necessary health care once coverage drops from 100 percent to zero percent. Similarly, this paper's reimbursement structure replicates the infamous Medicare Drug Plan "doughnut hole." The "doughnut hole" is an area where coverage falls from a high level of reimbursement to full out-of-pocket costs. Full coverage resumes after a certain amount of expenditure at the full out-of-pocket cost. Controversy surrounds this drop in coverage because advocates are concerned that patients will cut health expenditures drastically, endangering their health. The Congressional Committee on Ways and Means (2006) reports that patients are willing to pay substantially higher premiums by enrolling in alternative Medicare insurance plans without a coverage gap, on

average 250 percent higher. If patients are willing to pay a higher premium to make the coverage gap less severe, what would be the welfare implications of changing the reimbursement schedule to mitigate the severity of the drop in reimbursement rates?

In response to these questions, the policy experiment I conduct is replicating the environment of a coverage gap using my elasticity estimates, and then changing the reimbursement rates on each side to soften the drop from full coverage to full out-of-pocket. That is, I increase the marginal price above zero (full coverage) before the gap, and decrease the marginal price below one (full out-of-pocket) during the gap. I then calculate the resulting consumer surplus, producer surplus, and total welfare in each of these scenarios.

Because I conduct this policy experiment only within the range of health expenditures used in the estimation above, I assume constant elasticity of demand across the local window surrounded the nonlinearity. This is reasonable for several reasons. First, I consider only local changes in reimbursement rates. Thus, policy changes are taking place in a window around the nonlinearity, exactly the area where the elasticity estimates were calculated. Secondly, the range of elasticity estimates is not large over multiple years of estimation. During these years the range of health expenditures used for estimation changed with the deductibles and HSA limits set by the plan. Similar elasticity values were found over a slightly larger range than any

single year's threshold.

The market demand for health care thus follows the form for constant elasticity of demand, $H = AP^{\eta}$, where H is market demand for health care, A is a constant I will calculate from the data, P is marginal price, and η is the preceding section's elasticity estimate.

The first scenario is the current policy reimbursement schedule, with marginal price $p = 0$ before the nonlinearity and marginal price $p = 1$ after the nonlinearity. The second scenario considers changing the marginal price on each side from the endpoints of 0 and 1 by an amount τ where $0 < \tau < 1$.

Figure 1.5 illustrates the market demand curve for health expenditures, and the change in consumer surplus which would result from introducing a positive marginal price of τ into the full coverage area. The total consumer surplus at $p = 0$ is the area under the curve from the origin to $h(p = 0)$. This is the sum of areas A, B, and C. The total change in consumer surplus by increasing the price from $p = 0$ to $p = \tau$ is a loss of areas B and C.

I calculate the change in consumer surplus from introducing a small marginal price of τ into a previously full coverage region using the following formulas, the equivalent of B + C:

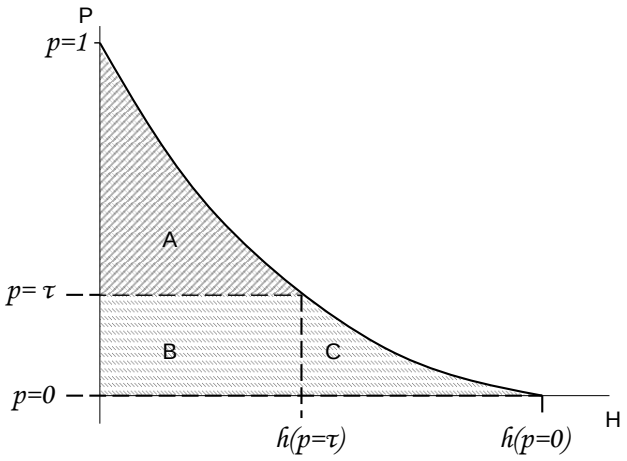


Figure 1.5: Change in consumer surplus from increasing $p = 0$ to $p = \tau$

$$\begin{aligned} CS &= Benefit - Cost \\ CS(p=0) &= \int_{h(p=1)}^{h(p=0)} (\frac{A}{H})^{\frac{1}{\eta}} dH - 0 \cdot h(p=0) \\ CS(p=\tau) &= \int_{h(p=1)}^{h(p=\tau)} (\frac{A}{H})^{\frac{1}{\eta}} dH - \tau \cdot h(p=\tau) \end{aligned}$$

Change in consumer surplus:

$$\Delta CS = -\left[\int_{h(p=\tau)}^{h(p=0)} (\frac{A}{H})^{\frac{1}{\eta}} dH + \tau \cdot h(p=\tau) \right]$$

Next, Figure 1.6 shows the effect on consumer surplus at the other end

of the marginal cost range. Here, I introduce a nonzero reimbursement rate into a region that was previously full out-of-pocket cost. To facilitate comparison later on, I use the same τ amount that was added in Figure 1.5. This changes the marginal price from $p = 1$ to $p = 1 - \tau$. The figure normalizes the area of consumer surplus at $h(p = 1)$ to zero. The gain in consumer surplus is the shaded area D. Calculating this change in consumer surplus uses the formulas above modified for the endpoints proscribing area D.

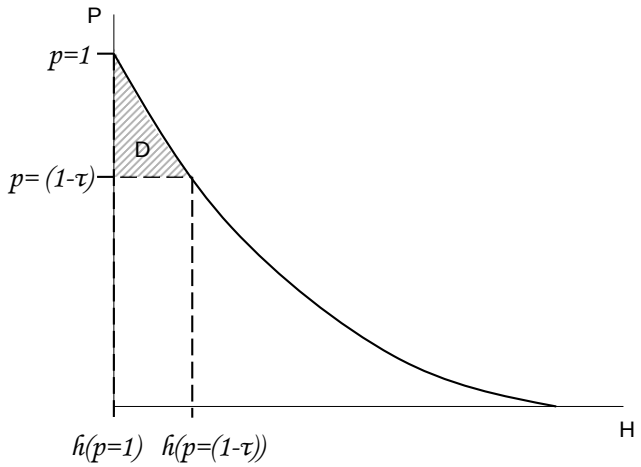


Figure 1.6: Change in consumer surplus from decreasing $p = 0$ to $p = \tau$

On the side of the insurer, because the health expenditure choice of each individual patient is the amount billed to the insurance company, change in insurer cost will take into account the out-of-pocket costs that the insurer receives and the total expenditures of the patient. Patient health expenditure

in the base scenario is $h(p=0)$, total cost in the new scenario with $p=\tau$ is $h(p=\tau)$, and total patient expenditure in the new scenario with $p=1-\tau$ is $h(p=1-\tau)$. The insurer receives out-of-pocket payments at each price level, $h(p=0)\cdot 0$, $h(p=\tau)\cdot\tau$, and $h(p=1-\tau)\cdot(1-\tau)$, respectively. In the next section, I report Producer Surplus as -*Insurer Cost*.

The equations for the net change in insurer's cost from introducing a positive marginal price of $p=\tau$ are:

$$\begin{aligned} \textit{Insurer Cost} &= \textit{Patient health expenditures} \\ &\quad -\textit{Patient out-of-pocket payment} \\ \textit{Insurer Cost}(p=0) &= h(p=0) - 0\cdot h(p=0) \\ \textit{Insurer Cost}(p=\tau) &= h(p=\tau) - \tau\cdot h(p=\tau) \end{aligned}$$

Change in producer surplus:

$$\Delta \textit{Insurer Cost} = h(p=\tau)\cdot(1-\tau) - h(p=0)$$

In the next section, changes in welfare are evaluated for a specific policy change of τ.

1.7.2 Policy Experiment Results

Using the elasticity estimates of Section 6 and the data, I solve the welfare equations listed above for a value of $\tau = 0.2$. This is an increase from $p = 0$ to $p = 0.2$ and a decrease from $p = 1$ to $p = 0.8$. The constant A is calculated using the constant elasticity of demand equation $H = AP^{\eta}$. To match A to the observed expenditures, I first take the health expenditures at $p = 0$ and $p = 1$, which are given in the data because the base scenario is a change from full coverage to full out-of-pocket costs. I then calculate the market demand equation for a series of marginal prices in between the two extremes, and choose the value of A which best replicates the beginning and end of the expenditure distribution. I use the H level and marginal price at the nonlinearity point and the elasticity estimates to back out the constant. Table 1.14 displays the resulting consumer surplus and change in insurer cost that results from the base scenario of current levels of reimbursement rates and the levels that result from each of the policy experiments described above. The welfare calculations result from an elasticity of -0.25, a conservative estimate from the range in my data, and the constant calculated from 2005 data.

In the left half of Table 1.14, the total change in welfare is negative from having patients pay a marginal price of $p = 0.2$ instead of free care. Imposing a positive price of $p = 0.2$ decreases consumer welfare by 901

Table 1.14: Policy Experiment Welfare Calculations

Increase from 0 to 0.2			
Consumer Surplus		Insurer Cost	
ΔCS	-901	ΔInsurer cost	109
ΔTotal Welfare			-792
Decrease from 1 to 0.8			
Consumer Surplus		Insurer Cost	
ΔCS	3	ΔInsurer cost	-2
ΔTotal Welfare			1

units. Change in Insurer Cost only considers the change in the insurer's cost compared to the patient's out-of-pocket payments. After introducing a marginal price increase, the total change in producer surplus is a gain of 109 units. Overall, the change in welfare is negative, and this is mainly driven by the large decrease in consumer surplus.

The right half of Table 1.14 shows the results from softening full out-of-pocket costs with $p = 1$ to a positive reimbursement rate with marginal price $p = 0.8$. Because these changes occur at the narrow portion of the market demand curve, the magnitude of the welfare changes is much smaller. Decreasing the out-of-pocket cost to a patient increases consumer surplus by 3 dollars. Correspondingly, because demand response was small to the decreased marginal price, the loss in producer surplus is also small, at 2 dollars.

The total effect of the two changes is the sum of both halves of Table 1.14.

The overall negative welfare impact is mostly driven by the large demand response to an increase in marginal price above $p = 0$. Consumer welfare is most sensitive in this scenario by changes at low levels of out-of-pocket cost.

The fact that patient expenditure is more sensitive to changes in reimbursement at low levels of out-of-pocket cost is relevant to the current policy reform. The most recent health care reform plan proposes taxing "gold-plated" insurance plans. These plans have patient out-of-pocket expenses at low levels. The current proposal is to tax 35 percent of a plan's value if it falls into the "gold-plated" category. Depending on the relative elasticities of the insurer's supply and the patient's decision to enroll, some portion of this 35 percent will be passed on in the form of less generous reimbursement rates. As the policy experiment above shows, the largest impact on total expenditures is on the portion with the most generous reimbursement rate. Thus, outside of any revenue motivations, the proposed tax will in fact be focused on the portion of insurance reimbursement in which expenditure decreases the most rapidly.

1.8 Moral Hazard Estimation

1.8.1 Setup

It is an artifact of history that health insurance is employer-based and tax-exempt. Critics of the tax-exemption claim that it leads employers to

offer plans that are too generous. Plans that are too generous exaggerate the magnitude of moral hazard. Moral hazard is the situation in which patients alter their expenditure behavior after they have purchased medical insurance because they are no longer paying the full cost of their care. Moral hazard arises in this framework because each patient has private information on their illness severity, θ. Insurers do not have the private information on θ. Plans with very high reimbursement rates are cited as inducing greater overconsumption and moral hazard because the patient bears very little of the cost of his health care, and the insurer has no way of monitoring the private information of the health shock.

To think consider how to measure overconsumption, think of the "free money" that is given to employers: the taxes that are not taken out. This "free money" can only be spent on health care because it is tied to health insurance tax exemptions. However, what if the "free money" were just given to patients to spend however they like? If patients choose to spend some of it on other things, then there is not only moral hazard but also some deadweight loss. Subsidies directed toward health care may cause overconsumption. If these subsidies exist with the only goal of increasing patients benefit, the government can dispense the same budget but increase patient benefit by not tying the subsidy to health expenditure.

I conduct a counterfactual experiment to measure the extent of this moral hazard and deadweight loss. First, the estimation Equation 1.4.6 in

Section 4 gives the amount of health care spending an individual with illness severity θ_i would spend on either side of the nonlinearity. So we have the health expenditure for reimbursement rate $r' = 1$ and $r' < 1$. Denote the optimal choice of expenditure at a given reimbursement rate r' as $h^*_{r'}$. A patient's utility at the choice of $h^*_{r'=1}$, full reimbursement, is then:

$$\bar{U} = u(h^*_{r'=1}, \theta) + c$$

Now consider the social planner's problem. If we impose less than full reimbursement, that is $r' < 1$, each patient's resulting utility must be less than or equal to $\bar{U}$, the utility achieved previously with full reimbursement. If we want to bring each patient's utility back up to the level of $\bar{U}$, we can give each patient a transfer T into their budget constraint, which the patient may use for health expenditures or outside good consumption. The patient's maximization problem would then be to choose a $h^*_{r'<1}$ for the new reimbursement schedule of $a(h) < 1$, but with an additional term in his budget constraint representing a income transfer T. The maximization problem is then:

$$\begin{aligned} \max_{h_{r'<1}} U &= u(h^*_{r'<1}, \theta) + c \\ \text{s.t.} \qquad & c + h(1 - r') \leq y + T \end{aligned}$$

Choose a utility function which satisfies the assumptions listed in Section 3 where the patient maximizes his budget constraint. Then the patient's utility maximization problem is:

$$U = y + T - h(1 - r') + u(h^*_{r'<1}, \theta) \tag{1.8.1}$$

The social planner would like to make the patient's utility the same at this new reimbursement rate $r' < 1$ and transfer system as it was before with full reimbursement. That means a utility level of $\bar{U}$. In equation 1.8.1, the following variables are data: $h^*_{r'<1}, y, r'$, and θ. The unknowns in the equation are T, the utility function parameters, and g. I discuss the utility function parameters and g in the next section. Equation 1.8.1 can then be used to solve for the social planner's T that will make the utility under the new scenario equal to $\bar{U}$. Utility parameters can be recovered by using the general utility function laid out in Section 3, and the resulting optimal decision rule, Equation 1.3.6. We can estimate the utility function parameters in the data by solving each patient's decision rule given the side of the discontinuity and the choice of health care.

The social planner's transfer T represents the amount the patient would accept to be indifferent between the two reimbursement scenarios. In the second scenario, the amount of reimbursement for the patient was reduced. Therefore, we can calculate the cost savings that resulted by the combination

of reduced expenditure due to increased out-of-pocket expense, as well as the increased payments by patients for each unit of health expenditure. The cost savings between the two scenarios is the total cost of the expenditure at full reimbursement minus the total cost of the expenditure at less-than-full reimbursement times the payments made by patients. Denote the cost savings as S and the formulas is thus:

$$\text{Cost Savings } = S = h^*_{r'=1} - h^*_{r'<1}(r' < 1) \tag{1.8.2}$$

Now there are two items to compare, the utility-equalizing transfer T and the cost savings S from lowering reimbursement rates. If the cost savings are greater than the transfer necessary to make the patient indifferent, then the potential exists for the government to reduce the size of the reimbursement subsidy and still make patients better off. If the $S < T$ then this also shows that patients were purchasing more care under the generous reimbursement scenario than they would have if the subsidy were not tied to health expenditures. Thus, the difference between S and T also gives a measure of the amount of moral hazard.

The final step of the moral hazard estimation is to estimate the amount of deadweight loss that the subsidies create by inducing generous insurance plans which cause patients to overconsume in health care versus the spending on an outside good they would have otherwise preferred. To measure this

loss in utility, I perform the following experiment. What if, at the less generous reimbursement with T scenario, the patient is forced to buy the original amount of health expenditure he would buy with full reimbursement. This means that some of his transfer T is now diverted from outside good consumption and tied to health expenditure instead. I then measure the resulting loss in utility.

1.8.2 Moral Hazard Estimation Results

Figure 1.7 shows the difference between the savings in health expenditure S and the compensating transfer T. If the cost savings of lowering reimbursement rates is greater than the amount of the compensating transfer, the size of this difference shows the magnitude of the inefficiency of health expenditures in this region. Each of the three histograms illustrates the distribution of $S - T$ in each year. The distribution is slightly different in each year because the free care region changed. For 2002, the $S - T$ value ranges between \$200 and \$300. As a percentage of the total spending, in 2002 the window of expenditures for the counterfactual was between \$500 and \$1,000. So the $S - T$ term is around a third of the total value of spending. For 2003, the $S - T$ value varies from \$180 to \$300. In 2003, the area of the counterfactual was between \$750 and \$1250. Finally, for 2005, the $S - T$ value varies between \$400 and \$500. The area of the counterfactual in 2005 was between \$600 and \$1,000.

These three histograms show that the magnitude of inefficiency of providing full reimbursement is large. The fact that the $S-T$ terms are such a large proportion of the overall spending during the full reimbursement stage indicates that patients would currently be willing to accept less composite good consumption than its equivalent in lost health expenditures. The loss in health expenditure does not have to be fully replaced by composite good consumption. That is, the marginal increase in utility of additional units of health care is much lower than if the patient could have been given the same income to spend on the composite good.

1.9 Conclusion

This paper addresses two goals. The first goal is measuring the elasticity of health expenditures to a patient's marginal price within an insurance plan. The second goal is to calculate welfare effects of policy experiments to inform the debate on changing patients' out-of-pocket costs and offer an estimate as to the extent of moral hazard inefficiencies in a full reimbursement system.

This paper presents an estimation method that draws upon the advantages of Regression Discontinuity design. The method isolates patients of similar illness severity levels, but who face different prices. The observed distribution of health expenditures reveals the underlying distribution of unobserved heterogeneity in illness severity. This empirical inverse cdf of

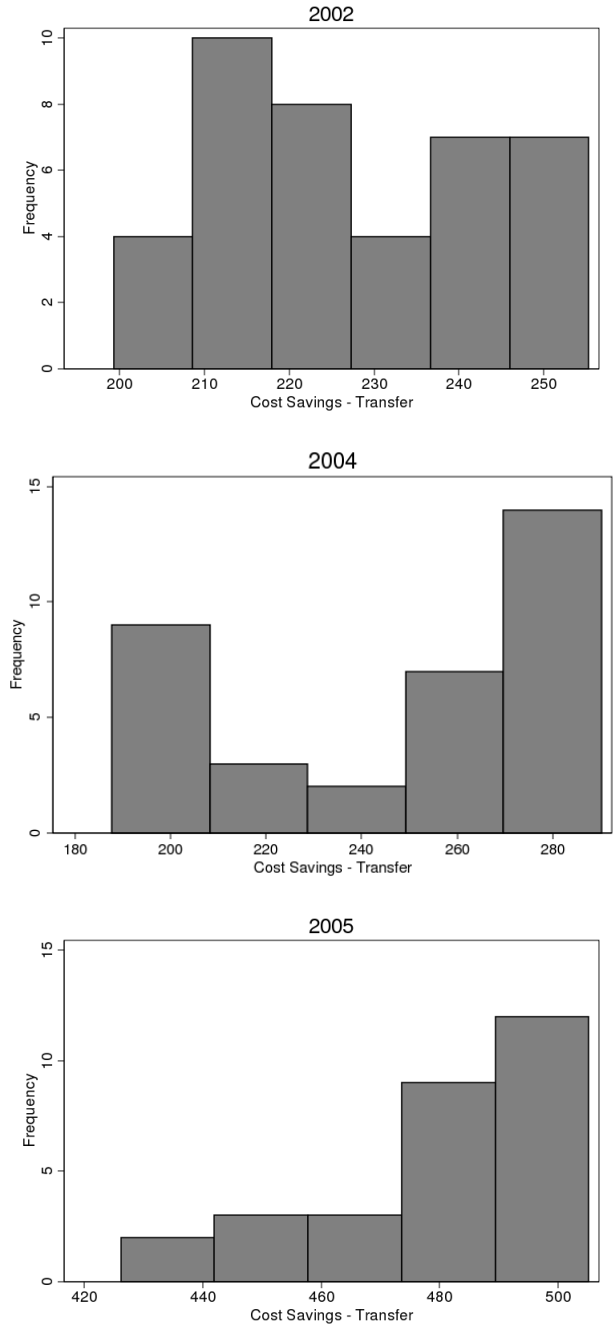


Figure 1.7: Savings in health expenditure S - Compensating transfer T

health expenditures identifies the slope relationship between increasing illness severity and resulting health expenditure choices. This slope is calculated on each side of a discontinuous change in marginal prices, a feature of the insurance plan's nonlinear reimbursement schedule. The estimation uses a flexible specification of a local linear regression. As the slope of these local linear regressions between health expenditures and unobserved illness characteristics approaches the discontinuity from each side, the underlying unobserved illness characteristics become similar. Thus, at the limit toward the discontinuity point from each side, the difference in the slope identifies the patient response to a change in prices. The resulting elasticity estimates are consistent across all three years of study, approximately -0.26. The elasticity estimates apply to patient behavior in employer-based insurance over ranges of expenditures less than $1,000.

The estimation method above was used for two policy counterfactuals. The first policy experiment reduced the severity of the drop in a coverage gap, such as the Medicare Drug Plan "doughnut hole." The constant elasticity which was calculated over the range of expenditures contributes to the conclusion that patient expenditures are very responsive to small changes from full reimbursement to a positive marginal cost. The second policy counterfactual calculated a compensating transfer for patients if a policy changed from full reimbursement to positive reimbursement. I find the compensating transfer that would have to be paid is much greater than the resulting cost

savings in providing health care.

The major contributions of this paper are threefold. First, I create a flexible estimation method that can be used in datasets with non-linearities. This is important because non-linearities are present in many important policies that have recently been debated, including deductibles, Medicare Parts A and B and the "doughnut hole" of the Medicare Drug Plan. A framework to examine these plans without restrictive behavioral assumptions did not previously exist. Secondly, I produce an elasticity estimate applicable to patients in employer-sponsored insurance plans, the most common type of insurance coverage in the U.S. This estimate updates previous studies using experimental datasets, and applies to a population affected by recent health reform. Finally, using this framework, I calculate the welfare effects and a measure of moral hazard stemming from recent health reform initiatives to increase cost sharing for patients.

Chapter 2

Moral Hazard, Adverse Selection and Health Expenditures: A Semiparametric Analysis

2.1 Introduction

A large theoretical literature predicts that adverse selection and moral hazard may generate inefficient outcomes in insurance markets due to asymmetric information (see e.g., Arrow 1963, Pauly 1968, Akerlof 1970, Zeck-

hauser 1970, Spence and Zeckhauser 1971, Spence 1973, Pauly 1974, Rothschild and Stiglitz 1976, Wilson 1977, 1980). Rothschild and Stiglitz (1976) develop a model in which adverse selection is present because individuals have private information about their health status. Similarly, Pauly (1968) shows that moral hazard may be present in health insurance markets because consumers do not bear the full cost of health care expenditures.

The predictions of theoretical models of insurance markets can depend on whether adverse selection or moral hazard is more important. Because of the complexity of insurance markets, theoretical models frequently emphasize one of these distortions at the expense of the other. In addition, optimal policy addressing inefficient outcomes depends crucially on the relative importance of these distortions. The most useful empirical analysis of asymmetric information must be able to identify the importance of moral hazard and adverse selection, but also to separately identify the two effects in order to guide both theory and public policy.

It is well recognized that it is empirically difficult to distinguish between moral hazard and adverse selection and consequently there is little consensus on which of these two sources of inefficiency is more important. A common method to detect asymmetric information is to examine the correlation between risk outcomes and a measure of the generosity of a contract. In the context of health insurance, Cutler and Zeckhauser (2000) review an extensive literature that finds evidence of adverse selection based on the positive

correlation between generosity of the insurance contract and adverse outcomes, and moral hazard based on the coinsurance elasticity of the demand for medical care. However, as pointed out by Chiappori and Salani (2003), among others, under moral hazard the generosity of the contract will lead to adverse risk outcomes while under adverse selection the causality is reversed, leading to observational equivalence between the two hypotheses.[1] [2]

In this paper, we propose an approach to assess the importance of moral hazard and adverse selection in health insurance markets. This approach establishes a model of individual choice given insurance offerings and uses semiparametric estimation to measure separately the magnitude of moral

[1] A test of asymmetric information based on the positive correlation between the generosity of the contract and adverse risk outcomes may also breakdown if there is heterogeneity in risk preferences (De Meza and Webb 2001). For example, if individuals who are highly risk averse also put more effort in to lowering their risk, and vice versa those who are less risk averse take fewer precautions, then there could be a negative correlation between the generosity of the contract and risk outcomes. Evidence of heterogeneity in risk preferences has been found by Finkelstein and McGarry (forthcoming) in the long term care insurance market and Cohen and Einav (2005) in the automobile insurance market. Chiappori, Jullien, Salani and Salani (forthcoming) develop a non parametric test to detect the presence of asymmetric information based on the correlation between the generosity of the contract and risk outcomes that addresses this limitation. However their test does not distinguish between adverse selection and moral hazard. Finkelstein and Poterba (2006) develop a test for adverse selection that avoids the limitation of heterogeneity in risk preferences. Their test is based on using data on observable characteristics of individuals that are correlated with the outcomes but are not used by insurers in pricing contracts.

[2] Abbring, Chiappori, Heckman and Pinquet (2003) have suggested that one could exploit the dynamic consequences of experience rating in insurance markets to distinguish between adverse selection and moral hazard. However, U.S. health insurance markets are regulated to restrict experience rating, which precludes this proposed empirical strategy. Another alternative would be to use field experiments to investigate the presence of asymmetric information. Karlan and Zinman (2005) provide an example of such an experiment in credit markets in South Africa. However the feasibility of implementing such experiments in the context of health insurance, especially in the U.S., is questionable.

hazard and the sorting across plans due to adverse selection. There is a growing body of empirical work on the structural estimation of models of medical utilization and health insurance choice (e.g., see the important work of Cameron et al. 1988, Gilleskie 1998, Harris and Keane 1999, Cardon and Hendel 2001, Vera-Hernandez 2003, Blau and Gilleskie 2003, Khwaja 2001, forthcoming). However, we differ from this earlier literature in three ways. First, we allow for both adverse selection and moral hazard due to asymmetric information in the estimation procedure. Second, earlier work uses methods from discrete choice estimation and relies on parsimoniously specified parametric models. However, akin to the empirical literature on auctions, our estimation strategy is semiparametric. This is important because theory provides little guidance about which parametric distributions for latent health shocks are a priori most plausible. Semiparametric estimation also is a major contribution because it allows us to specify plans' reimbursement schedules flexibly. Our estimation allows for common characteristics such as deductibles and copays, which could not be captured by more restrictive specifications in previous literature.

Most notably, our method improves on previous approaches because it only necessitates a single, relatively weak identification assumption. A major critique of previous structural approaches is that they rely on strong identifying assumptions. In this paper, the sole identifying assumption is that health shocks are fully independent of time. That is, the percentiles of

the underlying health status distribution are invariant across the three years of our data. This assumption is reasonable given a large enough population of study. In particular, our data is from the largest employer in the state of Minnesota which has retained the top position for many years with a stable population of employees over our period of study. Additionally, we use consecutive years in our estimation, which further supports the weakness of this identifying assumption by limiting the possibility of long-term changes in the local population or restructuring of the employee base.

An important contribution of this paper is furthering analysis of patient behavior by using a modeling framework. Einav, Finkelstein, and Levin (2010) describe how comparative claims testing cannot separate out adverse selection and moral hazard, necessitating a modeling framework in the absence of experimental data. A model of patient preferences is important as opposed to so-called reduced form approaches in order to allow testing for efficiency of markets and the welfare impact of market interventions. Beyond documenting the presence of moral hazard and adverse selection, a modeling framework permits counterfactuals allows for policy recommendations. Additionally, the structure of health insurance in particular necessitates a more structured approach than reduced form methods. This is because the price for health care services is often based on the amount of health care consumed, for example deductibles or copays. Thus, any basic regression analysis between health care quantities and incentives will face endogeneity

bias.

The estimation method proceeds in the following steps. We first lay out a model of individuals' choices given a menu of insurance plan offerings. The structural model is consistent with the theoretical models proposed by Spence and Zeckhauser (1971) and Blomqvist (1997). Risk averse agents have preferences over a composite commodity and health status, and maximize utility subject to a budget constraint. Agents have private information about their latent health status which is unobserved to the insurer leading to adverse selection. The agents do not pay the full costs of their health care coverage and therefore face a moral hazard problem. We estimate the model using a unique data set that includes confidential information about employee health insurance characteristics, health status and medical expenditures and from a large self-insured employer in Minnesota. Using this data, we estimate risk parameters semiparametrically and use these parameters to back out the latent health status distribution. Given the estimated health status distribution, utility parameters, and observed consumption choices, we are able to estimate moral hazard with a counterfactual replicating a social planner's optimal consumption allocations. The final step in the estimation uses the estimated latent health distribution across all plans and tests for sorting within this distribution into different plan offerings.

To foreshadow our findings, our model generates many results that are consistent with the existing empirical literature. First, our estimates of

risk aversion are consistent with much of the previous literature. Second, consistent with the literature we find substantial evidence of moral hazard or "overconsumption" of medical care (e.g., Manning et al 1987, Newhouse 1993) Third, we propose a nonparametric test for adverse selection, and find that individuals are indeed sorting across different insurance categories based on their latent health status.

Our research is novel in that it develops a tractable estimation procedure under minimal parametric assumptions to simultaneously examine adverse selection and moral hazard in health insurance contracts. This method is also important as it provides a framework for similar analysis in other contexts, especially with cross-section data, where distortions exist due to asymmetric information. The rest of the paper is organized as follows. The model is discussed in Section 2.2 and data in Section 2.3. Details of the estimation procedure are provided in Section 2.4. The estimation results including analysis of moral hazard and adverse selection are discussed in Section 2.5, and Section 2.6 concludes.

2.2 Model

In this section, we specify a model of consumer demand for health services. Our framework is motivated by the models of Spence and Zeckhauser (1971), Bloomqvist (1997) and Cardon and Hendel (2001). In the model,

consumers, or their physicians acting to maximize consumer utility, choose the amount of health services m and consumption of a composite commodity c subject to a budget constraint. This budget constraint requires that out-of-pocket expenses for m plus the value of c must be less than or equal to income y.

Our model differs from the standard neoclassical model of consumer choice in two important respects. First, the optimal choice of m and c depends on a patient's latent health status, which we denote as the scalar θ. We interpret θ as a preference shock for health services, that is, the higher the value of θ, the higher the utility from health services. While θ is observed to the consumer, it is not observed by the insurer. As a result, our model allows for adverse selection, which plays a central role in the economic analysis of health insurance and insurance markets more generally.

Second, the budget constraint in our model is more complicated than the neoclassical model of the consumer. Health insurance plans introduce nonlinearities in the budget constraint through features such as deductibles and coverage gaps. Also, the consumer's out-of-pocket expenses will not be known with certainty at the time that m is chosen. In our framework, we allow this out-of-pocket expense to be stochastic, generalizing previous research which typically assumes that the consumer's costs of health care are known at the time m is chosen.

In this framework, we approach the question of moral hazard and adverse

selection from the patient's problem for several reasons. First, consumer characteristics are important to the question of sorting between insurance plans and insurance design. Second, although we could also incorporate the choice of health care as a joint decision process between a patient and her doctor, there is strong evidence that patients' health care choices do respond to the patient's health insurance incentive structures (See for example, Manning et al. 1987, Keeler and Rolph 1988, Kowalski 2009, Marsh 2010.) Finally, it is important to note that our identifying assumptions do not depend on whether the choice of health care was modeled as a single or a joint decision, rather the identification depends on a population-wide health status distribution.

We develop the components of our model in detail below in the next subsections.

2.2.1 Consumer Preferences

In our model, the consumer's utility function can be written as:

$$U(c, m; \theta, \gamma) = (1 - \theta)\frac{c^{1-\gamma_1}}{1-\gamma_1} + \theta\frac{m^{1-\gamma_2}}{1-\gamma_2}$$

The consumer's utility depends on the consumption of m and c. We allow for the consumption to be additively separable in these two terms for analytical convenience. However, as will be made clear below, we easily allow

a non-separable specification. The parameter θ lies between [0,1] and indexes the weight that the consumer places on consumption of health services m and non-health services c. If θ is close to one (zero), the consumer receives more utility from c (m), all else held fixed.

Consumers in our environment are heterogenous in the sense that θ varies across individuals. We shall let $g(\theta)$ and $G(\theta)$ denote the density and cdf of the health shock across individuals in our data set. In principal, we could let g depend on observed covariates associated with health status such as age, education level or income. However, the ex ante choice of such covariates involves the risk of misspecification. As we shall discuss in our section on estimation, we do not need to specify this part of the model. However, in our results section, we shall regress θ on risk factors in order to ask whether our model has generated reasonable results.

We allow for constant relative risk aversion in c and m. The parameters γ_1 and γ_2 describe the consumer's attitude towards risk. As we shall describe below, this is important in our framework because the out-of-pocket expenses associated with m are stochastic. In principal we could allow for heterogenous preferences in γ_1 and γ_2 and our discussion below suggests that such preferences could be identified. However, this would complicate our estimation procedure and we abstract away from this issue, leaving it to future work.

2.2.2 Budget Constraint

In our model, consumers are constrained by the fact that out-of-pocket expenditures on health plus consumption of all other goods must be less than or equal to income. In our model, the price to the consumer of m depends on two things. The first is p_j, the fixed premium of the health plan j. In our data, there will be three plans. The largest is a Health Maintenance Organization (HMO) and the second and third are Preferred Provider Organizations (PPO). We shall describe these plans in more detail below.

We take j as exogenously given to the consumer. We could imbed our framework in a richer model that includes the choice of a health plan in an earlier stage. However, as we shall discuss below, the optimal choice of m will be sufficient to identify our model parameters and we abstract from this complication. Also, the choice of j does not influence the consistency of our estimates under our assumptions about the data generating process. However, in our results section, we shall use our estimates of the model to ask how consumers with different values of θ sort across plans j and if their observed sorting is consistent with leading theories of adverse selection.

The second part of our model that determines the consumer's cost is the reimbursement rate determined by insurance plan j, which we label as the scalar a_j. If a consumer chooses medical expenditure of m, the insurer

will cover $a_j m$ of these expenses. As a result, the consumer must pay for $m(1 - a_j)$ out-of-pocket.

The patient's budget constraint is:

$$c + m(1 - a_j) \leq y - p_j$$

The budget constraint specifies that the consumer's total expenditure on the composite commodity plus health, $c + m(1 - a_j)$, must be less than her income after deducting medical premiums, $y - p_j$.

A difficulty the consumer faces is that a_j will be uncertain at the time that m is chosen. Medical plans are long and typically quite complicated documents that are frequently written by insurers, their executives and attorneys. It is unlikely that a typical consumer invests the resources to understand these medical plans to discover what is covered in all states of the world. Furthermore, the final determination of reimbursement is done by administrators and is often the outcome of a negotiation between the health provider and the insurer.

Obviously, the determination of a_j is a complicated object. What is important to model is the determination of a_j from the perspective of the consumer since we are ultimately interested in consumer demand. The other issues raised above, while interesting, are secondary to our research question.

Therefore, we shall model a_j as a random variable with a distribution

$f_j(a_j|m)$. This allows the generosity of the benefits to depend on the plan j chosen by the consumer. Furthermore, we allow a_j to depend on m. This is natural in the context of consumer health. For example, many plans require a patient pay a fixed fee for a doctor's visit. Plans may display features such deductibles or coverage gaps which do not reimburse a certain range of expenditures. Our framework allows for these complications by allowing the reimbursement rate to depend on m.

We shall use the observed distribution of reimbursements and flexible, nonparametric methods to identify $f_j(a_j|m)$. In our opinion, this is preferable to a strict, ex ante specification of consumer's expectations about a_j. We allow these expectations to be data driven and consistent with the standard economic assumption that consumers have rational expectations, in the sense that their beliefs about a_j must be consistent with the observed outcomes. In principal, it might be interesting to allow consumers to have biased or some other form of irrational beliefs about the determination of a_j. However, as we shall discuss below, our model is very flexible and comes close to exhausting the degrees of freedom in the data. The identification of such irrational beliefs would therefore be tenuous. There is no agreement or evidence which would provide a plausible basis for the a priori specification of how consumers bias their beliefs. As a consequence, we use the more common assumption that consumers have beliefs that are consistent with ex post outcomes.

2.2.3 Expected Utility and First Order Conditions

As discussed above, the consumer makes her choice under uncertainty. A time line for the consumer's choice is:

1. A value of θ is drawn from g
2. After the consumer observes θ, the consumer makes a choice of m
3. The reimbursement rate a_j is realized from the distribution $f_j(a_j|m)$
4. Since preferences are strictly increasing, the budget constraint binds and c is determined by the equation $c = y - p_j - m(1 - a_j)$.

Let $EU(m)$ denote the consumer's expected utility at point 2 in the time line above.

$$EU(m; p_j, y, \theta, \gamma) = \int (1-\theta) \frac{(y - p_j - m(1 - a_j))^{1-\gamma_1}}{1 - \gamma_1} f_j(a_j|m) da_j + \theta \frac{m^{1-\gamma_2}}{1 - \gamma_2} \tag{2.2.1}$$

The above expression substitutes the choice of m into the consumer's budget constraint for c. In computing expected utility, the consumer integrates over a_j using the distribution $f_j(a_j|m)$. If the realization of a_j is close to one (zero), the value of c will be larger (smaller) all else held equal. Obviously, the value of $EU(m)$ depends crucially on γ, the consumer's attitude towards risk. For example, the more risk averse the consumer is toward

uncertainty in consumption of c, we would expect her to consume less m. The utility also depends on income and premiums through $y - p_j$. For example, households with lower income are more adversely impacted by a low realization of a_j.

The optimal choice of m is determined by a first order condition that sets the derivative of $EU(m)$ in m equal to zero. Formally, we can write this condition as:

$$\begin{aligned}\frac{\partial EU(.)}{\partial m} =& \frac{\partial}{\partial m}\left[\int (1-\theta)\frac{(y-p_j-m(1-a_j))^{1-\gamma_1}}{1-\gamma_1} f_j(a_j|m)da_j\right] + \frac{\partial}{\partial m}\left[\theta\frac{m^{1-\gamma_2}}{1-\gamma_2}\right] \\ =& \frac{1-\theta}{1-\gamma_1}\int [-(1-\gamma_1)(1-a_j)(y-p_j-m(1-a_j))^{-\gamma_1} f_j(a_j|m) \\ & + (y-p_j-m(1-a_j))^{1-\gamma_1}\frac{\partial f_j(a_j|m)}{\partial m}]da + \left[\theta\cdot m^{-\gamma_2}\right]\end{aligned} \tag{2.2.2}$$

The first line of the equation 2.2.2 sets out the components of the utility effects of a marginal increase in health care expenditure. The left bracketed term shows the marginal change in a patient's utility due to a marginal change in composition good consumption c, discounted by the patient's risk coefficient on consumption. The presence of the conditional probability of reimbursement $f_j(a_j|m)$ reflects the uncertainty in reimbursement and the effect this has on the income that is available for composition good

consumption. The right bracketed term shows the marginal increase in utility due to the consumption of health care, discounted by the patient's risk coefficient on health care. Each bracketed term is weighted by θ, so that the weight on a marginal increase in health care reflects the private value the patient places on health care. The total effect on expected utility is the combined terms, which are expanded in the second line of equation 2.2.2.

2.2.4 Inferring θ from observed choices

In our estimation method described below, we shall estimate the density of health shocks g and the preference parameters γ from the observed choices. Rewriting 2.2.2 in the following way will be essential in constructing our estimator:

$$\theta = \frac{I}{I - \frac{(1-\gamma_2)}{m^{\gamma_2}}} \tag{2.2.3}$$

where

$$I = \int [-(1-\gamma_1)(1-a_j)(y - p_j - m(1-a_j))^{-\gamma_1} f_j(a_j|m) + (y - p_j - m(1-a_j))^{1-\gamma_1} \frac{\partial f_j(a_j|m)}{\partial m}] da_j. \tag{2.2.4}$$

Our estimator will be based on logic similar to that used in auction

models, especially Campo, Guerre, Perrigne and Vuong (2009). The left hand side of the above equation is the consumer's private information, θ. For a fixed γ, the remaining terms on the right hand side of this equation are potentially observable. For example, from the data, we will be able to construct an estimate of $f_j(a_j|m)$ by observing the distribution of reimbursements conditional on medical expenditures m. Also, our data set will contain information about y, a particular consumer's income level, p_j medical premiums for plan j and m the total value of consumption of medical services.

Under suitable regularity conditions, for a fixed value of γ, we shall always be able to find a value of θ that rationalizes the consumer's choice. As a result, it follows that the assumption of utility maximization alone will not be adequate to identify both g and γ. This is similar to the analysis of the risk averse auction model in Campo, Guerre, Perrigne and Vuong (2009).

Our approach to identification will be to impose additional moment restrictions. In our application, we will use the moment restriction that g does not depend on time. Our data is a three year panel on incomes and health care choices from a large Minnesota based employer. Intuitively, this restriction means that the severity of illnesses for this large population of employees does not evolve over time. Below, we shall argue that this is reasonable since the population of employees is very large and there is no reason

to expect large fluctuations in the severity of illness as reflected in g among this group within a reasonably short three year time period. However, there will be considerable variation in individual incomes as individuals experience promotions or change jobs within the organization. Also, premiums and reimbursement rates $f_j(a_j|m)$ will vary from year to year.

2.2.5 Discussion

The two primary issues in health insurance that are addressed in this paper, moral hazard and adverse selection, result from the fact that patient's health, θ, is unobserved to the insurer, yet the insurer offers a menu of different plans from which a patient may choose.

Adverse selection results when a patient with a high value of θ has the ability to choose a plan j with corresponding beneficial features, such as high reimbursement or better access to care. Because θ is private information to the patient, the insurance plan cannot correct this selection. Equation 2.2.3 will be used to back out the distribution of θ for each plan and test how these different distributions reveal adverse selection across the available plans in the dataset.

Moral hazard comes through in the model because patients pay some percentage a_j for each unit of health care m, but do not pay the full cost incurred by the insurer. Insurers establish p_j and $f_j(a_j|m)$, but once patients have chosen plan j, the insurer cannot contract the amount of care m chosen

within plan j. Moral hazard results when the patient does not bear the full cost of her care. In equation 2.2.2, a patient's marginal increase in expected utility with respect to health care must be equal to the price ratio between health care and composite good consumption. The reimbursement a_j makes the relative price of health care lower than composite good consumption, leading to moral hazard.

2.3 Data

To estimate the model we use a detailed confidential claims level dataset from a large self-insured employer.[3] The claims level data is linked to enrollment and demographic history from the employer. Because the employer is self-insured, any medical care with an associated claim is included in the dataset for every employee in the firm. Full claims data for an entire year is available for the years 2002-2004. The employer has offices in several locations and the complete claims data covers over 19,000 employees and over 39,000 total beneficiaries.

The employer offers health insurance coverage for individuals, spouse/domestic partners and families. Employees had a choice of four types of plan during 2002-2004, a traditional health maintenance organization (HMO), a preferred provider organization (PPO), a tiered network product based on care

[3]We thank Robert Town and Caroline Carlin for sharing this data with us.

systems, and a consumer driven health plan (CDHP). The HMO featured generous coverage for network physicians and hospitals, but no coverage of out-of-network care. The PPO had nominal copayments for services in-network, with lower coverage out-of-network after meeting a deductible. In the tiered network product, employees chose from three cost tiers with standardized benefits but varying premiums which changed depending on the bids submitted by each of the facilities in the tiers. The CDHP option featured an employer-funded health savings account and a high deductible.

Claims level data was aggregated to yearly sum for each beneficiary. A claim entry includes all the information necessary to process the insurance payment. Each claim includes total approved health care spending, total reimbursement, and cost variables such as the coinsurance amount, copayment, amount of deductible used. The claim also includes information such as primary and secondary diagnosis codes, procedure codes, type of treatment facility, and type of provider. Cost and spending amounts are aggregated for each beneficiary over the year for a precise measure of total spending by the employer and total out-of-pocket cost to the beneficiary.

Claims data is linked to employee information to include demographic information. Age and gender are taken directly from the claims files. Salary was imputed based on birthdate, gender, home zipcode, work location, and job classification using a separate file provided directly from the employer. Plan characteristics such as premium, employee contribution, and cost struc-

ture were obtained directly from the employer's enrollment materials.

For the empirical analysis, we focus on the three most populous plans which comprise over 80 percent of enrollment. These plans are the HMO plan, hereafter referred to as "HP", and two of the plans in the tiered network product, referred to as "PC2" and "PC3". PC2 and PC3 have similar cost sharing parameters, however PC2 has a lower cost network of providers than PC3. We use only employees enrolled under single coverage, and only those employees with continuous enrollment over the entire year. Table 2.1 displays the enrollment statistics for each year and plan for single coverage full-year enrollment. Over the three-year period, the dataset has over 14,000 single-coverage enrollees. The HP plan has the largest enrollment by far, with over 80 percent of enrollees each year. Enrollments in the PC2 and PC3 plans are between a low of 230 up to 539.

Table 2.1: Plan Total Enrollees

Plan	**2002**	**2003**	**2004**	**Total**
HP	4,032	4,126	3,845	12,003
PC2	483	469	388	1,340
PC3	523	539	230	1,292
Total	5,038	5,134	4,463	14,635

The final variables included in the dataset are a current health status proxy and a measure for the probability of high future health costs. These two variables are created for each individual enrollee using the Johns Hopkins University ACG Case-Mix System (v6), which was developed by the

Health Services Research and Development Center. This is a commercial algorithm used to predict future illness severity and health care spending. Individual diagnosis code combinations are incorporated into clinical assessments and regression fitting to produce a summary measure. The health proxy variable is constructed around a national average of 1.0 with higher values indicating greater illness severity. The diagnosis codes and health status proxy are then used to create a measure of probability of high costs next year.

Table 2.2 displays summary statistics for each of the three plans for total health spending, total reimbursement, salary, age, the health status proxy, the probability the enrollee will be high cost next year, and percentage female. The average health care spending is less than $5,000 in all three of the plans. The tiered plans have higher average health care spending than the HMO plan. For each plan, the total reimbursement is slightly less than the total health spending, which reflects the enrollees' out-of-pocket costs.

Comparing the demographic variables across the plans, there is preliminary evidence of the existence of adverse selection. The HP HMO plan's enrollees are younger than PC2 and PC3, and the average of the health status proxy variable is also lower for the HMO plan. The HMO plan enrolls a higher percentage of males than the other plans and has the lowest average salary. While these differences in demographic characteristics do not prove adverse selection on unobservable illness levels, they offer evidence

that plans do vary in observable characteristics.

Table 2.2: Summary Statistics by Plan

Plan	Variable	Mean	Std. Dev.	p25	p75	N
HP	Total health spending	$2,845	$7,459	$306	$2,710	12,003
	Total reimbursement	$2,659	$7,139	$267	$2,470	12,003
	Salary	$42,972	$21,180	$30,659	$50,000	12,003
	Age	41.7	12.0	31.0	52.0	12,003
	Health status proxy	1.4	2.8	0.2	1.4	12,003
	Pr. future high cost	0.08	0.09	0.04	0.09	12,003
	Indicator for female	0.59	0.49	0.00	1.00	12,003
PC2	Total health spending	$3,964	$7,559	$774	$4,402	1,340
	Total reimbursement	$3,657	$7,555	$643	$3,972	1,340
	Salary	$45,946	$21,533	$32,802	$52,008	1,340
	Age	46.4	11.5	38.0	55.0	1,340
	Health status proxy	2.2	3.2	0.5	2.9	1,340
	Pr. future high cost	0.12	0.11	0.05	0.14	1,340
	Indicator for female	0.73	0.44	0.00	1.00	1,340
PC3	Total health spending	$4,700	$9,243	$864	$5,023	1,292
	Total reimbursement	$4,449	$9,151	$758	$4,639	1,292
	Salary	$50,357	$26,795	$33,615	$57,811	1,292
	Age	47.0	10.7	39.0	55.0	1,292
	Health status proxy	2.5	3.8	0.5	3.4	1,292
	Pr. future high cost	0.12	0.12	0.05	0.14	1,292
	Indicator for female	0.66	0.47	0.00	1.00	1,292
Total	Total health spending	$3,111	$7,664	$374	$3,029	14,635
	Total reimbursement	$2,909	$7,397	$328	$2,734	14,635
	Salary	$43,896	$21,874	$31,210	$50,898	14,635
	Age	42.6	12.0	32.0	52.0	14,635
	Health status proxy	1.6	3.0	0.2	1.5	14,635
	Pr. future high cost	0.09	0.10	0.04	0.10	14,635
	Indicator for female	0.61	0.49	0.00	1.00	14,635

Table 2.3 displays the health status proxy summary statistics for each year. The average value of the health status proxy over all plans is stable across the three years in the sample, at approximately 1.6. The other descriptors of the distribution are also consistent from year to year, with an

average standard deviation of 2.96 and the same values for the 25th and the 75the percentile in each year. We use this consistency across years in the underlying health status to construct our estimator, which is described in the next section.

Table 2.3: Health Status Proxy by Year

Year	**Mean**	**Std. Dev,**	**p25**	**p75**	**N**
2002	1.64	3.12	0.20	1.49	5,038
2003	1.60	2.81	0.20	1.49	5,134
2004	1.57	2.92	0.20	1.49	4,463
Total	1.60	2.96	0.20	1.49	14,635

To account for the skewed distribution of expenditures often found in health care data, we further restrict the data sample for estimation. First, we drop enrollees with zero yearly expenditures. This amounts to approximately 12 percent of the HP sample in each year, 7 percent of the PC2 sample, and 4 percent of the PC3 sample. Second, we drop enrollees whose expenditures fell in the top 20 percent of the entire sample of expenditures over all three plans. Finally, a very small number of enrollees, less than 1 percent in each year, were dropped if their out-of-pocket expenditures were larger than their salaries. Table 2.4 displays the resulting final estimation sample.

Table 2.4: Estimation Sample Size

Plan	**2002**	**2003**	**2004**	**Total**
HP	2,690	2,791	2,631	8,112
PC2	351	341	295	987
PC3	392	392	174	967
Total	3,433	3,533	3,100	10,066

2.4 Estimation

We use a semiparametric estimation strategy to identify the underlying utility parameters of the consumer's utility function and the resulting distribution of latent health status variables, θ. The estimation is conducted on over 10,000 enrollee-year observations that remain from the data selection process described above. An individual observation is a full-year single coverage enrollee i, in a plan j, for a given year t. The optimal decision rule, Equation 2.2.2, gives the formula for an individual θ, Equation 2.2.3. We estimate an empirical version of this equation. Components which come directly from the data are health care spending m_{ijt}, income y_{ijt}, and premiums p_{jt}. Health care spending and income are unique to an individual i in plan j in a given year t. Premiums are the same for all individuals in a plan j in a given year t. Components which are estimated from the data are $\hat{f}_{jt}(a_{jt}|m_{jt})$, $\frac{\partial \hat{f}_{jt}(a_{jt}|m_{jt})}{\partial m_{jt}}$, and $\hat{\gamma}_1$, $\hat{\gamma}_2$. The estimated version of Equation 2.2.3 is as follows:

$$\hat{\theta} = \frac{\hat{I}}{\hat{I} - \frac{(1-\hat{\gamma_2})}{m_{ijt}^{\hat{\gamma_2}}}} \tag{2.4.1}$$

where

$$\hat{I} = \int [-(1-\hat{\gamma_1})(1-\hat{a}_{jt})(y_{ijt} - p_{jt} - m_{ijt}(1-\hat{a}_{jt}))^{-\hat{\gamma}_1} \hat{f}_{jt}(a_{jt}|m_{jt}) + (y_{ijt} - p_{jt} - m_{ijt}(1-\hat{a}_{jt}))^{1-\hat{\gamma}_1} \frac{\partial \hat{f}_{jt}(a_{jt}|m_{jt})}{\partial m_{jt}}] da_{jt} \tag{2.4.2}$$

Estimation of Equation 2.4.1 proceeds in three steps:

1. Estimate conditional reimbursement distributions $\hat{f}_{jt}(a_{jt}|m_{jt})$ and $\frac{\partial \hat{f}_{jt}(a_{jt}|m_{jt})}{\partial m_{jt}}$.
2. Given $\hat{f}_{jt}(a_{jt}|m_{jt})$ and $\frac{\partial \hat{f}_{jt}(a_{jt}|m_{jt})}{\partial m_{jt}}$, find $\hat{\theta}$ in terms of $\hat{\gamma}_1$, $\hat{\gamma}_2$.
3. Solve for utility parameters $\hat{\gamma}_1$, $\hat{\gamma}_2$ using GMM.

2.4.1 Estimating Conditional Reimbursement Distributions

In the first step, we nonparametrically estimate the conditional reimbursement distributions for each plan in each year. First, the ex-post realized reimbursement as a percentage of total health spending is calculated for every enrollee. The empirical realized reimbursements are used to calculate the joint probability of a given reimbursement percentage and a given level

of health spending, $\hat{f}_{jt}(a_{jt}, m_{mt})$. The bandwidth was chosen using the optimal bandwidth rule of thumb suggested by Bowman and Azzalini (1997). To estimate the conditional distribution, we then calculate the health spending distribution $\hat{f}_{jt}(m_{jt})$ using the same bandwidth. The estimated conditional distribution $\hat{f}_{jt}(a_{jt}|m_{jt})$ is then the ratio of the joint distribution of reimbursement and spending over the distribution of spending. That is,

$$\hat{f}_{jt}(a_{jt}|m_{jt}) = \frac{\hat{f}_{jt}(a_{jt}, m_{jt})}{\hat{f}_{jt}(m_{jt})}$$

The conditional distribution is calculated for a grid that breaks reimbursement percentages into 128 categories, as well as the same number of health spending categories. Both more and fewer categories were used, with very little change in the resulting estimates. The conditional distributions $\hat{f}_{jt}(a_{jt}|m_{jt})$ are displayed by plan for each year in Figure 2.1, Figure 2.2, and Figure 2.3. Across all plans and years, the lowest reimbursement percentages occur at the lowest level of health spending, and vice versa. The HP plan has the lowest out-of-pocket pricing schedule, so the conditional probabilities corresponding with over 90 percent reimbursement are higher in HP than the other two plans. Finally, we obtain $\frac{\partial \hat{f}_{jt}(a_{jt}|m_{jt})}{\partial m_{jt}}$ by applying an approximate derivative to all grid points of the conditional distribution.

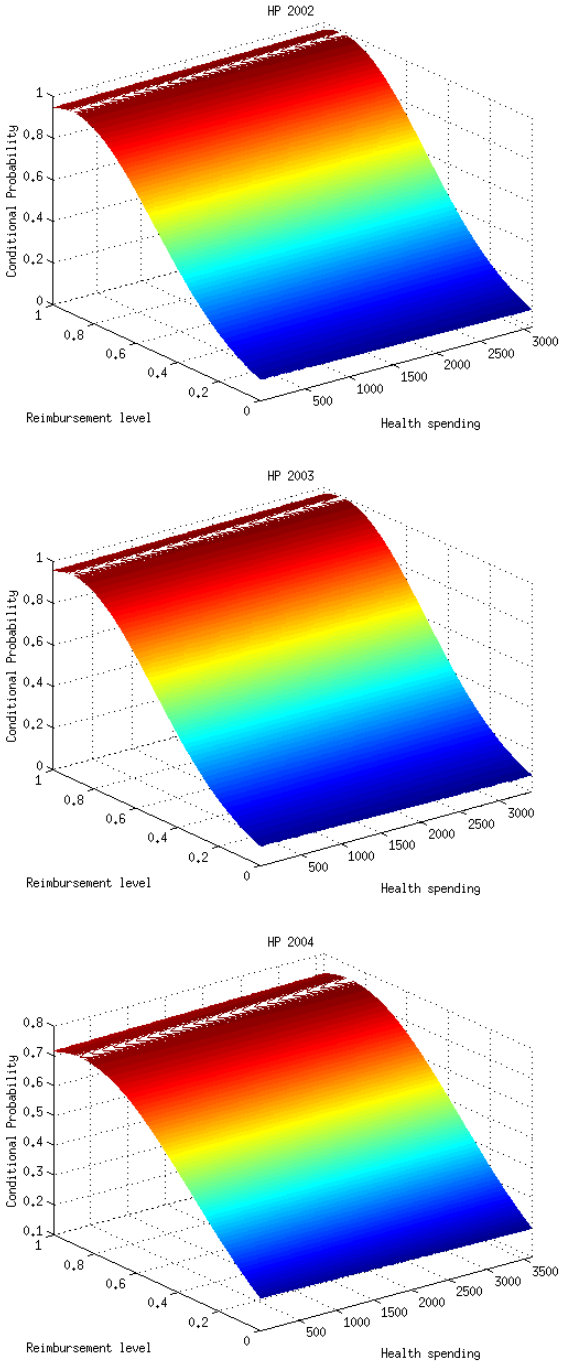


Figure 2.1: Conditional Reimbursement Distributions, HP Plan 2002-2004

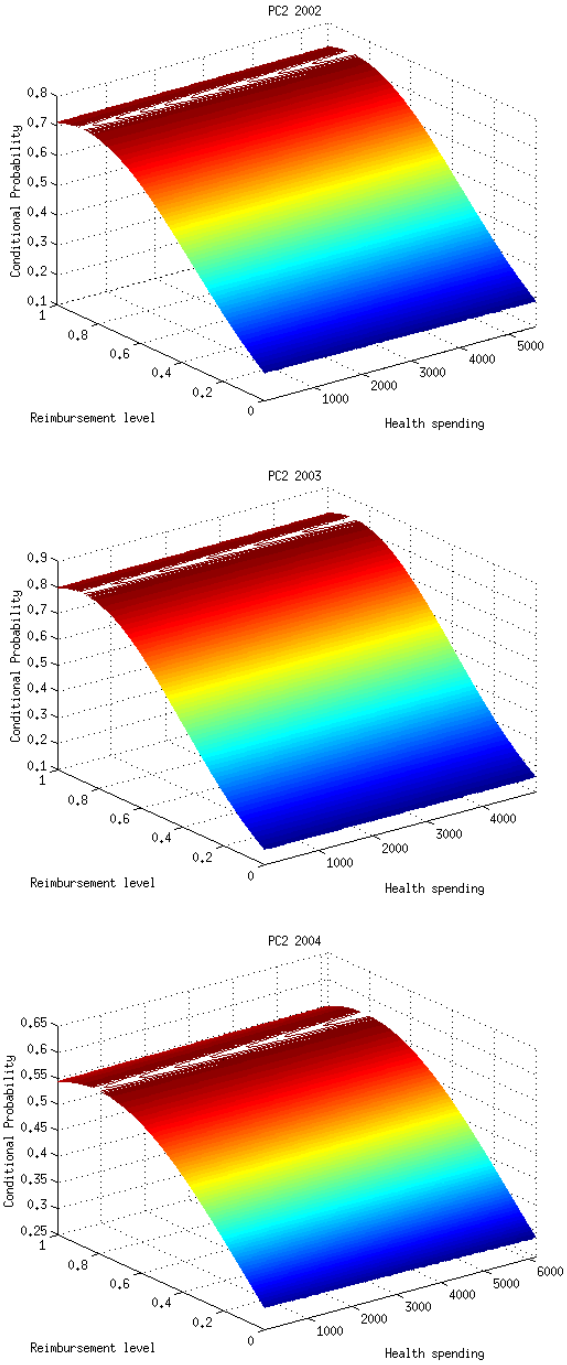


Figure 2.2: Conditional Reimbursement Distributions, PC2 Plan 2002-2004

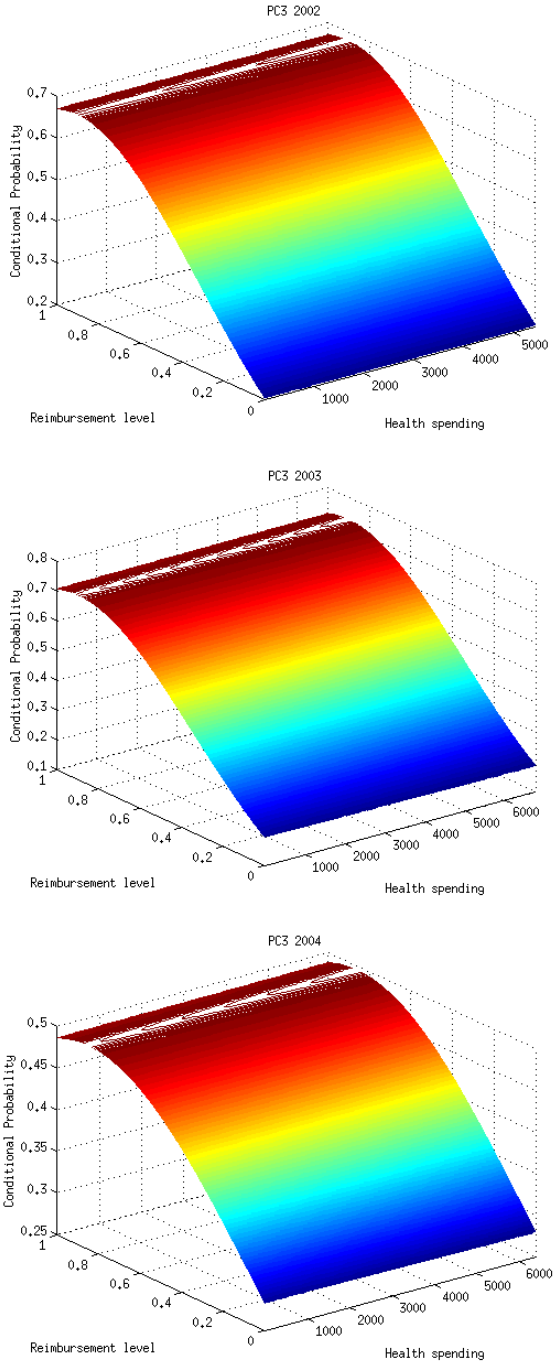


Figure 2.3: Conditional Reimbursement Distributions, PC3 Plan 2002-2004

2.4.2 Estimating the Theta Distributions

Once we estimate the conditional reimbursement probabilities, we use these distributions to estimate theta distributions. The underlying identification assumption is that the distribution of latent health status in the three plans is the same for each of the three years in our dataset. This is a reasonable assumption for the short range of years, 2002-2004. It is also supported by the proxy health status variable constructed in the data from diagnosis codes. This health proxy status variable has a similar mean and distribution across all three years, as reported above. Although we assume that the entire distribution of latent health status, θ, remains constant, individual θ_i may change, as well as plan-level variables such as premiums and cost sharing.

Equation 2.4.1 is the formula for the estimated $\hat{\theta}_i$ for each individual enrollee. Given the estimated conditional reimbursement probabilities and the data on spending and premiums, we can formulate this $\hat{\theta}_i$ in terms of the remaining parameters to be estimated, $\hat{\gamma}_1$ and $\hat{\gamma}_2$.

2.4.3 Solve for Utility Parameters

For each single-coverage full year enrollee i in a given plan j and a given year t, we now have an expression for the enrollee's latent health status in terms of the utility parameters of the enrollee's maximization problem. We

now solve for the values of the utility parameters which satisfy our identifying condition that the distribution of latent health status is the same in each year.

We use a grid search on the possible values of the two utility parameters to find the values that make the latent health status distribution most similar across 2002-2004. Varying size grids for $\hat{\gamma}_2 \in [1, 6]$ and $\hat{\gamma}_2 \in [1, 6]$ are used. For each combination of $\hat{\gamma}_1$ and $\hat{\gamma}_2$, we calculate entire distribution of resulting $\hat{\theta}_{ijt}$. We then use GMM to choose the pair of utility parameters that minimize the difference between the first four moments describing the latent health status distribution of a given year.

Four moments in each of the three years means 12 moment conditions. Let $E[h_n(w, \gamma_0)] = 0$ denote the moment conditions, where $h(w, \gamma)$ is a 12×1 vector function of the moments where $n = \{1, 2, .., 12\}$, $w = (m, y, p, \hat{f})$ denotes the matrix of observables which make up the $\hat{\theta}$ values, $\gamma = (\gamma_1, \gamma_2)$ is the vector of unknown utility parameters, and γ_0 is the value of γ in the data generation process. The weighting matrix is the identity matrix. For example, the first three moment conditions are the differences in mean for each combination of years. Let $\mu_{\hat{\theta},2002}(\hat{\gamma})$ indicate the mean of the latent health distribution $\hat{\theta}$ in the year 2002 for given values of $\hat{\gamma}$. Then the first three moment conditions are:

$$h_1(w, \hat{\gamma}) = \left(\mu_{\hat{\theta},2002}(\hat{\gamma}) - \mu_{\hat{\theta},2003}(\hat{\gamma})\right)^2$$
$$h_2(w, \hat{\gamma}) = \left(\mu_{\hat{\theta},2003}(\hat{\gamma}) - \mu_{\hat{\theta},2004}(\hat{\gamma})\right)^2$$
$$h_3(w, \hat{\gamma}) = \left(\mu_{\hat{\theta},2002}(\hat{\gamma}) - \mu_{\hat{\theta},2004}(\hat{\gamma})\right)^2$$

Similar moment conditions are formed for variance, skewness, and kurtosis of the $\hat{\theta}$ distribution in each year. The final estimator minimizes the sum of squared differences between the yearly distributions of $\hat{\theta}$. The pair of $\hat{\gamma}_1, \hat{\gamma}_2$ over the entire grid of combinations which minimizes the sum of the twelve moment conditions is the value of $\hat{\gamma}$ that solves the generalized method of moments minimization problem.

2.4.4 Asymptotics

The difficulty of computing standard errors and obtaining a valid statistical inference procedure in semiparametric models is well known. The estimator proposed in this paper is a semiparametric GMM estimator that depends on first stage nonparametric estimation of the conditional distribution of the rate of reimbursement given the amount of medical expenditure. Similarly to other semiparametric models with first stage nonparametric estimation of conditional mean or distribution functions, statistical inference

is a major challenge.

There are several conventional approaches to compute the correct standard errors to take into account the statistical uncertainty introduced by the first stage nonparametric estimation. The first one is to derive the asymptotic distribution of the estimator analytically, and replace the asymptotic variance with a consistent estimate based on the sample data. This is in principle possible by following the pathwise derivative calculation in Newey (1994). A second approach is resampling. Either bootstrap or subsampling will provide a valid inference procedure for this particular estimator. A third approach is to make use of an insight in Newey (1994), who shows that the asymptotic variance of the second stage estimators does not depend on how the first stage nonparametric estimation method is implemented. While the first stage estimator is currently implemented using kernel density smoothers, the second stage estimator will have approximately the same asymptotic variance even if the first stage is estimated using a sieve parametric approximation instead of the kernel smoother. If the implementation of the first stage estimator can be modified to be a sieve parametric approximation method, then the (overidentifying) moment conditions that are used in obtaining the estimator can potentially be modified to a set of exactly identifying moment conditions, for both the first stage sieve parameters and the second stage structural parameters of the model. According to Newey (1994), if this is possible, the approximate variance of the second stage es-

timator can be read off from the lower diagonal of the variance-covariance matrix of the entire generalized method of moment estimator that includes both the first stage and second stage estimators. Computing the overall variance-covariance matrix is straightforward using the conventional sandwich formula for GMM estimators.

Unfortunately each of these approaches has its own disadvantages. The pathwise derivation calculation in Newey (1994) is often tedious and prone to errors in the analytic computation. The resulting asymptotic variance estimate can also be complex and is sensitive to coding errors. Resampling methods require recomputing the estimators repeatedly over many bootstrap iterations. Given the nonlinear nature of the method of moment estimators, this might not be computationally feasible. Replacing the first stage kernel smoother with a sieve parametric approach also seems at odds with the current implementation of the semiparametric estimator, and might also lead to different point estimates for the second structural parameters. In addition, implementing a first stage sieve parametric approach appears to be more difficult than implementing the kernel smoother.

The intuition underlying the computation of the standard errors is as follows. In computing the standard errors for a two step estimator the first stage is treated in a parametric fashion (somewhat like in sieve estimation). Then the right standard error for the second stage can be read off from the lower diagonal components of the entire variance-covariance matrix of both

the first and second stage parameters. The paper by Ackerberg, Chen, and Hahn (2009) formally proves the validity of this approach.

In the following we formally outline the pathwise derivative argument of Newey (1994) which justifies the asymptotic normality of the semiparametric two step estimator. First note that the proposed estimator in this paper takes a general form for most two step semiparametric estimators:

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma} g_n(\gamma)' W_n g_n(\gamma),$$

where

$$g_n(\gamma) = \frac{1}{n} \sum_{i=1}^{n} g\left(z_i, \hat{f}_{a,m}(\cdot,\cdot), \gamma\right).$$

In the above, $\hat{f}_{a,m}(\cdot,\cdot)$ denotes the first stage kernel smoothing nonparametric estimate of the joint density function of a and m. The functional notation is used to emphasize that the second stage moment condition depends on the entire joint density because θ is recovered through an equation that depends on integrating against the conditional distribution of a given m and on the derivative of the conditional density $f_{a|m}(\cdot|\cdot)$.

It can be shown through standard Taylor expansion arguments that under suitable regularity conditions (which include the condition that the first stage nonparametric regression should make use of an undersmoothing se-

quence of bandwidth parameters):

$$\sqrt{n}(\hat{\gamma}-\gamma) = G^{-1}W\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g\left(z_i, \hat{f}_{a,m}(\cdot,\cdot), \gamma_0\right) + o_p(1).$$

In the above W is the probability limit of W_n, and

$$G = \frac{\partial}{\partial\gamma} g\left(z_i, f_{a,m}(\cdot,\cdot), \gamma_0\right).$$

When the nonparametric component $f_{a,m}(\cdot,\cdot)$ is known, asymptotic normality of $\hat{\gamma}$ follows immediately from a central limit theorem applied to the normalized sum of the moment conditions evaluated at γ_0:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g\left(z_i, f_{a,m}(\cdot,\cdot), \gamma_0\right) \xrightarrow{d} N(0,\Omega) \quad \text{where} \quad \Omega = Var(g\left(z_i, f_{a,m}(\cdot,\cdot), \gamma_0\right)).$$

However, when $f_{a,m}(\cdot,\cdot)$ has to be estimated in the first stage, its impact on the second stage asymptotic variance needs to be taken into account. Specifically, it can be shown understand additional regularity conditions that

$$\begin{aligned} & \tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n} g\left(z_i, \hat{f}_{a,m}(\cdot,\cdot), \gamma_0\right) - \tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n} g\left(z_i, f_{a,m}(\cdot,\cdot), \gamma_0\right) \\ = \quad & \sqrt{n}E\left[g\left(z_i, \hat{f}_{a,m}(\cdot,\cdot), \gamma_0\right) - g\left(z_i, f_{a,m}(\cdot,\cdot), \gamma_0\right)\right] + o_p(1). \end{aligned}$$

The pathwise derivation calculation in Newey (1994) seeks a linear in-

fluence function $\psi(z_i)$ such that

$$\sqrt{n}E\left[g\left(z_i, \hat{f}_{a,m}(\cdot,\cdot), \gamma_0\right) - g\left(z_i, f_{a,m}(\cdot,\cdot), \gamma_0\right)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(z_i) + o_p(1).$$

Conditional on knowledge of this influence function, it is easy to see that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g\left(z_i, \hat{f}_{a,m}(\cdot,\cdot), \gamma_0\right) \xrightarrow{d} N(0, Var(g\left(z_i, \hat{f}_{a,m}(\cdot,\cdot), \gamma_0\right) + \psi(z_i))).$$

Then we can conclude that

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1})).$$

2.5 Results and Analysis of Asymmetric Information

2.5.1 Test of Identifying Assumptions

We can test our identifying assumptions by using only a subset of the available years for estimation, and then use the estimated parameters from this subset to construct the health shock distributions for all the years in the dataset. If the resulting health shock distributions are similar both for the estimated subset years and for the predicted year, this lends strong support to our identifying assumption that the health shock distribution remains

constant over the three years in our dataset.

To perform this test, we estimated parameters based on only two years from the sample, for example 2002 and 2003, and then used the estimated $\hat{\gamma}$s to construct the estimated health shocks $\hat{\theta_{2002}}$ and $\hat{\theta_{2003}}$. The same $\hat{\gamma}$s that were based on 2002 and 2003 data are applied to the 2004 data to construct a predicted $\hat{\theta_{2004}}$. We then perform a Kolmogorov-Smirnov test for equality of distributions between each estimated $\hat{\theta}$ distribution versus the predicted $\hat{\theta}$ distribution. Table 2.5 lists the results for each of the three tests. The test for predicting the years 2003 and 2004 finds that the estimated health shock distributions cannot be statistically distinguished from the predicted health shock distributions. The KS test for equality of distributions is not rejected. For the test predicting the 2002 health shock distribution, we do see that the KS test rejects the equality of distribution between the predicted and the estimated years. Given that our sample size is very large– each year is over 3,000 observations – the statistical precision of this test is quite high,. However, economic significance may still be valid, despite statistical rejection, given the large sample sizes. The predicted 2002 vs. estimated health shock distributions are displayed in Figure 2.4. Comparing the three histograms we can see that the general shape of the distributions do remain very similar. Thus, our assumption of the health shock distribution remaining the same over three years seems to hold for each of the three prediction scenarios.

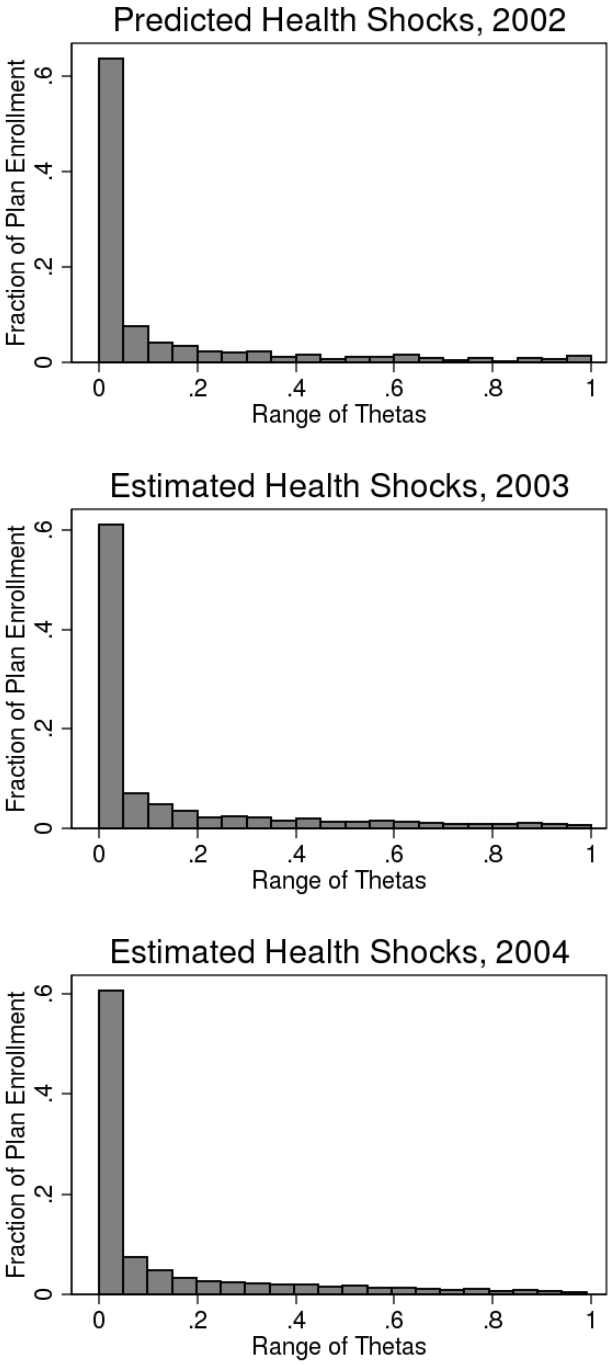


Figure 2.4: Predicted 2002 vs. Estimated 2003, 2004 Distributions

Table 2.5: Predicted vs. Estimated Health Shock Distributions

H_0: X_1 and X_2 are from the same continuous distribution.
H_A: X_1 and X_2 are from different distributions.

(X_1)		(X_2)	
Predicted Year		**Estimated Year 1**	**Estimated Year 2**
2002	KS stat	0.0343	0.0404
	Result	Reject	Reject
2003	KS stat	0.0262	0.0236
	Result	Do Not Reject	Do Not Reject
2004	KS stat	0.0313	0.0275
	Result	Do Not Reject	Do Not Reject

2.5.2 Utility Parameter Estimates

Our approach computes a coefficient of relative risk aversion to health as distinct from that for aggregate consumption. The utility parameters are estimated using a grid search over the potential parameter space. The moment conditions in the GMM were calculated for all combinations of γ_1, γ_2 in the grid and we chose the pair that minimized the GMM problem. The γs were computed over a grid between 1 and 6, over grids of increasing fineness. Table 2.6 displays values of $\hat{\gamma}$ for increasingly fine grids over the parameter space. The standard errors were computed using bootstrapping.

The resulting risk coefficients for γ_1 are in the range [1.88, 1.98] and in the range [3.12, 3.27] for γ_2. Higher values in the γ_1 range tend to be associated with correspondingly higher values of γ_2. This result implies that individuals are more risk averse with respect to health status than the aggregate consumption commodity. These estimates are within the range

Table 2.6: Estimated Risk Coefficients $\hat{\gamma_1}$, $\hat{\gamma_2}$

Grid Size	$\hat{\gamma_1}$	$\hat{\gamma_2}$
20	1.98	3.27
	(0.76)	(1.20)
40	1.88	3.12
	(0.86)	(1.35)
50	1.93	3.23
	(0.86)	(1.35)

Standard errors in parentheses.
Estimates from 200 bootstrap iterations.

found in the literature on consumption (see e.g., Zeldes 1989, Shea 1995, Hansen and Singleton 1982, Gourinchas and Parker 2002). Risk parameters between 1 and 2 have been found to approximate empirical data in work on risk in consumption versus investment (Prescott 1986). Kocherlakota's (1996) survey of risk parameters cites an empirical parameter range under the value of 10. Our estimated health risk coefficients are slightly higher than previous coefficients between consumption and investment, which is consistent with a belief that people are more risk adverse with respect to their health. Health often cannot be regained once lost.

2.5.3 Moral Hazard

The concept of moral hazard we measure addresses patients' counterfactual behavior in the absence of insurance coverage. The key component is a measure of "overconsumption" by the patient in the face of a generous

insurance coverage, when a patient's out-of-pocket expense is not matched to the full cost of providing the care. In the context of employer-sponsored insurance, the patient chooses his plan for a given year, and then chooses his health care consumption throughout the year based on the plan contract. We measure moral hazard as the difference in health expenditures due to the presence of this insurance contract. Patients' health care consumption behavior, or "excess" consumption, has been cited as a factor in rising health care costs. This measure of moral hazard focuses more on patient behavior than some of the previous empirical literature, such as Feldstein (1973) and Feldman and Dowd (1991).

We perform a counterfactual experiment that replicates a social planner's problem where patients must pay all health care expenditures out-of-pocket. The method backs out consumer health expenditures if patients were not subjected to a reimbursement schedule for their health care expenditures, but instead were compensated in their income for possible health care expenses. To calculate this amount of overconsumption, we first calculate the total reimbursement that was given to each patient in each year. That is, the observed amount $T = a_j \cdot m$. Next, we calculate the distribution of the health shocks, θ, in each year using the estimated utility parameters from the most detailed grid search of $[\gamma_1, \gamma_2]$.

The modified budget constraint is then:

$$c_T + m_T \leq y - p_j + T$$

where T represents the social planner's lump sum transfer, which is exactly equal to the amount of reimbursement to the patient that was observed in the data. The variables c_T and m_T are the composite good consumption and health expenditure in the counterfactual environment.

The patient's utility maximization problem is then:

$$U(m_T, p_j, y, \theta; \gamma) = (1-\theta)\frac{(y - p_j - m_T + T)^{1-\gamma_1}}{1-\gamma_1} + \theta\frac{m_T^{1-\gamma_2}}{1-\gamma_2} \qquad (2.5.1)$$

The patient's maximization problem sets his marginal rate of substitution between outside good consumption and health care expenditure equal to the ratio of prices. The patient no longer faces a reimbursement schedule a_j, but receives a lump sum transfer of T. The ratio of the price of health care to the price of the composite good is simply equal to one, because all health care is now purchased at full out-of-pocket price. In our estimation, we calculate the value of the MRS:

$$MRS = \frac{\theta}{(1-\theta)}\frac{(y - p_j + T - m_T)^{\gamma_1}}{m_T^{\gamma_2}}$$

We then solve for the value of m_T where $MRS = 1$. The difference

between m_T and the actual observed health expenditure is a measure of the amount that a patient is overconsuming as a result of the change in the price ratio. This is because m_T is the amount a patient chooses if forced to pay the full cost of care, given the same realized net income as we observed when insured. In the counterfactual scenario patients are no longer restricted to spend the amount of their reimbursement income, T, on health expenditure. If the patient chooses to spend part of the unrestricted income T on outside consumption, this measures how purchasing health care through insurance influences her behavior. In the estimation results, we will refer to m_T as first-best consumption. One advantage of this measure of moral hazard is that it allows changes in patient behavior for a wider range of reimbursement schedules, in particular nonlinear schedules. Traditional measures are typically based on a linear reimbursement schedule, e.g., Pauly (1968).

Our estimation finds that the magnitude of overconsumption in our health plans is substantial. For each year, we calculate the differences for each patient between the amount of health expenditures which were observed in the data and the first best level of expenditures, as described above. Figure 2.5 shows the resulting distributions of overconsumption estimates for each of the years 2002-2005. The horizontal axis is the difference between observed expenditure and first best expenditure. The vertical axis is a count of the number of patients. Many patients in all three years have very low

differences between observed and first best health expenditures. However, the difference for some patients is as high as $5,000. All three years have approximately the same range of overconsumption, between $0 and $5,000. The number of patients with differences near zero is approximately 1,800 patients with overconsumption in the range between $0 and $500. After this point, the distribution drops rapidly until reaching approximately $3,000 in overconsumption. After $3,000, the number of patients at each level drops off to less than a hundred in each bin.

Summing overconsumption over all patients, the total yearly estimate of overconsumption across plans is large. The sum of all patients' differences between observed and first best is $2,125,480 in 2002. The sum in 2003 is $2,504,380 and in 2004 is $2,270,059. To put individual overconsumption in perspective, Table 2.7 list summary statistics by year for overconsumption as a percentage of original health care expenditure. On average, patients' overconsumption was 45 percent of the patient's original health care expenditure in all three years. The median overconsumption percentage is also over 40 percent of the original choice of health expenditure, and the standard deviation is 10 percentage points or less in all years. Finally, another notable feature is that patients with overconsumption that is the largest percentage of original health care spending are the patients in the highest ranges of original health care spending, implying that overconsumption becomes a larger concern as health care expenditure increases.

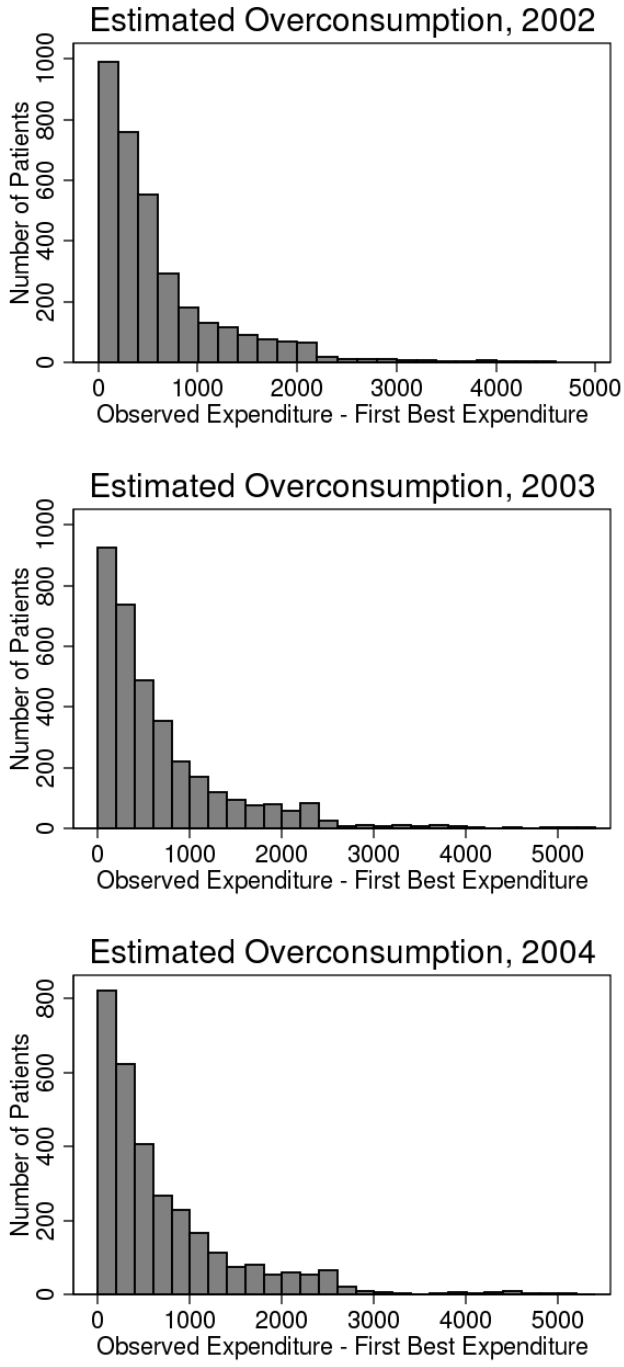


Figure 2.5: Estimated Overconsumption, 2002-2004

Table 2.7: Overconsumption as Percentage of Original Health Care Expenditure

Year	Mean	Median	Std. Dev.	N
2002	45.14	42.09	8.83	3,433
2003	46.22	42.14	8.85	3,533
2004	46.22	42.49	10.04	3,100

2.5.4 Adverse Selection

Adverse selection is a substantial concern in health insurance literature. As such, in this section, we propose a distribution-free test for adverse selection. The presence of adverse selection causes patients to sort across different plans based on their latent health status, e.g. (Rothschild and Stiglitz 1976). In our framework, this would imply that the distribution of the latent health status variable varies across health plans. To determine if the distribution of health status is different across the three plans, we first calculate the overall distribution of latent health status using the estimated utility parameters and observed health care and consumption choices. Figure 2.6 displays the distribution of health status, θ, for each of the years 2002-2004. The horizontal axis is between 0 and 1 in health status, and the vertical axis is the number of patients at each point in the interval. The distribution appears to be bimodal – many patients in each year had very low values of θ, but there is a small clustering at the θ value of 0.8. The large clustering at a θ value of zero is typical of U.S. health status distributions

more generally, which tend to have many relatively healthy individuals, but with a long right tail.

To place our health status measure in the context of observable characteristics, we examine the relationship between the estimated health status and various patient characteristics in our data. Table 2.8 reports the results of two regressions on log health status. Regression 1 includes age, an indicator variable equal to 1 for female, and log salary. Regression 2 also includes the health status proxy variable. The signs of the resulting coefficients generally match our expectations for individual characteristics's effects on health. Age has a positive relationship with log health status in Regression 1. A ten year increase in age is associated with a 0.17 percent increase in the magnitude of our estimated health status. Individuals with higher salaries are associated with lower levels of θ, or better health status, in both regressions. A negative relationship between income and illness is well-documented in both the economics and medical literatures, such as Smith (1999), Ettner (1996), Adler et al. (1994), and Deaton and Paxton (1998). Age is no longer significant when including the health status proxy variable. A one percent increase in the health status proxy variable is associated with a 1.232 percent increase in the value of our estimated health status. While the relationship between our estimated health status and the health status proxy is positive, the regression results show that our health status estimates are able to capture a more general measure of health status

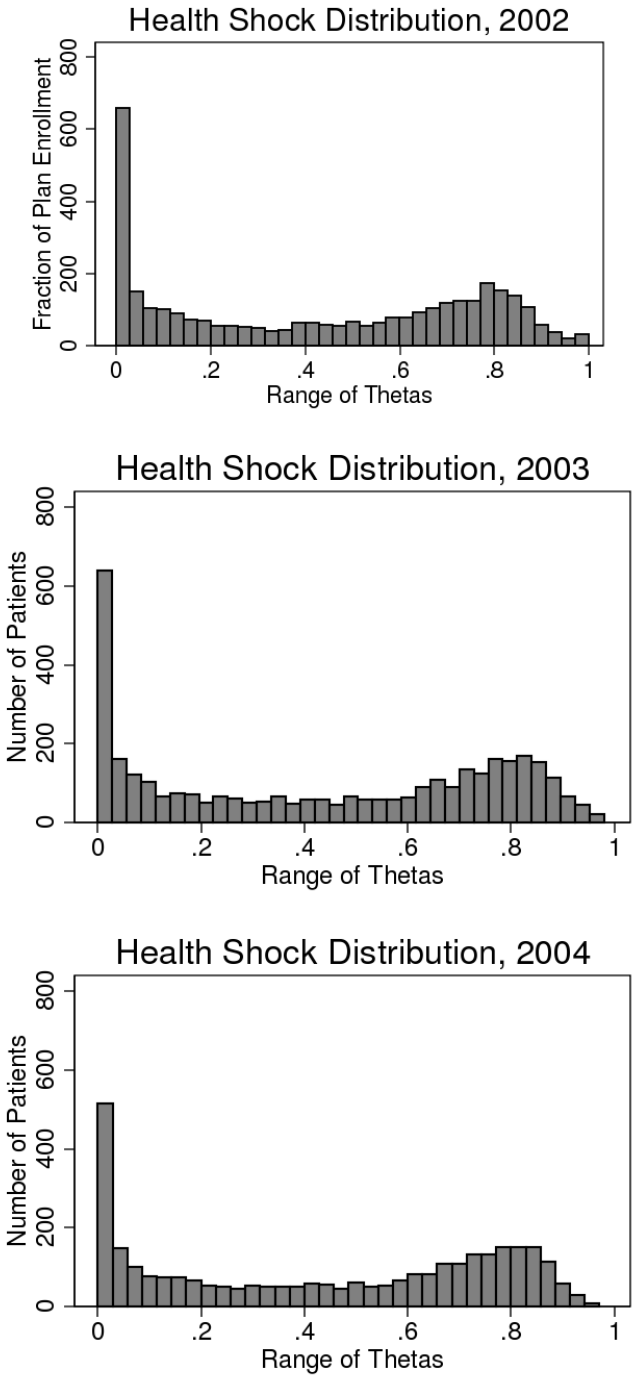


Figure 2.6: Health status Distribution, 2002-2004

than limited diagnosis codes.

Table 2.8: Estimated Health status Regression

Dependent variable: ln(health status, θ)		
Covariate	**Regression 1**	**Regression 2**
Age	0.017	0.002
	(0.087)	(0.070)
Indicator =1 if female	1.160	0.848
	(0.049)	(0.039)
ln(salary)	-0.822	-0.725
	(0.059)	(0.048)
ln(health status proxy)		1.232
		(0.016)
Constant	5.407	5.836
	(0.605)	(0.484)
R-squared	0.079	0.411
N	10,066	10,066

Standard errors in parentheses.

Using our estimate of the distribution of individual health status, θ, for each year, we break down the yearly distribution of health status by plan to examine whether individuals with greater health status appear to self-select into more generous plans. Figure 2.7 shows the distribution of health status types over all years in each of the three plans. The horizontal axis is the range of the θ from 0 to 1, and the vertical axis is the fraction of patients in each plan corresponding to the θ value. A simple visual analysis of the HMO plan, HP, versus the PPO plans, PC2 and PC3, shows that the HMO plan has a much larger fraction of very healthy individuals – those with θ near zero. Over 20 percent of the HP patients had a θ value less than 0.05,

compared with the PPO plans with approximately 15 percent of patients in the same range. Both PC2 and PC3 both show a larger clustering around the health status value of 0.8 than the HMO plan.

To formally test for adverse selection among the plans, we use Kolmogorov-Smirnov test statistics to compare the distributions between the HMO and PPO plans. Table 2.9 reports the K-S statistics from testing the null hypothesis that both plan's health status distributions come from the same continuous distribution. The three possible combinations of plan comparisons (HP and PC2, HP and PC3, PC2 and PC3) are presented. The null hypothesis that both plans were drawn from the same continuous distribution of θs is rejected for every year and plan combination except for PC2 and PC3 in 2002 and 2003. These results are significant at the 1 percent level. The insignificance of the result between PC2 and PC3 may be expected because the pricing structure between the two plans is the same, with the only difference being the network providers. These results confirm the presence of adverse selection within our sample, showing that the patient population is indeed different across the three plans.

Given the results in Table 2.9 confirming the presence of adverse selection, the next step is to test how the patient population sorts between plans. Table 2.10 looks in greater detail at the health status distribution between plans. The tradeoff between HMO plans and PPO plans is that PPO plans provide greater flexibility in provider choice in exchange for more cost shar-

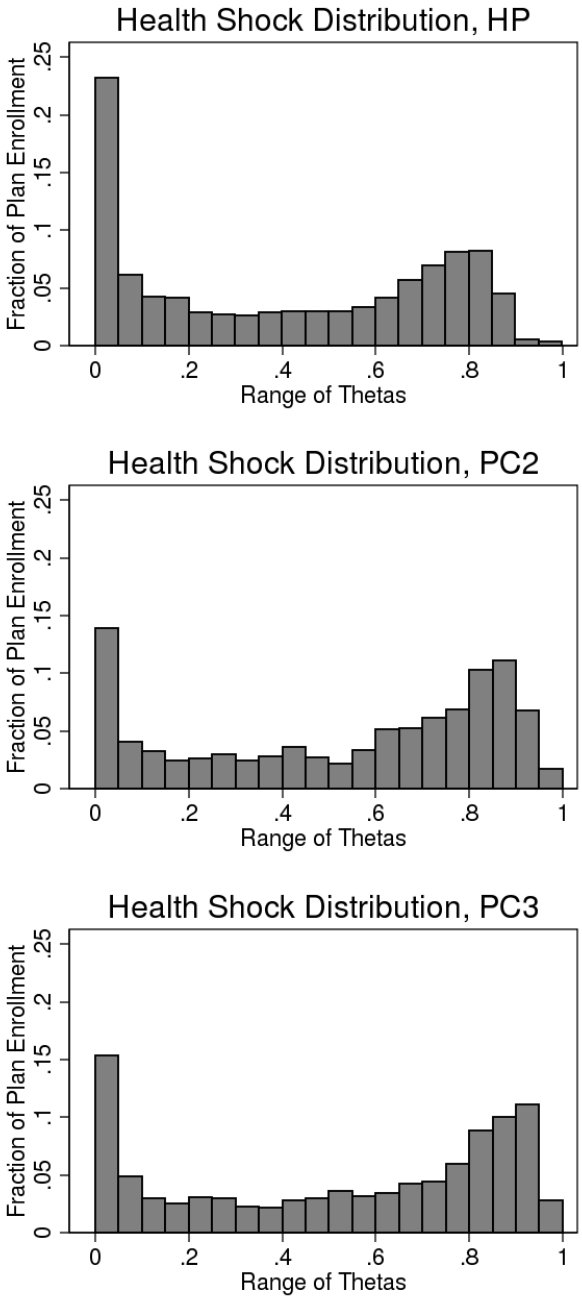


Figure 2.7: Health Status Distribution, by plan

Table 2.9: K-S Test Statistics, Inequality

H_0: X_1 and X_2 are from the same continuous distribution.
H_A: X_1 and X_2 are from different distributions.

Plan (X_1 and X_2)		**2002**	**2003**	**2004**
HP and PC2	KS stat	0.1948	0.1778	0.2347
	Result	Reject	Reject	Reject
HP and PC3	KS stat	0.1638	0.3036	0.1864
	Result	Reject	Reject	Reject
PC2 and PC3	KS stat	0.0459	0.1819	0.1050
	Result	Do Not Reject	Reject	Do Not Reject

ing by patients. The HMO restricts provider choice the most tightly, but has the lowest premium and cost sharing structure of all the plan choices. Patients that have the greatest preference for flexibility in provider choice are likely to be patients with specific health concerns. Additionally, health patients who do not anticipate using much care should choose the lowest cost plan. Relatively healthy patients who do not expect to incur more than a yearly checkup may also be satisfied with the more limited network of providers in the HMO plan. Thus, if a patient expects a low value of θ, that is relatively good health status, she will choose the cheapest option – the HMO plan. If a patient expects more a more severe θ, she may select into the PC3 plan for the most flexibility. A more severe health status enters utility as a larger θ. Therefore, if a plan enrolls a population with a more severe health status distribution, the cumulative distribution function of the plan with lower values of health status should stochastically dominate the

cdf of a plan with a larger health status values.

The first two tests of Table 2.10 show that the results of the K-S test for the comparison of the cdfs between the HMO and PPO plans support this hypothesis. When comparing the HMO plan, HP, with either PPO plan, PC2 or PC3, the null hypothesis of equality of distributions is rejected in favor HP's cdf being larger than both the cdf for PC2 and PC3. This implies that the portion of relatively healthy patients is larger in the HP plan than in the PC2 or PC3 plan.

Within the tiered PPO plans, PC2 and PC3 differ only in the types of facilities that patients may visit. PC3 providers include the PC2 providers plus additional providers. Once again, patients with more complicated health conditions, a larger value of θ, may select into the PC3 plan to take advantage of the additional available providers. In comparing the two PPO plans in 2003, the hypothesis of equality of distributions is rejected in favor of PC2 having a larger cdf, meaning the distribution of patients is healthier in the PC2 plan. However, the same hypothesis cannot be rejected for the other two years, 2002 and 2004. This inability to reject may be due to the fact that, despite the difference in provider choice, these two plans do have the same cost structure.

Finally, to check the validity of the distribution tests for adverse selection, we check the same hypotheses of equality in distributions, but instead of across plans we test for equality of distributions across years for the same

Table 2.10: K-S Test Statistics, Larger

H_0: X_1 and X_2 are from the same continuous distribution.
H_A: X_1's cdf is greater than X_2's cdf.

Plan (X_1 and X_2)		**2002**	**2003**	**2004**
HP and PC2	KS stat	0.1948	0.1778	0.2347
	Result	Reject	Reject	Reject
HP and PC3	KS stat	0.1638	0.3036	0.1864
	Result	Reject	Reject	Reject
PC2 and PC3	KS stat	0.0097	0.1819	0.0204
	Result	Do Not Reject	Reject	Do Not Reject

plan. The results are displayed in Table 2.11. In general, the null hypothesis cannot be rejected that a plan's health status distribution is different across years. The two exceptions are the HP plan between 2003-2004 and the PC3 plan between 2002-2003. These results support the assertion that the differences across plans in Table 2.10 are due to fundamental distribution differences that persist across years. Overall, the K-S tests show that the underlying distributions of health status are different between the different types of plans offered to enrollees, and that these differences persist over several years of testing. Together, these differences are strong evidence of adverse selection.

2.6 Conclusions

In this paper, we present a new approach to measuring moral hazard and adverse selection in health insurance and choice of health spending. This

Table 2.11: K-S Test Statistics, Within plans

H_0: X_1 and X_2 are from the same continuous distribution.
H_A: X_1 and X_2 are from different distributions.

Years (X_1 and X_2)		**HP**	**PC2**	**PC3**
2002 and 2003	KS stat	0.0212	0.0645	0.1795
	Result	Do Not Reject	Do Not Reject	Reject
2003 and 2004	KS stat	0.0281	0.1098	0.1495
	Result	Do Not Reject	Do not reject	Reject
2002 and 2004	KS stat	0.0332	0.0984	0.1035
	Result	Do Not Reject	Do Not Reject	Do Not Reject

approach develops a model of demand for health insurance and medical utilization in the presence of unobserved heterogeneity in the latent health status of individuals. Our framework and resulting estimates offer several contributions to the existing work on adverse selection and moral hazard. These include separate estimation of moral hazard and adverse selection, the ability to flexibly incorporate insurance reimbursement schedules, and our empirical estimates of the two phenomena in detailed claims level data. Previous literature has been limited by difficulties separating the two effects of moral hazard and adverse selection. Through our estimation of the distribution of underlying health status of the population, we are able to perform a counterfactual which allows us to isolate the effect of insurance on health spending, and thus gives us a measure of moral hazard. Adverse selection is measured separately by testing for equality in distributions of this health status variable across plans. The second contribution comes

from our semiparametric approach which measures insurance reimbursement schedules nonparametrically. This measurement approach has two advantages. First, nonparametric estimation allows for complex insurance plans including copays, deductibles, and other nonlinear features. Secondly, our nonparametric estimation of a conditional probability of reimbursement also incorporates the reality that plans are complex and difficult for consumers to predict in every state of the world. This conditional probability approach is a great advantage over previous literature which assumes deterministic certainty in patients' understanding of their insurance policies. Finally, this paper presents a measure of the magnitude of both moral hazard and adverse selection in a self-insured employer insurance pool. Our estimates indicate that moral hazard may account for as much as 40 percent of current health spending in these plans, and that patients with a relatively less complex health status do sort into the lowest cost-sharing plan.

Although our proposed semiparametric method provides a more flexible and robust alternative for analyzing the empirical issues of adverse selection and moral hazard in health insurance, several limitations are acknowledged. The utility function specification we use is assumed to be separable in the consumption of the composite good and medical care. While this specification captures the risk aversion features of consumer utilities in health status, it rules out more flexible interactions between the utility derived from composite good consumption and health status. It is sometimes argued that

the marginal utility for composite good consumption might decrease in the case of severe illness (e.g. Viscusi and Evans 1990). We note, however, that Spence and Zeckhauser (1971) and Blomqvist (1997) use a similar specification, and Campo, Guerre, Perrigne and Vuong (2003) also require similar restrictions on utility in an auctions context.

In conclusion, in spite of these limitations our research is novel in that it develops a tractable estimation procedure under minimal parametric assumptions to simultaneously examine adverse selection and moral hazard in health insurance contracts. Our research is also important as it provides a framework for similar analysis in other contexts, especially with cross section data, where distortions exist due to asymmetric information.

Bibliography

Bajari, P., H. Hong, M. Park, and R. Town (2010): "Regression Discontinuity Designs with an Endogenous Forcing Variable and an Application to Contracting in Health Care," Working paper.

Cameron, C., and P. Trivedi (2005): *Microeconomics: Methods and Applications*. Cambridge University Press, New York, NY.

Centers for Medicare and Medicaid Services (2009): "Medicare & You 2010," Annual report.

Cherkin, D., L. Grothaus, and E. Wagner (1989): "The effect of office visit copyaments on preventative care services in an HMO," *Medical Care*, 27(7), 669–679.

Chernew, M. E., R. A. Hirth, and D. M. Cutler (2009): "Increased Spending on Health Care: Long Term Implications for the Nation," *Health Affairs*, 28(5).

Committee on Ways and Means (2006): "The Prescription Drug 'Doughnut Hole', 7 Million at Risk, Preimiums Over 250% Higher for Full Coverage," Briefing, U.S. House of Representatives.

Duan, H., W. Manning, C. Morris, and J. Newhouse (1983): "A comparison of Alternative Models for the Demand for Medical Care," *Journal of Business and Economic Statistics*, 1(2), 115–126.

Eichner, M. (1998): "The Demand for Medical Care: What People Pay Does Matter," *American Economic Review Papers and Proceedings of the Hundred and Tenth Annual Meeting of the American Economic Association*, 88(2), 117–121.

Feldstein, M. S. (1971): "Hosptial Cost Inflation: A study of Nonprofit Price Dynamics," *American Economic Review*, 61(5), 853–872.

Fuchs, V., and M. Kramer (1972): "Determinants of Expenditures for Physicians' Services in the United States, 1948-1968," (117).

Hosek, S. D., S. Mahnaovski, J. S. Ringel, and B. A. Vollaard (2005): "Elasticity of Demand for Health Care: A Review of the Literature and Its Application to the Military Health System," Discussion paper.

Huang, L., and R. N. Rosett (1973): "The Effect of Health Insurance

on the Demand for Medical Care," *Journal of Political Economy*, 81(2), Part 1: 281–305.

Janssen, R. (1992): "Time Prices and the Demand for GP Services," *Social Sciences and Medicine*, 34(7), 725–733.

Kaiser Family Foundation (2006): "Employer Health Benefits: 2006 Summary of Findings," Annual Report 7528.

——— (July 2009): "Health Care and the Middle Class: More costs and less coverage," in *Focus on Health Reform*, ed. by D. Rowland, C. Hoffman, and M. McGinn-Shapiro, no. 7951.

Keeler, E., and J. E. Rolph (1988): "The Demand for Episodes of Treatment in the Health Insurance Experiment," *Journal of Health Economics*, 7, 337–367.

Kowalski, A. (2009): "Censored Quantile Instumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care," Working Paper 15085, National Bureau of Economic Research.

Manning, W. G., J. P. Newhouse, H. Duan, E. B. Keeler, and A. Leibowitz (1987): "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *American Economic Review*, 77(3), 251–277.

New York Times (October 17, 2009): "Making Sense of High Deductible Plans," B6.

Newhouse, J. P. (1993): *Free for All: Lesson from the RAND health insurance experiment*. Havard University Press, Cambridge, MA.

Newhouse, J. P., C. E. Phelps, and M. S. Marquis (1980): "On Having Your Cake and Eating It Too: Econometric Problems in Estimating the Demand for Health Services," *Journal of Econometrics*, 13, 365–390.

Phelps, C. E. (1992): *Health Economics*. HarperCollins, New York.

Scitovsky, A. A., and N. McCall (1977): "Coinsurance and the demand for physician services: Four years later," *Social Security Bulletin*, 40(5), 1–41.

Scitovsky, A. A., and N. M. Snyder (1972): "Effect of Coinsurance on Use of Physician Services," *Social Security Bulletin*, 36(6), 3–19.

Time Magazine (August 31, 2009): "A Health Care Glossary," pp. 26-27.

US News and World Reports (February 3, 2009): "Medicare Drug Plan 'Doughnut Hole' Could Impact Seniors' Health," Health Day.

van der Klaauw, W. (2008): *'Regression-Discontinuity Analysis', The New Palgrave Dictionary of Economics*. Palgrave Macmillon, second edn.

Wall Street Journal (June 12, 2007): "Health Savings Plans Start to Falter," D1.

Chapter 2

[1] Abbring, J. H., P. A. Chiappori, J. H. Heckman and J. Pinquet. Adverse Selection and Moral Hazard in Insurance: Can Dynamic Data Help to Distinguish? Journal of the European Economic Association. 2003; 1(Papers and Proceedings): 512-521.

[2] Ackerberg, D.; X. Chen, and J. Hahn. A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators. Working Paper. UCLA. 2009.

[3] Adler, Nancy E.; T. Boye, S. Cohen, S. Folkman, and R.L. Kahn. Socioeconomic Status and Health: The challenge of the gradient. American Psychologist. 1994; 49(1):15-24.

[4] Akerlof, George. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. The Quarterly Journal of Economics. 1970; 84(3): 488-500.

[5] Arrow, K.J. Uncertainty and the Welfare Economics of Medical Care. American Economic Review. 1963; 53:941-973.

[6] Athey, S. and P. Haile. Identification of Standard Auction Models. Econometrica. 2002; 70:2107-2140.

[7] Bajari, Patrick and Matt Kahn. Estimating Housing Demand with an Application to Explaining Racial Segregation in Cities. Journal of Business and Economic Statistics. 2005; 23(1): 20-33.

[8] Blau, D. and Donna B. Gilleskie. The Role of Retiree Health Insurance in the Employment Behavior of Older Males. Working Paper. UNC. 2003.

[9] Blomqvist, A.G. Optimal Nonlinear Health Insurance. Journal of Health Economics. 1997; 16(3):303-321.

[10] Bowman, A.W. and Azzalini, A. (1997). Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations. Oxford University Press, Oxford.

[11] Cameron. A. C.; P. K. Trivedi, Frank Milne, and J. Piggott. A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia. Review of Economic Studies. 1988; 55(1):85-106.

[12] Campo, Sandra M.; I. Perrigne, and Q. Vuong. Asymmetry and Joint Bidding in OCS Wildcat Auctions. Journal of Applied Econometrics. 2003; 18:179-207.

[13] Campo, Sandra M.; E. Guerre, I. Perrigne, and Q. Vuong. Semiparametric Estimation of First-Price Auctions with Risk Averse Bidders. Working Paper. UNC. 2003.

[14] Cardon, James H. and Igal Hendel. Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey. RAND Journal of Economics. 2001; 32(3):408-27.

[15] Cawley, J., and T. Philipson. An Empirical Examination of Information Barriers to Trade in Insurance. American Economic Review, 1999; 89: 827-46.

[16] Chiappori, Pierre-Andre and B. Salani. Testing for Asymmetric Information on Insurance Markets. Journal of Political Economy. 2000; 108(1):56-78.

[17] Chiappori, Pierre-Andre and Bernard Salani. Testing Contract Theory: a Survey of Some Recent Work, in Advances in Economics and Econometrics - Theory and Applications, Eighth World Congress, M. Dewatripont, L. Hansen and P. Turnovsky, ed., Econometric Society Monographs, Cambridge University Press, Cambridge, 2003: 115-149.

[18] Chiappori, Pierre-Andre, Bruno Jullien, and Franois Salani. Asymmetric Information in Insurance: General Testable Implications. RAND Journal of Economics. Forthcoming.

[19] Cohen, Alma, and Liran Einav. Estimating Risk Preferences from Deductibel Choice. NBER Working Paper No. 11461. 2005.

[20] Courty, P., and Hao Li. Sequential Screening. Review of Economic Studies. 2000; 67:697-717.

[21] Currie, Janet and Brigitte. C. Madrian. Health, Health Insurance and the Labor Market. Ashenfelter, Orley and David Card (eds) Handbook of Labor Economics Vol. 3C. Amsterdam ; New York : Elsevier; 1999: 3309-3416.

[22] Cutler, David M. and Richard J. Zeckhauser. The Anatomy of Health Insurance. Culyer, Anthony J. and Joseph P. Newhouse (eds) Handbook of Health Economics Vol. 1A. Amsterdam ; New York : Elsevier; 2000: 563-643.

[23] Dai, Chifeng, Tracy Lewis, and Giuseppe Lopomo. Delegating Management to Experts. RAND Journal of Economics. Forthcoming.

[24] De Meza, David, and David Webb. Advantageous Selection in Insurance Markets. RAND Journal of Economics. 2001; 32(2):249-262.

[25] Deaton, Angus and Christina Paxton. Aging and Inequality in Health and Income. American Economic Review. 1998: 88(2):248-253.

[26] Ettner, Susan L. New Evidence on the Relationship Between Income and Health. Journal of Health Economics. 1996; 15:67-85.

[27] Einav, Liran, Amy Finkelstein, and Jonathan Levin. Beyond Testing: Empirical Models of Insurance Markets. Annual Review of Economics. 2010; 2:311-36.

[28] Fan, J. and Gijbels, I. Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability. London: Chapman & Hall. 1996.

[29] Feldman, Roger and Bryan Dowd. A New Estimate of the Welfare Loss of Excess Health Insurance. The American Economic Review. 1991; 81(1): 297-301.

[30] Feldstein, Martin. The Welfare Loss of Excess Health Insurance. Journal of Polical Economy. 1973; 81(2)251-280.

[31] Ferrall, Christopher and Bruce Shearer. Incentives and Transactions Costs Within the Firm: Estimating an Agency Model Using Payroll Records. Review of Economic Studies. 1999; 66:309-338.

[32] Finkelstein, Amy and James Poterba. Adverse Selection in Insurance Market: Policyholder Evidence from the U.K. Annuity Market. Journal of Political Economy. 2004; 112(1):183-208.

[33] Finkelstein, Amy and James Poterba. Testing for Adverse Selection with "Unused Observables." NBER Working Paper No. 12112. 2006.

[34] Finkelstein, Amy and Kathleen McGarry. Multiple Dimensions for Private Information: Evidence from the Long-Term Care Insurance Market. American Economic Review. Forthcoming.

[35] Gertler, Paul and Jonathan Gruber. Insuring Consumption Against Illness. American Economic Review. 2002; 92(1): 51-70.

[36] Gilleskie, Donna B. A Dynamic Stochastic Model of Medical Care Use and Work Absence. Econometrica. 1998; 66(1):1-45.

[37] Gourinchas, P. O. and J. Parker. Consumption Over the Life Cycle. Econometrica. 2002. 70(1): 47-89.

[38] Gruber, Jonathan. Health Insurance and the Labor Market. Culyer, Anthony J. and Joseph P. Newhouse (eds) Handbook of Health Economics Vol. 1A. Amsterdam ; New York : Elsevier; 2000; pp. 645-706.

[39] Guerre, E.; I. Perrigne, and Q. Vuong. Optimal Nonparametric Estimation of First-Price Auctions. Econometrica. 2000; 68(3):525-574.

[40] Hansen, L. P. and K. J. Singleton. Generalized Instrumental Variables Estimation of Non Linear Rational Expectations Models. Econometrica. 1982; 50(5):1269-1285

[41] Harris, K. and M. Keane. A Model of Health Plan Choice: Inferring Preferences and Perceptions From a Combination of Revealed Preferences and Attitudinal Data. Journal of Econometrics. 1999; 89: 131-157.

[42] Heckman, James. J. and Thomas E. MaCurdy. A Life Cycle Model of Female Labor Supply. Review of Economic Studies. 1980; 47(2): 47-74.

[43] Hong, Han and Elie Tamer. Inference in Censored Models with Endogenous Regressors. Econometrica. 2003; 71(3):905-932.

[44] Hubbard, R. Glenn, Jonathan Skinner, and Stephen P. Zeldes. Precautionary Savings and Social Insurance. Journal of Political Economy. 1995; 103(2): 360 99.

[45] Juster, F.T. and R. Suzman. The Health and Retirement Study: An Overview. Journal of Human Resources. 1995; Supplement (JHR 30-S): S7-S56.

[46] Karlan, D. and J. Zinman. Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment. Working Paper. Yale University 2005

[47] Keeler, E., J. Newhouse and C. Phelps. Deductibles and the Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule under Uncertainty. Econometrica 1977; 45: 641-655.

[48] Keeler, E., and J. Rolph. The Demand for Episodes of Treatment in the Health Insurance Experiment. Journal of Health Economics 1988; 7:337-367.

[49] Khwaja, A. Health Insurance Habits and Health Outcomes: A Dynamic Stochastic Model of Investment in Health. Ph.D Dissertation, University of Minnesota. 2001.

[50] Khwaja, A. Estimating Willingness to Pay for Medicare Using a Dynamic Life-Cycle Model of Demand for Health Insurance, forthcoming, Journal of Econometrics.

[51] Kocherlakota, Narayana. "The Equity Premium: It's Still a Puzzle", Journal of Economic Literature, March 1996, Vol 34.

[52] Lleras-Muney, A. The Relationship Between Education and Adult Mortality in the U.S. Review of Economic Studies. 2005;.72(1): 189-221.

[53] Manning, W.G.; J.P. Newhouse, N.Duan et al. Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. American Economic Review. 1987; 77(3):251-277.

[54] Newey, W. K. The Asymptotic Variance of Semiparametric Estimators. Econometrica. 1994; 62(6):1349-1382.

[55] Newhouse, J.P. Free For All? Lessons from the Health Insurance Experiment. Cambridge: Harvard University Press. 1993.

[56] Paarsch, Harry J. and Bruce Shearer. Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records. International Economic Review. 2000; 41(1):59-92.

[57] Pauly, M. The Economics of Moral Hazard: Comment. American Economic Review. 1968; 58:531-536.

[58] Pauly, M. Overinsurance and Public Provision of Insurance: the Roles of Moral Hazard and Adverse Selection. Quarterly Journal of Economics. 1974; 88(1):44-54.

[59] Pauly, M. Insurance Reimbursement. Culyer, Anthony J. and Joseph P. Newhouse (eds) Handbook of Health Economics Vol. 1A. Amsterdam ; New York : Elsevier; 2000; pp. 537-560.

[60] Powell, James. Least Absolute Deviations Estimation for the Censored Regression Model. Journal of Econometrics. 1984; 25:303-325.

[61] Prescott, E., Theory ahead of business cycle measurement, Federal Reserve Bank of Minneapolis Quarterly Review, 10, Fall, 9-22. 1986

[62] Rothschild, M. and J.E. Stiglitz. Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. Quarterly Journal of Economics. 1976; 90(4):630-649.

[63] Sieg, Holger. Estimating a Dynamic Model of Household Choices in the Presence of Income Taxation. International Economic Review. 2000; 41 (3), 637-668.

[64] Shea, J. Union Contracts and the Life-Cycle/Permanent-Income Hypothesis. American Economic Review. 1995; 85(1): 186-200.

[65] Smith, James P. Healthy Bodies and Thick Wallets: The dual relation between health and economic status. Journal of Economic Perspectives. 1999; 13(2):145-166.

[66] Spence, M. Job Market Signaling. The Quarterly Journal of Economics. 1973; 87(3): 355-374.

[67] Spence, M. and R. Zeckhauser. Insurance, Information, and Individual Action. American Economic Review. 1971; 61(2):380-387.

[68] Viscusi, K. and W. Evans. Utility Functions that Depend on Health Status: Estimates and Economic Implications. American Economic Review. 1990; 80: 353-374.

[69] Wilson, C. A Model of Insurance Markets with Incomplete Information. Journal of Economic Theory. 1977; 16:167-207.

[70] Wilson, C. The Nature of Equilibrium in Markets With Adverse Selection. Bell Journal of Economics. 1980; 11(1):108-130.

[71] Vera-Hernandez, Marcos. Estimation of a Principal-Agent Model: Moral Hazard in Medical Insurance. The Rand Journal of Economics. 2003. 34(4):670-693.

[72] Zeldes, S. P. Consumption and Liquidity Constraints: An Empirical Investigation. Journal of Political Economy. 1989; 97(2):305-346.

[73] Zeckhauser, R. Medical Insurance: a Case Study of the Tradeoff Between Risk Spreading and Appropriate Incentives. Journal of Economic Theory. 1970; 2(1):10-26.

[74] Zweifel, Peter and Willard G. Manning. Moral Hazard and Consumer Incentives in Health Care. Culyer, Anthony J. and Joseph P. Newhouse (eds) Handbook of Health Economics Vol. 1A. Amsterdam ; New York : Elsevier; 2000; pp. 409-459.

Chapter 3

Appendix

3.1 Appendix A: Appendix to Chapter 1

In this appendix, I lay out a flexible parametric form for the utility function of the agent and solve explicitly for the patient's decision rule.

Consider the following general utility function, which satisfies Utility Conditions 1.1-1.5 from Section 1:

$$\begin{aligned} U(h,\theta,c) &= u(h,\theta)+c \\ &= \gamma\theta+\alpha\theta h^{\beta}-\tau h+c \end{aligned} \tag{3.1.1}$$

Where γ is a parameter on the health shock, α is the parameter on the interaction of the health shock and health expenditures, β is the patient's

risk parameter, and τ is the parameter on the time inconvenience from going to the doctor or from utility-reducing levels of health care.

Conditions 1.1, 1.3, and 1.5 are clearly satisfied. Condition 1.2 is satisfied with $h_\theta^{max} = \left[\frac{\alpha\beta\theta}{\tau}\right]^{\frac{1}{1-\beta}}$. Condition 1.4 is satisfied with γ small enough, $\gamma < \alpha h^\beta$.

The corresponding optimal decision rule is:

$$h^* = \left[\frac{\alpha\beta\theta}{1 - r' + \tau}\right]^{\frac{1}{1-\beta}} \tag{3.1.2}$$

Recall the reimbursement schedule with a deductible and resulting MC schedule as described in Section 1:

$$MC = 1 - r' = \begin{cases} 1 & \text{if } h \leq \bar{h} \\ 0 & \text{if } h > \bar{h} \end{cases}$$

Given the reimbursement schedule and a draw of θ, the corresponding optimal h^* in each marginal reimbursement r' section is:

$$h^* = \begin{cases} \left[\frac{\alpha\beta\theta}{\tau+1}\right]^{\frac{1}{1-\beta}} & \text{if } h^* \leq \bar{h} \\ \left[\frac{\alpha\beta\theta}{\tau}\right]^{\frac{1}{1-\beta}} & \text{if } h^* > \bar{h} \end{cases} \tag{3.1.3}$$

Notice that the optimal h^* that is chosen in the first part of the reimbursement schedule, where all health expenditure must be paid fully out-of-pocket, is smaller than the h^* that is chosen in the second part of the

reimbursement schedule, where all additional units of expenditure are fully covered by the insurance plan.

Because the decision rule for the choice of h is strictly increasing in θ as θ approaches the nonlinearity, the decision rule can be written in terms of $\bar{\theta}$, where $\bar{\theta}$ is the value of θ that corresponds to the nonlinearity, $\bar{h}$. Therefore, the indicator function $1\{h \leq \bar{h}\}$ can be written as $1\{\theta \leq \bar{\theta}\}$, and vice versa for the right-hand side of the nonlinearity.

Equation 3.1.3 can be transformed into a linear equation using a Taylor approximation around $\bar{\theta}$. The linear approximation is:

$$h = \left[\frac{\alpha\beta\bar{\theta}}{\tau + 1\{\theta \leq \bar{\theta}\}}\right]^{\frac{1}{1-\beta}} + \frac{1}{1-\beta}\left[\frac{\alpha\beta\bar{\theta}^{\beta}}{\tau + 1\{\theta \leq \bar{\theta}\}}\right]^{\frac{1}{1-\beta}}(\theta - \bar{\theta}) \qquad (3.1.4)$$

To construct a local linear regression from Equation 3.1.4, estimation occurs in the neighborhood of the $\bar{\theta}$ term, with the following coefficients:

$$\begin{aligned} a(\theta) &= \left[\frac{\alpha\beta\bar{\theta}}{\tau + 1\{\theta \leq \bar{\theta}\}}\right]^{\frac{1}{1-\beta}} \\ b(\theta) &= \frac{1}{1-\beta}\left[\frac{\alpha\beta\bar{\theta}^{\beta}}{\tau + 1\{\theta \leq \bar{\theta}\}}\right]^{\frac{1}{1-\beta}} \end{aligned}$$