Moral Hazard and Adverse Selection in Private Health Insurance[*]

David Powell[†]
Dana Goldman[‡]

May 7, 2014

Abstract

Moral hazard and adverse selection create inefficiencies in private health insurance markets. We use claims data from a large firm to study the independent roles of both moral hazard and adverse selection. Previous studies have attempted to estimate moral hazard in private health insurance by assuming that individuals respond only to the spot price, end-of-year price, expected price, or a related metric. There is little economic justification for such assumptions and, in fact, economic intuition suggests that the nonlinear budget constraints generated by health insurance plans make these assumptions especially poor. We study the differential impact of the health insurance plans offered by the firm on the entire distribution of medical expenditures without assuming that individuals only respond to a specific parameterized price. We use a new instrumental variable quantile estimation technique introduced in Powell [2013b] that provides the quantile treatment effects for each plan, while conditioning on a set of covariates for identification purposes. This technique allows us to map the resulting estimated medical expenditure distributions to the nonlinear budget sets generated by each plan. Our method also allows us to separate moral hazard from adverse selection and estimate their relative importance. We estimate that 79% of the additional medical spending observed in the most generous plan in our data relative to the least generous is due to adverse selection. The remainder can be attributed to moral hazard. A policy which resulted in each person enrolling in the least generous plan would cause the annual premium of that plan to rise by over $1,600.

Keywords: Price Elasticity, Health Insurance, Quantile Treatment Effects, Adverse Selection, Moral Hazard
JEL classification: I11, I13, C21, C23

1 Introduction

Moral hazard and adverse selection create inefficiencies in health insurance markets and result in a positive correlation between health insurance generosity and medical care consumption. It is difficult to disentangle moral hazard from adverse selection and isolate the primary driving force behind this observed correlation. The policy implications are very different, however, depending on the relative magnitudes of each source. This paper separates moral hazard and adverse selection for the health insurance plans offered by a large firm. Our method relaxes the assumptions in the literature typically employed to estimate moral hazard in health insurance markets.

Rising health costs have prompted greater interest in mechanisms to reduce health care spending. There is a large literature studying how health insurance design encourages medical care spending and there is evidence that cost-sharing reduces health care consumption (see Baicker and Goldman [2011] for a review). Many recent federal policies have focused on increasing cost-sharing as a means to pass costs to the consumer and discourage additional consumption of medical care. The Affordable Care Act (ACA) promotes cost-sharing in several ways, such as the introduction of a "Cadillac tax" in 2018 which taxes plans with high premiums and, presumably, generous cost-sharing arrangements. Health savings accounts, established in 2003, encourage the purchase of high deductible plans which have less generous coverage at low levels of annual medical expenditures. On the other hand, policies encouraging the purchase of generous health insurance plans have been shown to have meaningful effects on medical care consumption. Cogan et al. [2011] find evidence that the tax subsidy - which allows individuals to purchase insurance in pre-tax dollars - leads to more medical care consumption. As insurance rates increase due to the ACA, it is especially important to understand how benefit design impacts spending.

At the same time, there is interest in understanding adverse selection in insurance markets. Adverse selection is another impediment to efficiency in health insurance. A large literature documents the difficulties in separating adverse selection and moral hazard (Chiappori and Salanie [2000], Chiappori [2000], Finkelstein and Poterba [2004]). We use plan introduction as an exogenous shock to plan generosity and then estimate the medical expenditure distribution of each plan if enrollment were random. The difference in the observed medical care distribution and this estimated distribution driven solely by moral hazard quantifies the magnitude of selection. Our empirical strategy does not require restrictive

structural assumptions to isolate adverse selection from moral hazard.

In this paper, we analyze administrative health claims data from a large employer in the United States for 2005 and 2006. This employer offered only one insurance plan in 2005. In 2006, they offered three different PPO plans. These plans varied in generosity based on their deductible, stop loss, and coinsurance rate. We estimate the impact of each plan on the entire distribution of medical care consumption. Our identification strategy first predicts which plan each family will choose in 2006 based on observable characteristics. These predictions interacted with availability of those plans act as instruments for plan choice. We condition on individual characteristics to account for the independent effects of the covariates. Identification originates from the introduction of the plan choice in 2006 and the differential but predictable effect that this introduction had on different enrollees. This strategy allows us to estimate the causal effect of each plan. Once we have estimated the distribution of expenditures for each plan, we can compare the observed distribution that selected into the plan to the estimated distribution if enrollment were random, separately identifying a useful measure of adverse selection.

Our empirical strategy is to estimate the medical expenditure distribution for each plan if selection into the plan were random. We estimate these distributions non-parametrically instead of assuming that individuals respond to a specific price in the plan. We can map this estimated distribution directly to the non-linear budget constraint created by the health insurance plan. We can observe whether the medical spending distribution is especially responsive to the plan for annual expenditure levels above the deductible or stop loss. This mapping can provide insight into which aspects of a plan, such as the deductible or coinsurance rate, impact health care decisions. While there many possible mechanisms through which a plan can encourage additional consumption, a basic first step is simply to create this mapping between the plan and the distribution.

There is a long-standing interest in how responsive people are to the generosity of their health insurance plan. The RAND Health Insurance Experiment (Manning et al. [1987]) estimates are still widely considered the standard in this literature, though the validity of results from the 1970's to today's health care system is questionable. More recent studies have also estimated the relationship between insurance cost-sharing and health care consumption. However, there is no consensus about how to parameterize a health insurance plan. The RAND estimates assume that individuals respond only to the spot price - the out-of-pocket portion of the next $1 of medical care consumed. More recent studies have

assumed that people respond to a measure of the expected end-of-year price (Eichner [1997], Duarte [2012]) or the actual end-of-year marginal price (Eichner [1998], Kowalski [2012a]). These latter studies assume both that individuals have perfect foresight and that they only respond to the cost-sharing rate of the last $1 of medical care consumption.

Recent work has asked what the relevant price is in the presence of a non-linear health insurance plan. This literature has studied whether the spot price is a sufficient statistic or whether individuals are forward-looking in their medical care consumption decisions. Research has found evidence that future prices do impact current medical care consumption decisions. Aron-Dine et al. [2012] finds that individuals with the same spot price but different expected end-of-year prices have different consumption patterns, implying that individuals exercise at least some foresight in health consumption decisions. Einav et al. [2013b] finds similar evidence with prescription drugs using nonlinearities in Medicare Part D plans. In fact, additional evidence suggests that individuals act with even more foresight than simple within-year decisions. Alpert [2012] provides longer-term evidence using the announcement of Part D in 2003 that individuals may delay drug purchases when future prescription drug coverage becomes more generous. While this growing literature suggests that individuals are not entirely myopic, it does not imply that it is reasonable to assume that individuals can or do respond *only* to their end-of-year price.

In a related literature, a limited number of studies model individuals or households as potentially responding to the entire budget set generated by a health insurance plan. These studies (including Cardon and Hendel [2001], Einav et al. [2013a], Kowalski [2012b]) do not use variation across plans for identification, require strong structural assumptions to generate identification, and assume perfect foresight.

In the end, we are interested in how people respond to the price of medical care, but it is unclear how to define "price." In this paper, we study the impact of different health insurance plans on the entire distribution of medical care consumption. This test allows us to circumvent parameterizing plans by potentially uninformative metrics, imposing restrictive behavioral assumptions, or requiring individuals to solely respond to specific types of prices. The results can be interpreted as the medical expenditure distribution that we would observe if each person in the data were enrolled in the plan or, put differently, if there were no systematic selection into the plan. We will be able to explicitly test the assumption that individuals respond to the realized end-of-year marginal price.

We see our paper as making three important contributions. First, using a new

quantile estimation technique introduced in Powell [2013b], we generate - to our knowledge - the first estimates of the impact of the end-of-year price on the distribution of medical care expenditures. The literature has frequently estimated a mean effect or relied on conditional quantile techniques. Conditional quantile techniques are difficult to interpret in this context. By conditioning on variables, such as age, when estimating the 90th quantile, the estimates provide the elasticity at the 90th quantile of the distribution for a fixed age. People at the 90th quantile of the conditional distribution, however, are possibly near the bottom of the medical care distribution (e.g., at younger, healthier ages). It is difficult to interpret the estimates from a conditional quantile estimator as providing information about the impact of prices on the unconditional (on covariates) distribution. We use a quantile technique which allows for conditioning on covariates to improve identification, but the results can be interpreted as the impact on the outcome distribution.

Second, our method allows us to be agnostic about how health insurance plans impact medical care consumption. We estimate the impact of each plan on the distribution of medical expenditures with no parameterizations of the plans. We compare these distributions to the estimated distributions when we impose restrictions commonly made in the literature. We test the equality of the two distributions, allowing us to perform a straightforward test of the usefulness and accuracy of the restrictive assumptions.

Finally, we separate adverse selection and moral hazard, providing magnitudes for both. We observe plans that are similar but with clear ranks in terms of generosity. Because we estimate the causal distribution for each plan, we can compare the observed distribution - which is a function of both moral hazard and adverse selection - with the estimated distribution that we would counterfactually observe if there were no adverse selection. This difference identifies the magnitude and distribution of selection. Note the importance of our approach for this contribution as well. By estimating the impact of a plan non-parametrically (i.e., without parametric restrictions on how plans affect individual behavior), it is straightforward to compare the observed distribution with the estimated causal distribution. There is widespread interest in adverse selection of health insurance (Bundorf et al. [2012], Cardon and Hendel [2001], Carlin and Town [2008], Geruso [2013]), yet it is surprisingly rare in the health literature to find plan-specific estimates of the magnitude of adverse selection. Handel [2013] studies the effects of inertia on adverse selection and reports metrics of the per-person cost to a plan. In this paper, we report per-person selection costs for each plan and show the distribution of selection in each plan. We estimate adverse selection without restrictive structural assumptions. We also show that using the prior year's medical care expenditures

as a measure of selection overestimates the magnitude of adverse selection selection. We estimate that adverse selection accounts for 79% of the difference in medical care expenditures between the most and least generous plans.

In the next section, we discuss the merits of an approach that does not parameterize moral hazard by a response to a specific price with some basic economic reasoning. Section 3 discusses the data and empirical strategy. Section 4 details the estimator and the parameters that are estimated. Section 5 presents the results and we conclude in Section 6.

2   Theory

There are many reasons to believe that the entire budget constraint potentially matters when studying the effect of health insurance plans on each part of the medical care consumption distribution. We highlight three of these reasons.

First, we can consider a model with a standard utility function $U(c, m)$ where both consumption of goods c and medical care m are valued and preferences are convex. Assume that the person has perfect foresight and decides at the beginning of the year exactly how much medical care to consume. We can draw the budget constraint generated by a typical health insurance plan. In Figure 1, we include a deductible which generates a kink in the budget constraint. A stop loss point would generate a similar kink. The shape of the indifference curve follows directly from convex preferences. In this setup, it is possible that there is not a unique optimum due to the non-convexity of the budget constraint. This possibility is shown graphically in Figure 1. Say that we observe an individual on the second segment of the budget constraint (to the right of the kink). Given standard assumptions on preferences, we cannot rule out the possibility that small changes in the first segment of the budget constraint would change the individual's optimal health care consumption. The implication is that it would be inappropriate to assume that an individual only responds to the marginal price. While non-convexities in the budget constraint appear in other contexts, health insurance poses a situation where they are the norm and should not be ignored.

Second, it is not clear that we should assume perfectly convex preferences in the context of health care consumption. Episodes of care can generate consumption behavior which appears inconsistent with convex preferences. Individuals may decide between not receiving a specific treatment versus initiating an expensive set of treatments. Keeler et al. [1977] and Keeler and Rolph [1988] include arguments that any price of care variable must

account for these episodes. Again, the implication is that changes in one segment of the budget constraint may impact behavior on other segments.

Third, perfect foresight is a strong assumption, especially in the context of medical care consumption, which is a function of unforeseen health shocks. This assumption requires that all individuals know on January 1 the exact marginal price that they will be facing at the end of the year. It is likely that we observe many individuals facing a different marginal price at the end of the year than they would have anticipated. Consequently, it would be inappropriate to model individuals as responding to the end-of-year prices.

On the other hand, we think that it is unwise to assume that individuals are not forward-looking at all. Instead, we believe that the policy parameters of interest are the responses to the entire health insurance plan. A health insurance plan can encourage consumption through several mechanisms, such as reducing the marginal cost of care, reducing the spot price of care for part or all of the year, or reducing the price of larger episodes of care. While understanding the role of each of the mechanisms is interesting, we are primarily concerned with estimating the overall impact of a plan, which is an improvement upon simply assuming away some possible mechanisms. We take a necessary step back relative to the literature to understand the role of benefit design in impacting medical care consumption.

For example, a plan with a smaller deductible may encourage additional annual consumption because individuals reach the deductible earlier in the year and react partially to the reduced spot price. Individuals may also initiate expensive treatments because the total out-of-pocket payments are smaller given the low deductible. Even individuals that end up consuming relatively little care within a year may react to a low stop loss point because it reduces their expected payments in the year. Regardless of the mechanism, we are interested in how this plan impacts annual health care expenditures. We believe that understanding this overall impact is an important step within a literature that has frequently imposed restrictive assumptions or explicitly assumed away many of the listed mechanisms. The cost of our approach is that we are limited in the types of inferences that we can make since we only observe a limited number of health insurance plans offered by our firm. However, we believe that there is little theoretical justification for the parameterizations of plans frequently made in the literature. Our approach will provide important evidence about how health insurance plans affect medical care consumption.

3   Data and Empirical Strategy

3.1   Background

We study the impact of employer-sponsored health insurance on medical care spending. Traditional employer-sponsored health insurance plans are defined by three characteristics: the deductible, the coinsurance rate, and the stop loss. These parameters dictate cost-sharing based on annual medical care expenditures. Consumers pay the full cost of their medical care until they reach the deductible. They, then, are only responsible for a fraction of their costs, referred to as the "coinsurance rate." In our sample, we observe plans with coinsurance rates of 0.1 and 0.2. Finally, consumer risk is bounded by the stop loss - the maximum annual out-of-pocket payments by the consumer. After stop loss, the consumer faces a marginal price of zero for additional medical care.

These plans are defined by individual annual expenditures, but it is also common for plans to include a family deductible and family out-of-pocket maximum. In our analysis, we want to map the distribution of expenditures to the non-linear budget set created by the health insurance plan and these family-level parameters obscure this mapping. Consequently, we select on families with only one or two enrollees because they cannot reach the family deductible or stop loss and we are capable of mapping the entire individual budget constraint. For example, the individual deductible for the most generous plan is $200 and the family deductible is $400. By limiting our analysis to families with only one or two members, we can ignore family-level parameters. This is beneficial because we know that a person consuming $50 of medical care is below the deductible. If we included larger families, then this individual could be facing the coinsurance rate or even a marginal price of 0 due to high medical care consumption of family members.

In our data, we study a firm that offered only one plan in 2005, which we label as Plan A. In 2006, the firm offered three PPO plans of varying generosity, which we label as Plans B, C, and D. Plan B is the most generous 2006 plan with a low deductible, coinsurance rate, and out-of-pocket maximum. Plan C is less generous and Plan D is the least generous plan. Table 1 provides the relevant parameters for each of the plans in our data.

In Figure 2, we show the empirical non-linear budget constraints generated by our plans, holding constant income minus the premium across plans. Our goal in Figure 2 is to simply illustrate the shapes of the budget constraints for the first $18,000 of annual medical

expenditures. Most importantly, there are significant differences in the kink points where the slopes of the respective budget constraints change.

3.2   Identification Strategy

Our identification strategy relies on the introduction of plans with varying generosity and the differential effect that this introduction had on enrollees based on covariates. We use data which provide demographic information such as family size, age, sex, and relationship to employee. We also observe medical expenditures in our data for 2005 and 2006. We create "cells" based on the demographics - 2005 family size, age, sex, and relationship to employee (employee, spouse, or dependent). The mean cell size in our data is 58 people.

We use the changes in plan options in 2006 for identification. Each person in our sample was enrolled in Plan A in 2005. We can assume that many families would have liked to have been enrolled in a plan with different generosity. In 2006, they were given a different set of options and sorted according to their preferences. We use the created cells to predict which plan each family will enroll in. We estimate the probability of enrollment into each plan based on family characteristics. Given that the cells are based on variables that should independently affect medical care consumption, we condition on the cells themselves to isolate the differential effect of plan availability for identification.

The probabilities are our predictions of plan choice and the instruments are these probabilities interacted with plan availability. These instruments are only exogenous conditional on the covariates used to make the original prediction. Otherwise, we would simply be predicting that a family that prefers to enroll in the most generous plan based on its covariates also tends to consume systematically different amounts of medical care. Instead, we condition on indicators based on the same cells (2005 family size, age, sex, and relationship to employee). We think that people with covariates that predict high medical care consumption are also more likely to enroll in Plan B in 2006. This is not problematic given our empirical strategy because we compare each cell's distribution of medical care in 2006 to its distribution in 2005. The only factor that has changed is plan availability.[1]

The identification strategy, then, is to compare the distribution of expenditures for people with certain characteristics in 2005 to the distribution of expenditures for people with the same characteristics in 2006. In 2005, those people have a probability of enrollment

---

[1]We discuss later that we are also not assuming stationarity.

9

in Plan B, Plan C, or Plan D of 0. In 2006, they have positive predicted probabilities for those plans. These predicted probabilities vary by cell. Construction of our instrument is discussed more explicitly in Section 4.2.

3.3 Data

We use administrative claims data from a large firm included in the MarketScan Commercial Claims & Encounter and Benefit Plan Design Databases. The firm is a large manufacturing firm, and the employees reside in 44 different states. The workers are not unionized and are predominantly salaried (94%) and work full-time (84%). These data provide basic demographic information for each person and detailed information about inpatient and outpatient medical claims, including out-of-pocket and total costs. The data also provide information about plan choice and plan structure. We observe claims from the firm for 2005 and 2006 and we restrict our attention to employees for whom we observe insurance choice and spending in both 2005 and 2006. The benefit of observing multiple years is that we can compare the distributions of expenditures in 2005 and 2006, which will be important to our empirical strategy. We observe all individuals enrolled in a plan even if they do not consume any medical care. This is important given that not consuming any medical care is part of the distribution that we are interested in.

Another benefit of this firm is that the plans are identical in all ways except for the deductible, coinsurance rate, and stop loss. Consequently, given exogenous variation in plan enrollment, we can attribute differences in consumption behavior to the differences in these plan parameters. Furthermore, there was only one plan in 2005. Each person experienced the same plan generosity in 2005 and we can use a year fixed effect to account for the medical expenditure distribution in 2005.[2]

Summary statistics are presented in Table 2. We present the summary statistics by 2006 plan. As one might predict, the most generous plan attracts an older population with higher mean medical expenditures in the previous year. The mean age and 2005 medical expenditures decrease with plan generosity. Note also the medical expenditures are very skewed and the mean potentially masks important distributional factors.

We use the plan parameters and the individual's annual medical expenditures to

---

[2]While the lack of choice in 2005 is convenient because it allows us to represent the 2005 distribution by a year fixed effect (that varies throughout the distribution), the identification strategy would work similarly given multiple plans in 2005.

assign end-of-year prices to each person. An individual below the deductible is assigned a price of one. An individual above the deductible but below the stop loss is assigned the coinsurance rate (which varies by plan). An individual above the stop loss is assigned a price of 0.

3.4 Sample Selection

We select our sample on families with two or fewer members. As explained earlier, we want to exclude individuals that may potentially meet the family deductible or out-of-pocket maximum, and these thresholds can only be met by families with at least three members. Family-level parameters add a layer of complexity and it would be difficult to map the distribution of expenditures to the nonlinear budget set generated by the plan when people with the same medical expenditures may face different marginal prices due to family-level expenditures.

Next, we exclude children from our analysis and only use employees and their spouses. We also only use families that were enrolled for both 2005 and 2006, and we require them to remain in the same plan for all of 2006. Our analysis sample includes 10,094 families (15,115 people).

4   Empirical Model and Estimation

We use a quantile framework in our analysis for three reasons. First, a significant proportion of our analysis sample consumes no medical care within a year. This censoring can bias mean estimates. Quantile estimates are robust to censoring concerns without making strong distributional assumptions. Second, the distribution of medical expenditures is heavily-skewed. Mean regressions techniques may primarily reflect behavioral changes for people at the top of the expenditure distribution. In general, mean regression estimates are not necessarily representative of the impact at any part of the distribution. Third, a primary goal of this paper is to understand how insurance plans affect medical care consumption. If individuals are responding to the end-of-year marginal price, then we should observe that plans have a larger causal impact in the parts of the distribution above the deductible than the parts of the distribution below the deductible. Estimating a distribution, then, is important as we can map the quantile estimates to the plan parameters - the deductible and the stop loss - and observe whether the plan has larger impacts at parts of the distribution

11

where the end-of-year price is lower.

We are interested in estimating two equations. In the first equation, we assume that individuals only respond to the end-of-year marginal price. In the second equation, we assume that individuals respond to the plan, but we place no restrictions on this response. We will use the QTE framework introduced in Powell [2013b]. There are several advantages of this framework over the traditional conditional quantile frameworks and we will discuss the benefits in the context of each equation in sections 4.3 and 4.4. We discuss the IV-GQR estimator first.

4.1   IV-GQR Estimation

This paper uses IV-GQR, an estimator that generalizes more conventional quantile estimation techniques such as quantile regression (QR, Koenker and Bassett [1978]) and instrumental variables quantile regression (IV-QR, Chernozhukov and Hansen [2006]). We discuss the benefits of IV-GQR over traditional quantile estimators in this section and will focus on its utility relative to IV-QR, given instruments Z, treatment variables D, and control variables X. We will specify D in proceeding sections but discuss the estimator more generically here.

Traditional quantile estimators allow the parameters of interest to vary based on a nonseparable disturbance term, frequently interpreted as unobserved "proneness" (Doksum [1974]). In our context, this disturbance term can be interpreted as an individual's underlying tendency to consume medical care due to health, preferences for medical care, etc. As more covariates are added, however, the interpretation of the parameters in traditional quantile models changes as some of the unobserved proneness becomes observed. It is common in applied work to simply add covariates in a quantile regression framework. To illustrate why this is problematic, let us consider a case where prices are randomized. With randomized prices, one could simply perform a quantile regression of medical expenditures on prices. If we are interested in how prices impact the top of the distribution, we could estimate a quantile model for $\tau = 0.9$. However, we might want to condition on covariates as well. Adding these covariates in a traditional quantile framework changes the interpretation because the "high quantiles" now refer to people with high levels of medical care given their covariates. Many of these people may be at the bottom of the medical care distribution.

We note that other estimators were developed with similar motivations. Firpo et al.

[2009] introduced "unconditional quantile regression" (UQR) for reasons similar to those provided for IV-GQR. Powell [2013b] details many of the differences between the two estimators. We highlight three primary differences which are especially important in our context. First, UQR does not allow for endogenous variables. Second, it provides only a first-order approximation (Chernozhukov et al. [2013]). Third, UQR estimates the effect of small changes in covariates on the existing distribution. The existing distribution, however, is already treated. Consequently, UQR cannot be used to estimate the quantile functions of interest. IV-GQR estimates the expenditure distribution given various sets of policy variables.

It is also argued that one can estimate the conditional quantile functions and then integrate out the covariates using the procedure described in Machado and Mata [2005]. The Machado and Mata [2005] method is useful for decompositions, but it still relies on conditional quantile estimation. The procedure restricts the effect of the treatment variables to be the same for the bottom of the conditional distribution for a 25 year old and the bottom of the conditional distribution for a 60 year old. The method then integrates out the covariates (such as age) to determine the counterfactual unconditional distribution under different distributions of covariates. The conditional quantile restriction is still enforced though. A primary motivation of IV-GQR is to relax this assumption and provide more flexible methods for estimating quantile treatment effects.

Let $U^* \sim U(0,1)$ be a rank variable which represents proneness to consume medical care (normalized to be distributed uniformly). Powell [2013b] models proneness for the outcome variable as an unknown and unspecified function of "observed proneness" (X) and "unobserved proneness" (U): $U^* = f(X, U)$ where we also normalize $U \sim U(0,1)$. The specification of interest can be written as

$$Y = D'\beta(U^*), \quad U^* \sim U(0,1). \tag{1}$$

Following Chernozhukov and Hansen [2008], we are interested in estimate the Structural Quantile Function (SQF):

$$S_Y(\tau|d) = d'\beta(\tau). \tag{2}$$

The SQF defines the τth quantile of the outcome distribution given the policy variables if U^* and D were independent or, put differently, if each person in the data were subject to the policy variables $D = d$.

However, it is common and frequently necessary in applied work to condition on

additional covariates. IV-QR requires those covariates to be included in the structural model, altering the SQF. The specification is assumed to be:

$$Y = D'\tilde{\beta}(U) + X'\tilde{\delta}(U), \quad U \sim U(0,1). \tag{3}$$

The parameters are no longer assumed to vary by proneness, only the unobserved component of the disturbance term. The SQF is

$$S_Y(\tau|d,x) = d'\tilde{\beta}(\tilde{\tau}) + x'\tilde{\delta}(\tilde{\tau}). \tag{4}$$

where $\tilde{\tau}$ refers to the τth quantile of U, not U^*. A primary motivation of employing quantile techniques is that they allow for a nonseparable disturbance term. Adding covariates in the above way, however, separates this term into different components, undermining the original motivation. Put differently, adding control variables in a traditional quantile framework requires altering the structural quantile model. This property is undesirable in our application. Instead of treating the covariates in the same way as the policy variables, the IV-GQR estimator treats them differently. The covariates are allowed to inform the distribution of the disturbance term. An older person is likely to have a different distribution for U^* than a younger person. The IV-GQR estimator uses this information.

Table 3 provides concise comparisons between the IV-QR and IV-GQR estimators. With IV-QR, it is possible to estimate the SQF of interest (equation (2)) under the assumption that $U^*|Z \sim U(0,1)$. IV-GQR relaxes this assumption ($U^*|Z, X \sim U^*|X$), which will be necessary with our empirical strategy since our instruments are only conditionally independent. In short, IV-GQR compares conditional (on cells) distributions, but the parameters refer to the unconditional distribution.[3]

4.2 Instruments

Prices are mechanically related to medical expenditures due to the structure of the health insurance plans. An individual that consumes additional medical care may pass the deductible or stop loss and lower the marginal price of care. Similarly, plan choice is not random. Individuals that are predicting high medical care expenditures likely select into more generous plans.

---

[3]We use "unconditional" to mean unconditional on the covariates (cells). The resulting distribution depends on the treatment variables (price or plan choice).

We predict plan choice based on the demographic information in our data. Using age, sex, and family size, we can predict which plan the household chooses in 2006. This predicted probability is correlated with health status and medical care preferences through the demographic variables, but our estimation strategy will allow us to condition on these same demographic variables because we observe these households in 2005, when they were restricted to Plan A. The underlying experiment is to assume that for a given "cell" defined by age, sex, family size, and relationship to the employee, the proneness to consume medical care does not change over time in a systematic manner. The resulting change in medical care consumption (relative to the other cells) is due to plan generosity only. We believe that this is a plausible assumption in our context. We use these instruments to shock prices and, in separate regressions, plan enrollment. The instruments are (1) the probability of choosing Plan B (conditional on demographics) interacted with $\mathbf{1}(\text{Year} = 2006)$ and (2) the probability of choosing Plan C (conditional on demographics) interacted with $\mathbf{1}(\text{Year} = 2006)$. These are set to 0 in 2005 since those plans were not available. Thus, we do not use actual prices or plan enrollment for identification but, instead, the availability of new plans and the differential probability of enrollment based on demographics. Identification originates from changes in the probability of enrollment.

4.3   Price Elasticity

Our empirical strategy is to estimate the relationship between per-person medical care expenditures and health insurance generosity. The literature has commonly parameterized an insurance plan with one price measure. In our framework, we write the log of annual medical care expenditures as a function of the end-of-year price:

$$\ln M_{it} = \phi_t(U_{it}^*) + \delta(U_{it}^*)\left[\ln P_{it} \times \mathbf{1}\left(P_{it} > 0\right)\right] + \gamma(U_{it}^*)\mathbf{1}\left(P_{it} = 0\right), \quad U_{it}^* \sim U(0,1) \quad (5)$$

where U^* represents proneness for medical care consumption. We normalize this structural disturbance term to be uniformly distributed. The term is a rank variable which ranks people based on their consumption of medical care for a given price. P represents the end-of-year price. We are interested in estimating the Structural Quantile Function:

$$S_{\ln M}(\tau | P) = \phi_t(\tau) + \delta(\tau)\left[\ln P \times \mathbf{1}\left(P > 0\right)\right] + \gamma(\tau)\mathbf{1}\left(P = 0\right). \quad (6)$$

In this equation, $\delta(\tau)$ represents the price elasticity for the τth quantile of the distribution. Elasticities are only valid for positive prices so we include a separate term for people facing an end-of-year price of 0.

In our data, we have information about each person such as age, sex, family size, and relationship to the employee. It should be useful to condition on these variables as well and, in fact, it is necessary given our empirical strategy since our shocks to price and plan choice are only exogenous conditional on observed covariates. We let X represent our covariates. In a traditional (IV-QR) quantile framework, including these covariates changes the interpretation of the parameters. As an example, assume that we are only conditioning on age. IV-QR restricts the effect to be the same for 5th percentile of the distribution for individuals age 25 as the 5th percentile of the distribution for individuals age 60. This restriction is problematic given that the 5th percentiles of each group are very different.

We use the IV-GQR estimator introduced in Powell [2013b] to generate estimates of the price elasticity of medical care for the (unconditional) distribution. Note that traditional quantile methods cannot estimate the effect of prices on the distribution of medical expenditures when other covariates are included in the quantile regression. The IV-GQR estimator allows us to condition on covariates for identification purposes while still estimating equation (6).

4.4   Plan Elasticity

A primary motivation for this paper is to estimate individuals' responsiveness to health insurance plans without parameterizing the plan in a restrictive manner. We believe that this is especially worthwhile given the lack of evidence to support the parameterizations found in the literature. The estimation of QTEs using IV-GQR becomes even more important when we estimate these plan elasticities. We assume specification

$$\ln M_{it} = \phi_t(U_{it}^*) + \sum_k \beta_k(U_{it}^*) \left[\mathbf{1} \left(\text{Plan}_{it} = k \right) \right]. \tag{7}$$

The corresponding SQF is

$$S_{\ln M} = \phi_t(\tau) + \sum_k \beta_k(\tau) \left[\mathbf{1} \left(\text{Plan}_{it} = k \right) \right]. \tag{8}$$

Our goal is to estimate the distribution of medical care for each plan. The SQF will provide the resulting distribution for each plan if everyone in the sample were enrolled in that plan or there were no systematic selection into the plan. We can graph the resulting distribution for each plan along with the deductible and stop loss for that plan to observe whether the distribution responds to these parts of the plan. Conditional quantile estimators are uninformative in this context because we cannot map the quantiles to specific expenditure levels. A conditional quantile estimate would provide the impact of the plan for that quantile given a fixed age, sex, etc. For different covariates, this estimate would refer to different expenditure levels. For a 60 year old, a given quantile estimate may refer to a value above the stop loss. But the same quantile estimate may refer to a value near the deductible for a younger individual. We are interested in how the plan affects medical care spending and, consequently, we require that the estimates map to the same part of the cost-sharing schedule for each person in the data. IV-GQR provides such estimates.

Note the relative importance of the IV-GQR estimator in each context. When estimating the price elasticity, it is difficult to interpret traditional quantile estimates. When estimating the plan elasticity, traditional quantile estimates are essentially uninformative. We want the plan elasticity estimates to map directly to the plan which implies that we want to know the impact of the plan on the unconditional expenditure distribution.

4.5   Estimation

We implement the IV-GQR estimator to estimate equations (6) and (8). Focusing on the plan elasticity model, our model is

$$\ln M = \alpha(U^*) + \sum_k \beta_k(U^*) \left[\mathbf{1} \left(\text{Plan} = k \right) \right], \quad U^* \sim U(0,1) \tag{9}$$

$$Y = \max(\ln M, C), \tag{10}$$

$$U^* | Z, X \sim U^* | X, \tag{11}$$

$$\mathbf{1} \left(\text{Plan} = k \right) = \phi_k(Z, X, V) \quad \text{for all } k. \tag{12}$$

We make no assumptions on the functional form $\phi_k(\cdot)$ and no restrictions are placed on the disturbance term V which partially determines plan choice. Many individuals do not consume any medical care and we model these individuals as having censored medical ex-

penditures. We observe Y instead of $\ln M$ for these individuals. Quantile estimation is, generally, robust to censoring. We estimate the SQF for quantiles that are unaffected by censoring (i.e, quantiles where the SQF predicts $M > 0$).[4] Practically, we set the outcome variable for observations with no medical expenditures to a very low value. The exact number chosen has no impact on the final estimates.

The IV-GQR estimator simultaneously uses two moment conditions. We write the quantile function as $D'\beta(\tau)$ where $D'\beta(\tau)$ refers to the SQF defined by equation (6) or equation (8), depending on the specification being estimated.

$$E\left\{ Z\Big[\mathbf{1}(Y \leq D'\beta(\tau)) - \hat{\tau}_X\Big] \right\} = 0, \tag{13}$$

$$E[\mathbf{1}(Y \leq D'\beta(\tau)) - \tau] = 0. \tag{14}$$

where $\hat{\tau}_X$ is an estimate of $P(Y \leq D'\beta(\tau)|X)$. In words, IV-GQR uses X to determine the probability that the outcome variable is below the quantile function given the covariates. An older individual is less likely to have medical expenditures below the quantile function and the estimator uses this information. For comparison with a conditional IV-QR estimator, note that equation (13) is equivalent to IV-QR when $\hat{\tau}_X$ is replaced by τ. Put differently, when there are no covariates, IV-GQR reduces to IV-QR. This illustrates the benefit of covariates in the IV-GQR framework - it relaxes the assumption that $P(Y \leq D'\beta(\tau)|Z)$ is constant and, instead, allows X to affect this probability. As a reminder, our covariates are indicator variables based on the cells (age, sex, 2005 family size, relationship to employee) used to predict 2006 plan choice. Note that condition (14) ensures that estimates refer to the τth quantile of the unconditional (on covariates) distribution.

We use GMM to estimate the parameters of interest. The sample moment condition is

$$g_i(b) = Z_i\Big[\mathbf{1}(Y_i \leq D_i'b) - \hat{\tau}_X(b))\Big],$$

---

[4]Censoring is only problematic with quantile estimators if the quantile function itself is censored for any of the observations. Traditional quantile estimators include all variables in the quantile function so it is much more likely that at least some observations will be censored (e.g., if a variable has a large negative effect on the outcome and an observation has a high value of that variable, then the quantile function evaluated for that observation's covariates is likely censored). IV-GQR only includes the treatment variables - which take a limited set of values in our context - in the quantile function so the additional covariates cannot induce censoring issues. Our estimated SQFs at all values of the treatment variables imply positive medical expenditures and we are robust to censoring concerns.

$$\hat{g}(b) = \frac{1}{N} \sum_{i=1}^{N} g_i(b). \tag{15}$$

$$\widehat{\beta(\tau)} = \arg\min_{b \in \mathcal{B}} \hat{g}(b)' \hat{g}(b) \tag{16}$$

Where $\mathcal{B}$ is defined by

$$\mathcal{B} \equiv \left\{ b \quad | \quad \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(Y_i \leq D_i'b) = \tau \right\}.$$

Only b such that $\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(Y_i \leq D_i'b) = \tau$ are considered. This set defines $\mathcal{B}$. This constraint enforces equation (14) and has several computational benefits as discussed in Powell [2013b]. Most importantly, it makes simultaneous estimation of one of the parameters unnecessary, simplifying estimation. We use grid-searching to find $\widehat{\beta(\tau)}$ using equation (16). Thus, we "guess" b and evaluate the objective function for that guess. For each guess, we must estimate $P(Y_i \leq D_i'b|X_i)$. Powell [2013b] recommends a simple probit or logit model for this step due to computational conveniences and discusses how incorrect distributional assumptions do not necessarily bias the estimates. However, in our analysis, our covariates are dummy variables for each cell used to generate our instruments. These dummy variables saturate the space so no distributional assumptions are necessary. The estimator, in fact, reduces to a special case for fixed effects discussed in Powell [2013a].[5] Note that the conditional independence assumption for the Powell [2013a] estimator does not impose stationarity on the disturbance term. In our context, the distribution of the disturbance term is allowed to change over time within each cell. We place no restrictions on the conditional mean or variance of the underlying distribution of medical expenditures. The only assumption is that changes in this conditional distribution should be orthogonal to changes in the probability of enrollment in each plan. Given that we are identifying off the introduction of these plans and they were not introduced based on changes in variance for any specific group, this assumption seems plausible in our context.

We use subsampling (Politis and Romano [1994]) for inference.[6] All subsampling is performed at the family-level to account for possible intra-family clustering.

---

[5] Powell [2013a] shows that the estimates are consistent even for $T = 2$.

[6] Powell [2013b] recommends a weighted bootstrap, but given the size of our data set, we found that subsampling had computational advantages over bootstrap techniques.

4.6 Reported Parameters

We will present our results with graphs that show the parameters over the entire distribution. When applicable, our graphs will include the point where the distribution has passed the plan deductible or stop loss. Some caution in interpretation is necessary. Each point refers to the quantile in the distribution based on the end of the year expenditures. The estimates, then, are *not* comparing the behavior of a person right before and right after that person hits the deductible. Instead, the estimates below the deductible refer to people that never pass the deductible in that year while the estimate above the deductible refer to individuals that pass the deductible by the end of the year.

4.6.1 Price Elasticities

For the price elasticity estimates, we report the estimates for $\delta(\tau)$ and $\gamma(\tau)$. These estimates should be comparable in interpretation to those found in the literature.

4.6.2 Plan Elasticities

We report differences in the plan estimates, using one plan as a baseline. For example, we present a figure graphing the differences between the most generous and least generous plan, corresponding to $\beta_B(\tau) - \beta_D(\tau)$. We graph the estimates by quantile and mark which quantiles correspond to the deductible and stop loss thresholds for each plan. Presenting the results in this way allows us to test visually whether plans encourage additional expenditures for the part of the distribution that is above the deductible for the most generous plan but not for the least generous plan.

Furthermore, we can use the price elasticity estimates (Section 4.6.1) to simulate what the plan distributions would look like under the assumption that plans impact medical care consumption solely through the end-of-year price. We create a plan distribution defined

by a set of $\tilde{\beta}_k(\tau)$. We define the "parameterized" distribution of this plan by

$$
\tilde{\beta}_k(\tau) = \begin{cases}
\widehat{\phi_{2006}(\tau)} & \text{if} \quad \exp\left[\widehat{\phi_{2006}(\tau)}\right] < \text{Plan } k\text{'s Deductible} \\
\widehat{\phi_{2006}(\tau)} + \widehat{\delta(\tau)}\left[\ln(\text{Plan } k\text{'s Coinsurance Rate})\right] & \text{if} \quad \exp\left[\widehat{\phi_{2006}(\tau)}\right] \geq \text{Plan } k\text{'s Deductible} \\
\quad \text{and} \quad \exp\left[\widehat{\phi_{2006}(\tau)} + \widehat{\delta(\tau)}\left[\ln(\text{Plan } k\text{'s Coinsurance Rate})\right]\right] < \text{Plan } k\text{'s Stop loss} \\
\widehat{\phi_{2006}(\tau)} + \widehat{\gamma(\tau)} \\
\quad \text{if} \quad \exp\left[\widehat{\phi_{2006}(\tau)} + \widehat{\delta(\tau)}\left[\ln(\text{Plan } k\text{'s Coinsurance Rate})\right]\right] \geq \text{Plan } k\text{'s Stop loss}
\end{cases}
\tag{17}
$$

This is exactly the distribution that we would estimate if people responded purely to the end-of-year price. We can compare the resulting distribution generated by the estimates of $\beta_k(\tau)$ and $\tilde{\beta}_k(\tau)$. For inference, we employ a Cramér-von-Misses-Smirnov (CMS) test discussed in Chernozhukov and Fernández-Val [2005] which uses resampling to simulate the test distribution.

We estimate each quantile function separately. When creating the expenditure distributions caused by each plan, we use the Chernozhukov et al. [2010] method to rearrange quantiles when necessary.

4.6.3 Adverse Selection

We report adverse selection as the fraction of people that select into plan k that are below the estimated τ^{th} quantile for that plan, using the plan elasticity estimates (equation (8)). These estimates refer to the medical expenditures if the entire sample were exogenously enrolled in the plan, shutting down adverse selection. Consequently, we can look at the expenditure distribution of those actually enrolled in the plan. If the fraction of enrollees in the plan that have medical expenditures below $\widehat{\phi_{2006}(\tau)} + \widehat{\beta_k(\tau)}$ is smaller than τ, then this is evidence of adverse selection into that plan. Let N_k represent the number of people enrolled in plan k and $\mathcal{K}$ represent the set of people enrolled in plan k. We present the empirical probability

$$
\widehat{\psi_k(\tau)} = \frac{1}{N_k} \sum_{i \in \mathcal{K}} \mathbf{1}(Y_i \leq \widehat{\phi_{2006}(\tau)} + \widehat{\beta_k(\tau)}).
\tag{18}
$$

This equation represents the sample equivalent of the probability that an enrollee in plan k is below the τth SQF. $\widehat{\psi_k(\tau)} < \tau$ implies that the enrollees are consuming more medical care than expected and that the plan has adverse selection. We expect to find adverse selection for the most generous plan and relatively healthy people to enroll in the least generous plan.

We present graphs of the distribution of the adverse selection parameters by τ for each plan.

5 Results

5.1 First Stage

In the first step, we create instruments which predict plan choice. We use the demographic information in our data to predict which plan each family will select in 2006. In 2005, all families were constrained to choose Plan A. Identification originates from the availability of Plans B, C, and D in 2006 and the differential preferences for these plans. We predict these probabilities using the covariates. We condition on the same covariates in our regressions so that we are not simply capturing that households with preferences for generous plans are different than those with preferences for less generous plans.

It is first necessary that our predicted probabilities are actually predictive of plan choice, conditional on the covariates. Table 4 shows that there is a relationship. We construct the probability of choosing Plan B in 2006 and the probability of choosing Plan C in 2006. Plan D is the excluded category. We include year fixed effects and fixed effects for each demographic cell. We report partial F-statistics which represent the strength of the instruments in predicting each endogenous variable independent of the other. We find that the instruments have a strong relationship with the endogenous variables.

5.2 Price Elasticity Estimates

In this section, we provide estimates of equation (6). We present the results graphically. The price elasticity term $\delta(\tau)$ is presented with confidence intervals in Figure 3. We simultaneously estimate the effect of free marginal medical care. These results are presented in Figure 4. We present results only for quantiles in which the relevant parameters are identified. For example, at lower quantiles, people face the full price of care regardless of plan choice. Consequently, the price elasticity estimates are not identified until the expenditure distribution is above the deductible for the most generous plan (and, similarly, the distribution is smaller than the stop loss for the least generous plan). A similar point can be made for the effects of free marginal care. Note, however, this will not affect our interpretation when we use the price elasticity estimates to create the distribution inferred by the estimates in Section

22

5.3.[7]

The elasticity estimates are relatively constant throughout the distribution. We estimates an elasticity between -0.1 and -0.3 for most of the sample. However, we also find positive and significant estimates for a few quantiles, potentially suggesting misspecification and providing some of the first evidence that plan generosity affects medical care consumption through mechanisms other than the end-of-year marginal price. In general, however, the elasticities are similar to those found and reported by the RAND Health Insurance Experiment. An elasticity of -0.2 implies that a coinsurance rate of 0.2 would increase medical care consumption by 38%. The estimates in Figure 4 suggest that medical expenditures are very responsive to a marginal price of 0. The estimates are between 1 and 2.5, with the exception of quantile 50. An estimate of 1 implies that marginal price of 0 increases medical care by 172% while an estimate of 2.5 implies an increase of 1,118%.

5.3   Plan Elasticity Estimates

This section presents our main results. We estimate the SQF in equation (8) and then present the differences in the SQFs to show how the plans generate different distributions of medical care. In Figure 5, we present the differences in the distributions for the most generous plan (Plan B) relative to the least generous (Plan D). We also include markers signifying the deductibles and stop loss points for each plan. The figure shows the estimated distribution of Plan B (relative to Plan D) if there were no systematic selection into either plan and then maps that distribution to the kinks in the budget set generated by the plans' parameters. If people respond to the marginal end-of-year price, then we should see the plan elasticity increase immediately after the deductible. Note also that the coinsurance rates are different across these plans.

Before the Plan B deductible, we estimate little difference between Plans B and D (with a surprising exception that Plan D has higher expenditures at the very bottom of the distribution). Individuals with low medical care spending appear to be unaffected by differences in plan generosity. This finding makes sense in a model where individuals respond only to end-of-year prices or spot prices as the differences in plan generosity do not take effect at these low levels of annual expenditures.

---

[7]The problem with point identification at these low quantiles is that certain combinations of the two parameters could generate the same distribution. However, the distribution itself is point-identified even if the underlying parameters are not.

For annual medical expenditures exceeding the Plan B deductible, Plan B has higher expenditures, reaching an elasticity of 0.5 to 0.6 before the maximum. These estimates imply the Plan B causes individuals to consume 80% more medical care at these quantiles relative to their medical spending if they were enrolled in Plan D. The increase in medical care due to Plan B enrollment does not increase sharply at the deductible but, instead, begins to increase before the deductible and climbs steadily until reaching 0.5-0.6.

Due to the additional care that Plan B encourages individuals to consume, the Plan B distribution reaches the out-of-pocket maximum before the Plan D distribution reaches the deductible. Surprisingly, the maximum does not appear to change this elasticity. At the Plan B out-of-pocket maximum, the elasticity remains relatively constant. The elasticity gradually decreases as the distribution reaches the Plan D deductible. Despite the differences in coinsurance rates across these plans, there is still a large change in the relative prices of medical care at the Plan D deductible. We estimate a sharper effect of the Part D stop loss as the spending distribution due to Part D enrollment increases relative to the Part B distribution. When the marginal prices of both plans are 0, the distributions appear to be relatively equal, with the exception that Plan B enrollment has a positive effect at the very top of the spending distribution. Overall, Figure 5 provides evidence that the more generous plan appears to encourage additional medical care spending through most of the distribution, even in parts where the generosity differences are small.

While it is difficult to understand the mechanisms through which these plans affect the entire distribution, it is instructive to look at the distributions generated by the assumptions that individuals respond only to the end-of-year marginal price. We label this the "parameterized difference" and present the results in Figure 6. The resulting distributions are highly unrealistic and look very different from the less parametric results found in Figure 5. The comparison of these figures illustrates the value of our non-parametric approach. We will formally test the equality of these distributions in the next section.

We can perform the same exercise for Plans B and C. The difference in the resulting distributions is shown in Figure 7. Here, we see similar patterns as before. There is little difference in the distributions caused by the two plans until the Plan B deductible. The estimated coefficients after the deductible are between 0.2 and 0.3, implying a 22%-35% increase in spending due to enrollment in Plan B relative to Plan C. These estimates are smaller than the previous comparison. While Plan C is more generous than Plan D overall, there is no difference in generosity at this point in the cost-sharing schedule.

The relative distributions appear unaffected by the Plan C deductible or the Plan B maximum. As before, the distributions converge when the end-of-year price is 0 in both plans. However, we estimate that Plan B enrollment has a large effect on the top of the distribution, despite the equality of end-of-year prices. There are several possible reasons for this such as differences in the prices of episodes of care for expensive treatments.

The counterfactual distributions again illustrate that assuming that individuals respond only to the end-of-year price lead to very different conclusions. Figure 8 presents these results.

For the sake of completeness, we also compare Plan C to Plan D, though the conclusions can be inferred from the other comparisons. Figures 9 and 10 presents these estimates.

5.4   Equality of Distributions Tests

For each plan, we can also formally test the equality of the distributions generated by our non-parametric method (estimation of SQF (8)) and the parametric method which assumes that individuals respond solely to the end-of-year marginal price. While the distributions look very different, we would like to test these differences statistically.

We use a Cramér-von-Misses-Smirnov (CMS) type test and simulate the distribution of this test statistic using subsampling. The CMS test for Plan B rejects the equality of the two distributions. In fact, the test statistic is larger than any value in the simulated distribution. The equality of distribution is also rejected at the 1% level for Plan C and at the 10% for Plan D. Overall, the graphs and the CMS tests suggest that an assumption that individuals respond solely to the end-of-year price is a particularly poor one that cannot be justified empirically. Consequently, we use the non-parametric distributions to generate our adverse selection estimates.

5.5   Adverse Selection

Next, we present our metric of adverse selection. Without adverse selection, the observed plan distributions and the causal distributions would be the same, implying that $P(Y_i \leq \widehat{\phi_{2006}(\tau)} + \widehat{\beta_k(\tau)}) = \tau$. Graphically, we would see a 45-degree line for each plan. The intuition behind our metric is that once we have estimated the causal distribution of a plan, we can compare the observed distribution with the distribution if there were no systematic

selection. In other words, once we net out adverse selection, the difference in the observed distribution and this estimated distribution provides information about the magnitude of adverse selection.

We estimate our metrics and present them in Figure 11. We include the 45-degree line as well. If the adverse selection metric is above the 45-degree line, then that is evidence of favorable selection. For example, Plan D appears to attract an especially healthy population. With no systematic selection, we would expect 20% of the Plan D enrollees to have expenditures below the estimated 20th quantile of the SQF for Plan D, which is equal to $193.45. Instead, we observe that almost 40% of the enrollees have smaller expenditures than $193.45. This favorable selection extends throughout the distribution.

Plan B shows evidence of adverse selection, especially at the bottom of the distribution. We estimate that without selection, the 15th quantile of the medical care distribution for Plan B would be $90.45. Only 12% of Part B enrollees have smaller expenditures than this amount. The systematic selection into Plan B disappears close to the top of the expenditure distribution.

Plan C shows a mix of favorable and adverse selection. We estimate that without selection, the 90th quantile of the medical care distribution for Plan C would be $6073.91. Only 88% of Part C enrollees have smaller expenditures than this amount. However, at the estimated 75th percentile, we find that 80% of enrollees have smaller expenditures.

Selection for Plans B and C are difficult to observe in Figure 11. We present the same estimates in Figure 12. In Figure 12, we subtract the quantile so that all of the selection estimates are centered around 0. In other words, no systematic selection would imply that a plan is centered around 0. Here, we see that favorable selection is especially notable for Plan D at the bottom of the distribution. We can also more clearly see evidence of adverse selection for Plan B and favorable selection for Plan C.

More formally, we can compare the observed distribution with the estimated causal distributions using the same CMS test as Section 5.4. These distributions would be the same in the absence of systematic selection. We reject the equality of distributions at the 1% level for Plans C and D, implying that there is systematic selection. We can reject non-systematic selection at the 10% level for Plan B.

5.6   Relative Importance of Moral Hazard and Selection

While we have presented several metrics involving the distribution of medical expenditures, we can also look at the overall importance of the causal impact of the plan on mean expenditures and selection. Given estimates for SQF (8), we can integrate over all quantiles to arrive at the mean medical expenditures for each plan if there were no systematic selection into the plan. These metrics are the expected per-person medical expenditures for a given plan if everyone in our sample were subject to that plan. The calculation for Plan B is the following:

$$\hat{E}\left[\text{Per-Person Medical Expenditures in Plan B with Random Selection}\right] = \int_{\tau} \left[\widehat{\phi_{2006}(\tau)} + \widehat{\beta_B(\tau)}\right]$$
$$(19)$$

We label these "Per Person Expenditures with Random Selection" in Table 5 because all differences across plans are driven solely by moral hazard. The first row is the actual per-person expenditures which includes moral hazard and adverse selection. Note that, for the sake of consistency, we calculate the actual expenditures in a similar manner by using the values of the quantile endpoints and integrating over τ. Consequently, the numbers are slightly different from those found in Table 2.[8]

We also include "Adverse Selection" which eliminates the causal impact of the plan and describes the expenditures of the individuals selecting into the plan if the plan itself did not impact expenditures. We simply subtract the moral hazard estimate from the per-person expenditures estimate to estimate selection. In the previous section, we tested the equality of the observed and estimated (causal) distributions as a test for adverse selection. We also see evidence of selection in the mean estimates. Again, we find statistical evidence of systematic selection. Under the assumption that differences in premiums across plans only reflect differences in expected insurer payments, our selection estimates provide evidence about the ramifications of policies which change enrollment behavior. For example, the Cadillac tax may encourage enrollment in less generous plans. Our estimates suggest that if our entire sample enrolled in Plan D that the premium would increase by over $1,600.

Table 6 repeats the results in Table 5 but provides complementary evidence by using comparisons between plans. We estimate that enrollment in Plan B increases per-person medical expenditures by almost $900 relative to Plan D and almost $700 relative to Plan C.

---

[8]We should also highlight that the standard errors in Table 5 represent the standard errors for the mean estimates and are not comparable to the standard deviations found in Table 2.

We can also estimate differences in selection and calculate the fraction of the differences in observed per-person costs across plans that can be attributed to selection. We estimate that 79% of the additional spending in Plan B can be attributed to adverse selection.

5.7 Adverse Selection Metrics

Our empirical strategy allows us to calculate precise estimates of systematic selection into each plan. It is useful to compare this method to an alternative metric of selection - previous year's medical expenditures. To test for selection, it might seem reasonable to observe whether individuals with higher medical expenditures in 2005 choose Plan B. Since all individuals were enrolled in the same plan in 2005, the differences in 2005 expenditures across 2006 plans reflect differences in selection.

However, these differences in 2005 expenditures do not reflect the true magnitude of selection. Individuals have private information about *changes* in health. Furthermore, individuals with high medical expenditures may, on average, expect to require less care in the next year due simply to mean reversion and health improvements, but may still value the additional financial risk protection of the most generous plan.[9] Our results suggest that last year's medical expenditures overstate the magnitude of adverse selection. Referring back to Table 2, we find that the difference in 2005 medical expenditures between Plan B and Plan D enrollees is $3,893. However, in Table 5, we see that the difference in selection in 2006 expenditures is only $2,999. Similarly, the difference in 2005 medical expenditures between Plan B and Plan C enrollees is $2,441. But, in 2006 expenditures, selection accounts for only $1,457. These differences are economically meaningful and highlight the benefits of estimating adverse selection in the same year that the selection is occurring.

6   Conclusion

Understanding moral hazard and adverse selection in private health insurance is widely-recognized in the field as an important endeavor. While the literature has frequently estimated the effect of price on medical care consumption, it has typically resorted to parameterizing the mechanism through which individuals respond to cost-sharing. We show that these assumptions typically contradict economic reasoning, and we provide empirical evidence that these specifications perform poorly. In this paper, we estimate the impact

---

[9]All three 2006 plans provide full coverage above the stop loss point, but individuals may still value the financial risk protection at lower levels of annual expenditures.

of different health insurance plans on the entire distribution of medical care consumptions using a new instrumental variable quantile estimation method. These estimated distributions are the distributions caused by the plans in the absence of systematic selection into plans. We map these causal distributions to the parameters of the plans themselves. We find some evidence that the medical care distributions respond to the deductible and stop loss. However, we can statistically reject that individuals only respond to the end-of-year price.

We also estimate the magnitude of adverse selection. We find favorable selection in the least generous plan and adverse selection in the most generous. We estimate that adverse selection is responsible for almost $1,400 of additional per-person costs in the most generous plan, implying that an individual considering this plan would pay over $100 per month in additional premium payments simply to cover the expected costs of the population selecting into the plan. Similarly, a policy which resulted in our entire sample enrolling in the least generous plan would cause annual premiums for that plan to rise by over $1,600.

We estimate that adverse selection is responsible for 79% of the differences in expenditures between the most and least generous plans. Moral hazard accounts for the other 21%. In the absence of moral hazard, the difference across these plans would be $2,999 instead of $3,782. Finally, we find that using the previous year's medical expenditures as a metric of selection greatly overstates the magnitude of selection.

References

Abby Alpert. The anticipatory effects of Medicare Part D on drug utilization. 2012.

Aviva Aron-Dine, Liran Einav, Amy Finkelstein, and Mark R Cullen. Moral hazard in health insurance: How important is forward looking behavior? Technical report, National Bureau of Economic Research, 2012.

Katherine Baicker and Dana Goldman. Patient cost-sharing and health care spending growth. *Journal of Economic Perspectives*, 25(2):47–68, 2011.

M Kate Bundorf, Jonathan Levin, and Neale Mahoney. Pricing and welfare in health plan choice. *The American Economic Review*, 102(7):3214–3248, 2012.

James H Cardon and Igal Hendel. Asymmetric information in health insurance: evidence from the national medical expenditure survey. *RAND Journal of Economics*, pages 408–427, 2001.

Caroline Carlin and Robert Town. Adverse selection, welfare and optimal pricing of employer-sponsored health plans. 2008.

Victor Chernozhukov and Iván Fernández-Val. Subsampling inference on quantile regression processes. *Sankhyā: The Indian Journal of Statistics*, pages 253–276, 2005.

Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 2006.

Victor Chernozhukov and Christian Hansen. Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1):379–398, January 2008.

Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

Victor Chernozhukov, Iván Fernandez-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

Pierre-Andre Chiappori. Econometric models of insurance under asymmetric information. In *Handbook of insurance*, pages 365–393. 2000.

Pierre-André Chiappori and Bernard Salanie. Testing for asymmetric information in insurance markets. *Journal of political Economy*, 108(1):56–78, 2000.

John F. Cogan, R. Glenn Hubbard, and Daniel P. Kessler. The effect of tax preferences on health spending. *The National Tax Journal*, 64(3):795, 2011.

Kjell Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Statist.*, 2(2):267–277, 1974.

Fabian Duarte. Price elasticity of expenditure across health care services. *Journal of Health Economics*, 31(6):824–841, 2012.

Matthew J Eichner. The demand for medical care: What people pay does matter. *The American Economic Review*, 88(2):117–121, 1998.

Matthew Jason Eichner. *Medical expenditures and major risk health insurance.* PhD thesis, Massachusetts Institute of Technology, 1997.

Liran Einav, Amy Finkelstein, and Stephen P Ryan. Selection on moral hazard in health insurance. *American Economic Review*, 103(1):178–219, 2013a.

Liran Einav, Amy Finkelstein, and Paul Schrimpf. The response of drug expenditures to non-linear contract design: Evidence from Medicare Part D. Technical report, National Bureau of Economic Research, 2013b.

Amy Finkelstein and James Poterba. Adverse selection in insurance markets: Policyholder evidence from the U.K. annuity market. *Journal of Political Economy*, 112(1):183–208, 2004.

Sergio Firpo, Nicole M. Fortin, and Thomas Lemieux. Unconditional quantile regressions. *Econometrica*, 77(3):953–973, 05 2009.

Michael Geruso. Selection in employer health plans: Homogeneous prices and heterogeneous preferences. 2013.

Ben Handel. Adverse selection and inertia in health insurance markets: When nudging hurts. *The American Economic Review*, 2013.

Emmett B Keeler and John E Rolph. The demand for episodes of treatment in the health insurance experiment. *Journal of Health Economics*, 7(4):337–367, 1988.

Emmett B Keeler, Joseph P Newhouse, and Charles E Phelps. Deductibles and the demand for medical care services: The theory of a consumer facing a variable price schedule under uncertainty. *Econometrica: Journal of the Econometric Society*, pages 641–655, 1977.

Roger W Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, January 1978.

Amanda E Kowalski. Censored quantile instrumental variable estimates of the price elasticity of expenditure on medical care. Technical report, 2012a.

Amanda E Kowalski. Estimating the tradeoff between risk protection and moral hazard with a nonlinear budget set model of health insurance. Technical report, National Bureau of Economic Research, 2012b.

José AF Machado and José Mata. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20(4):445–465, 2005.

Willard G Manning, Joseph P Newhouse, Naihua Duan, Emmett B Keeler, and Arleen Leibowitz. Health insurance and the demand for medical care: evidence from a randomized experiment. *The American Economic Review*, pages 251–277, 1987.

Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):2031–2050, 1994.

David Powell. Quantile regression for panel data with exogenous or endogenous regressors and a nonseparable disturbance. Working papers, RAND Corporation Publications Department, 2013a.

David Powell. A new framework for estimation of quantile treatment effects: Nonseparable disturbance in the presence of covariates. Working papers, RAND Corporation Publications Department, 2013b.

Figures

Figure 1: Indifference Curve and Non-Linear Budget Constraint

Consumption (All Other Goods)

Indifference Curve

Two intersections

Total Annual Medical Expenditures
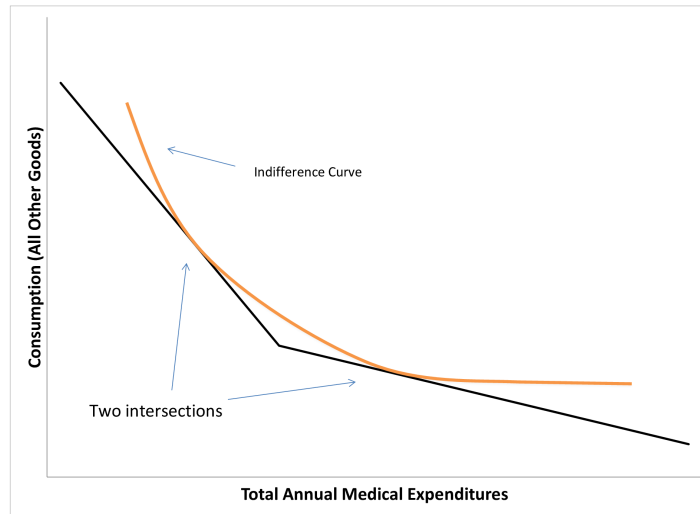
Notes: This figure graphs consumption of all other goods as a function of total annual medical expenditures. The indifference curve assumes convex preferences.

Figure 2: Budget Constraints Generated by Plans in Data
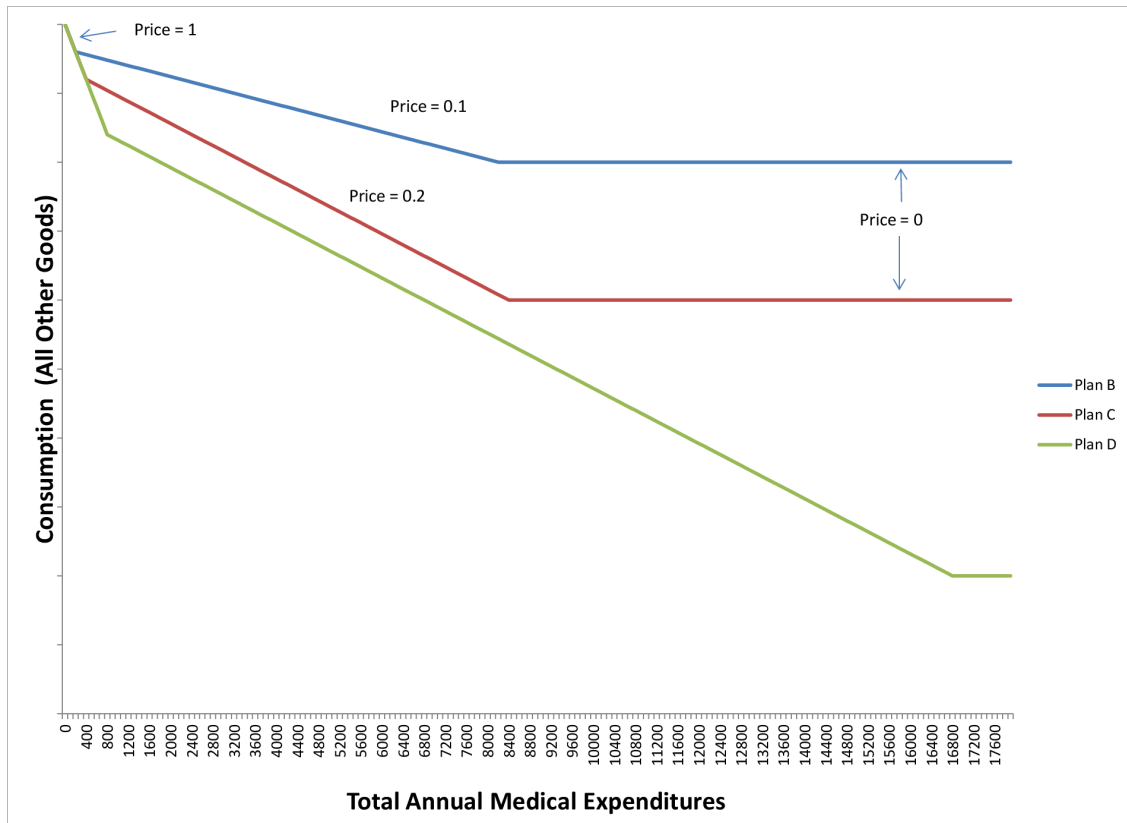
Notes: This figure graphs the nonlinear budget set generated by each 2006 plan, holding income minus the premium constant.

Figure 3: Price Elasticity Estimates
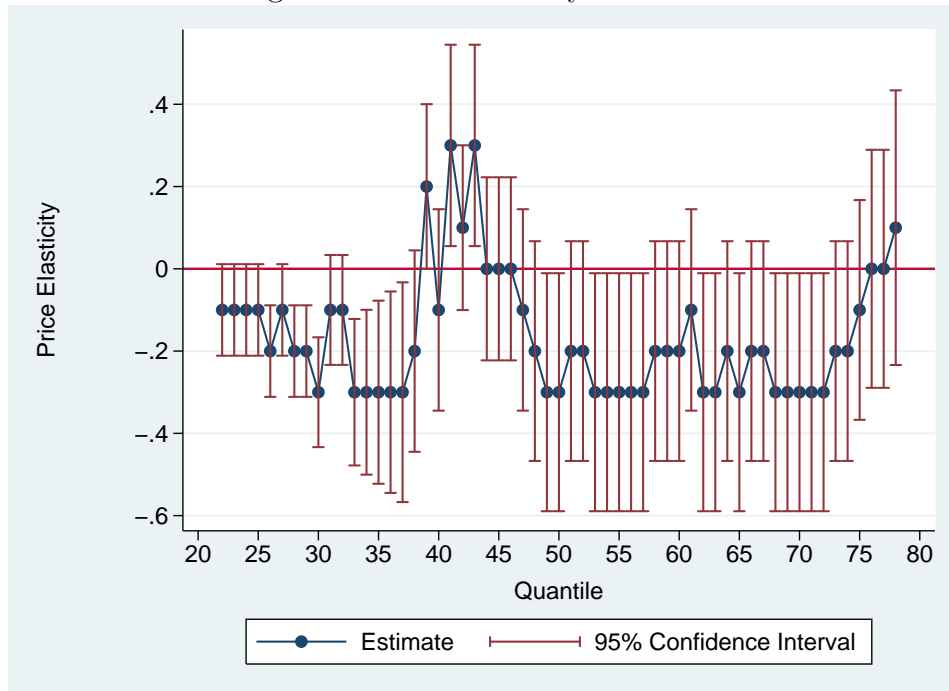
Notes: Using an instrumental variable quantile regression estimator, we estimate the price elasticity and the effects of a marginal price of 0, which are presented in Figure 4. Confidence intervals generated using clustered subsampling.

Figure 4: Effect of Price=0 on Medical Expenditures
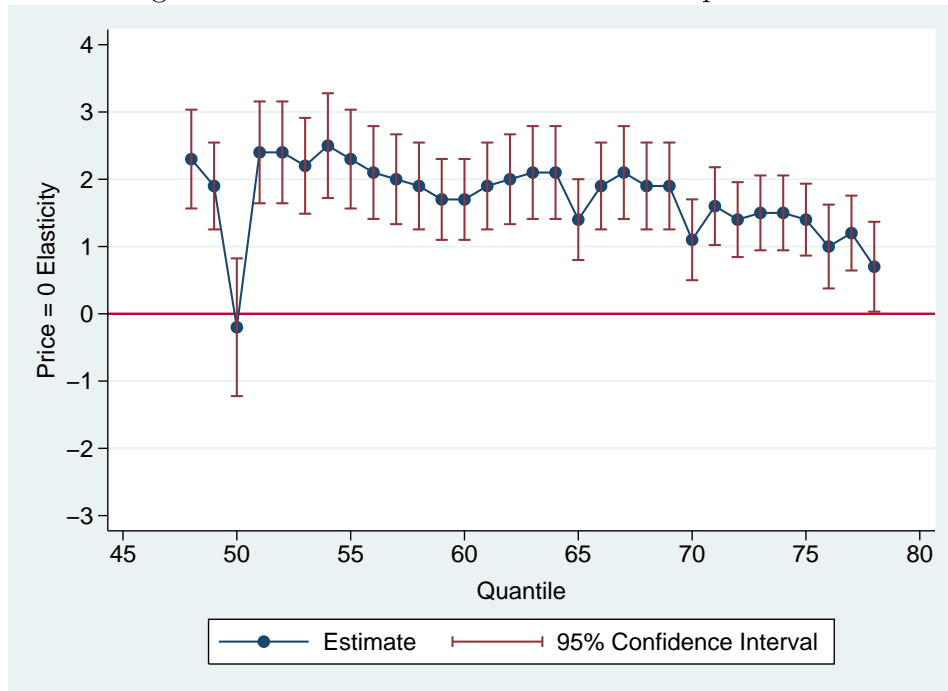
Notes: Using an instrumental variable quantile regression estimator, we estimate the price elasticity (presented in Figure 3) and the effects of a marginal price of 0, presented here. Confidence intervals generated using clustered subsampling.

Figure 5: Difference in Expenditure Distribution: Plan B vs. Plan D
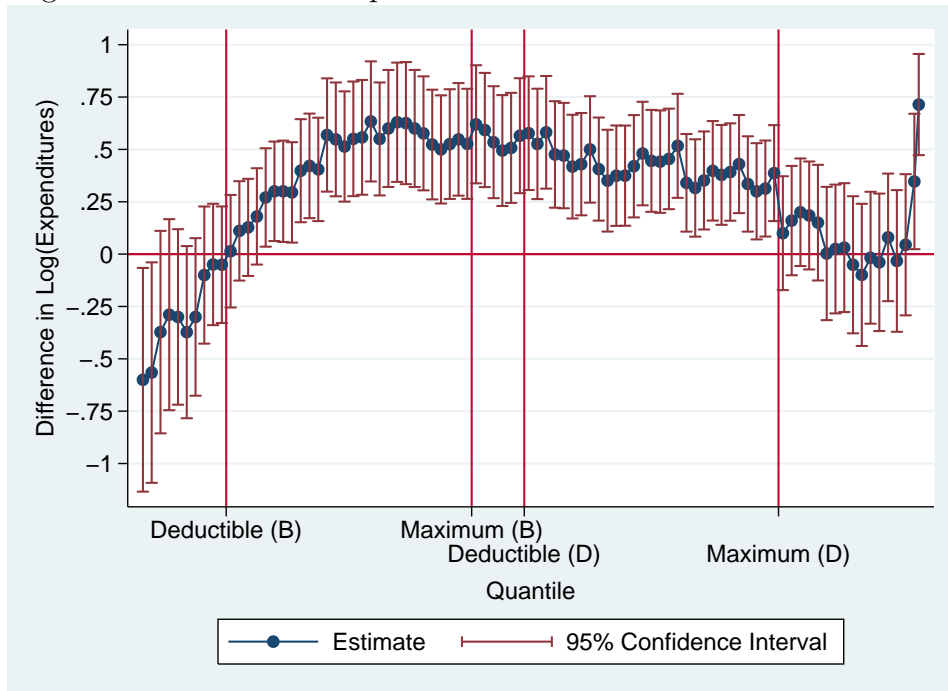
Notes: Using an instrumental variable quantile regression estimator, we estimate the distribution of Plan B and Plan D if enrollment into each plan were random. We graph the difference in these distributions here. Confidence intervals generated using clustered subsampling.

Figure 6: Parameterized Difference in Expenditure Distribution: Plan B vs. Plan D
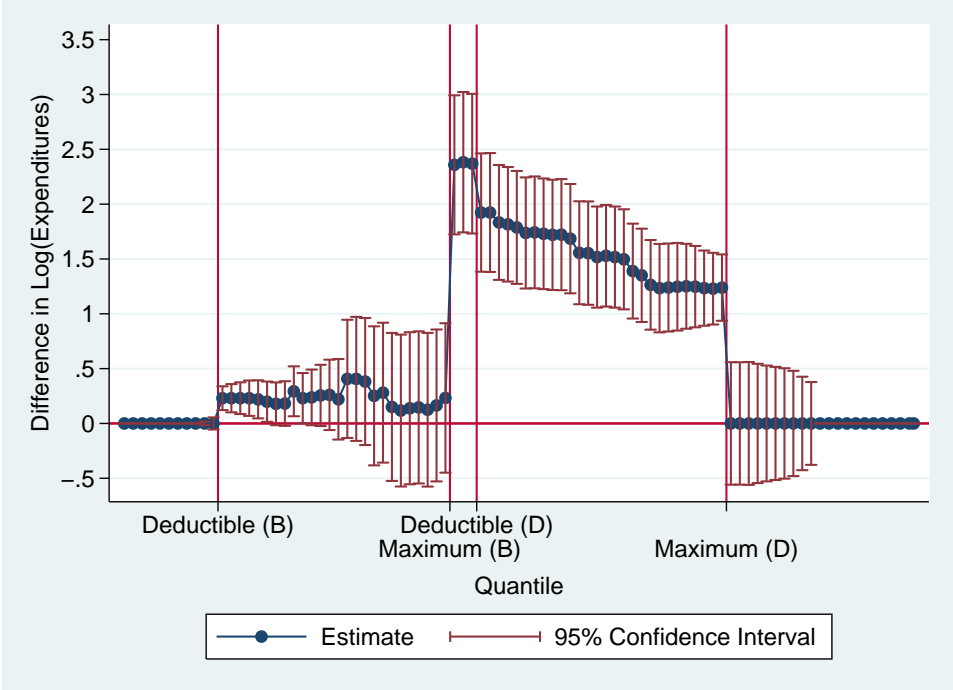
Notes: We use the price elasticity estimates shown in Figures 3 and 4 to generate the distributions that we would observe given random enrollment into Plans B and D under the assumption that individuals respond only to the end-of-year marginal price. Confidence intervals generated using clustered subsampling.

Figure 7: Difference in Expenditure Distribution: Plan B vs. Plan C
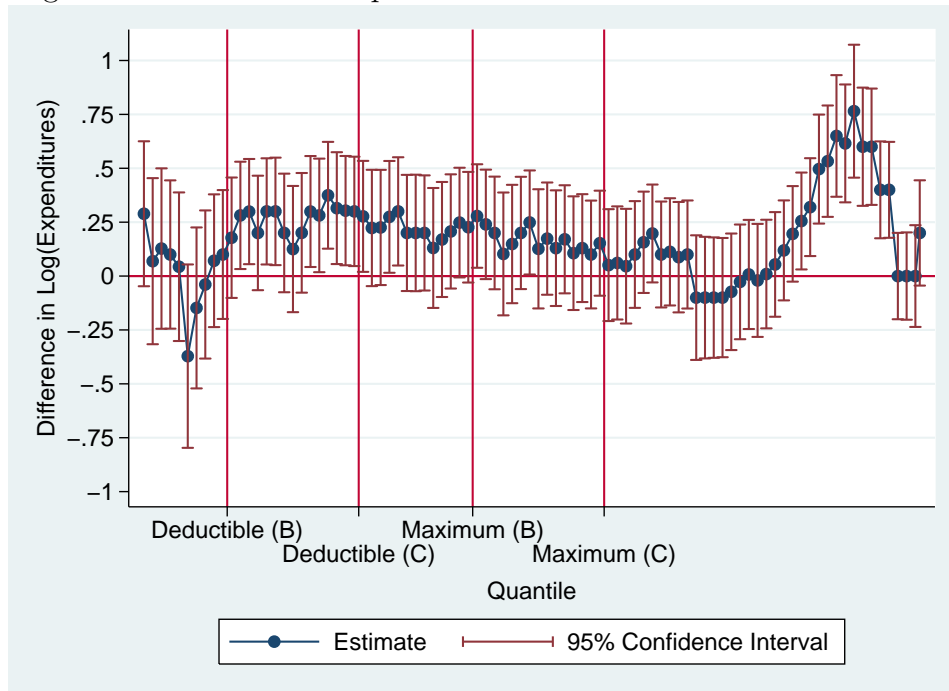
Notes: Using an instrumental variable quantile regression estimator, we estimate the distribution of Plan B and Plan C if enrollment into each plan were random. We graph the difference in these distributions here. Confidence intervals generated using clustered subsampling.

Figure 8: Parameterized Difference in Expenditure Distribution: Plan B vs. Plan C
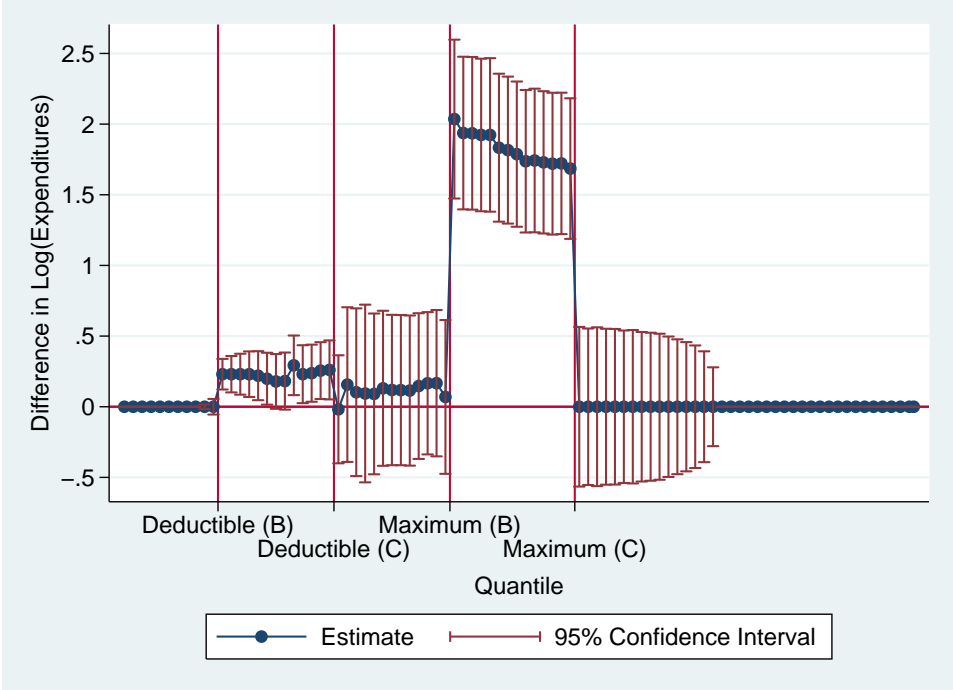
Notes: We use the price elasticity estimates shown in Figures 3 and 4 to generate the distributions that we would observe given random enrollment into Plans B and C under the assumption that individuals respond only to the end-of-year marginal price. Confidence intervals generated using clustered subsampling.

Figure 9: Difference in Expenditure Distribution: Plan C vs. Plan D
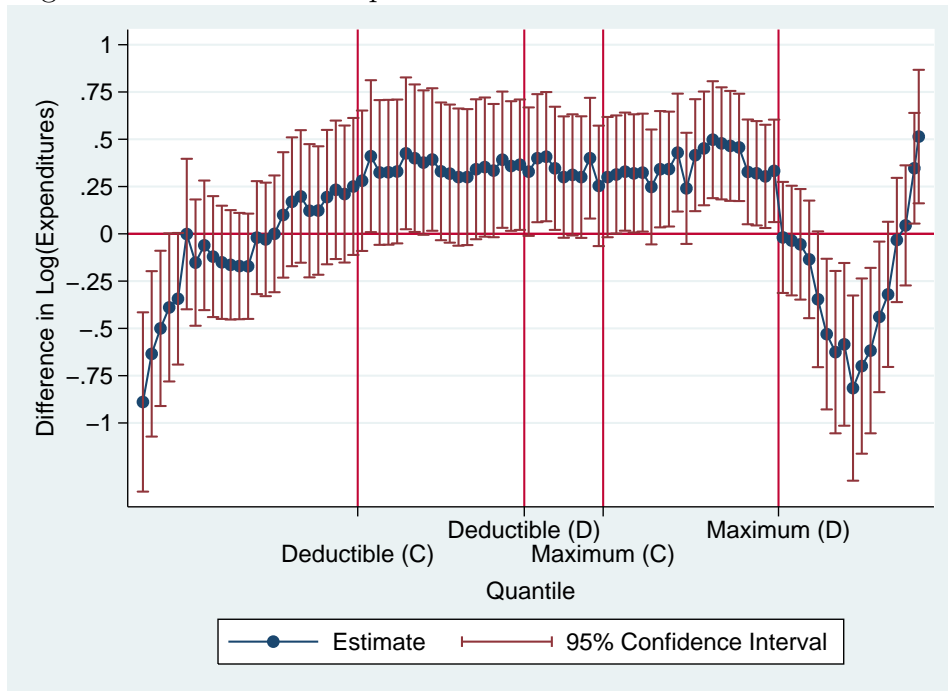
Notes: Using an instrumental variable quantile regression estimator, we estimate the distribution of Plan C and Plan D if enrollment into each plan were random. We graph the difference in these distributions here. Confidence intervals generated using clustered subsampling.

Figure 10: Parameterized Difference in Expenditure Distribution: Plan C vs. Plan D
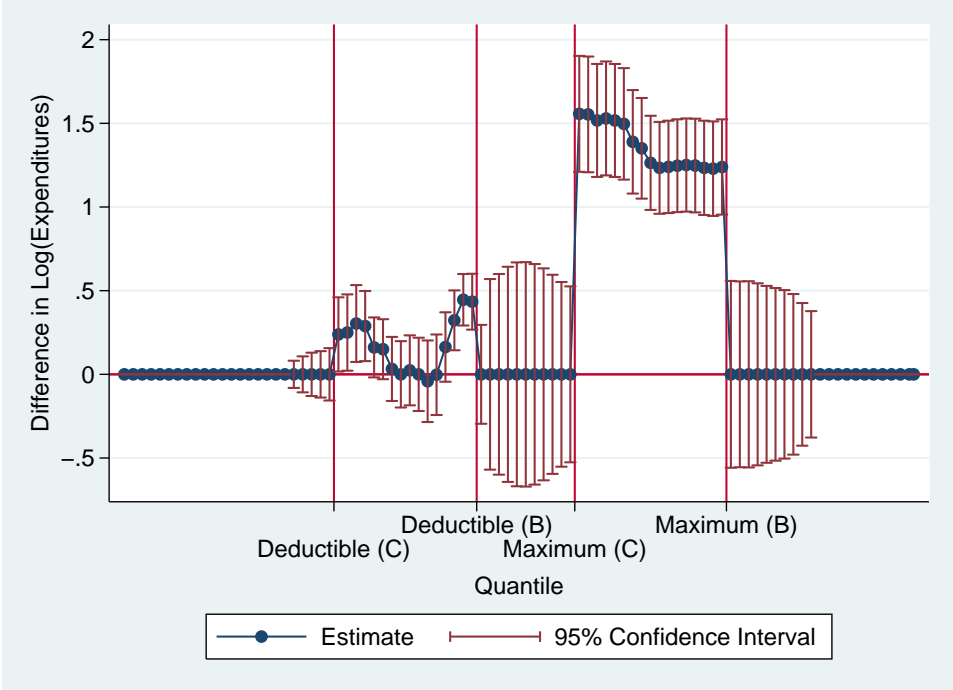
Notes: We use the price elasticity estimates shown in Figures 3 and 4 to generate the distributions that we would observe given random enrollment into Plans C and D under the assumption that individuals respond only to the end-of-year marginal price. Confidence intervals generated using clustered subsampling.

Figure 11: Adverse Selection
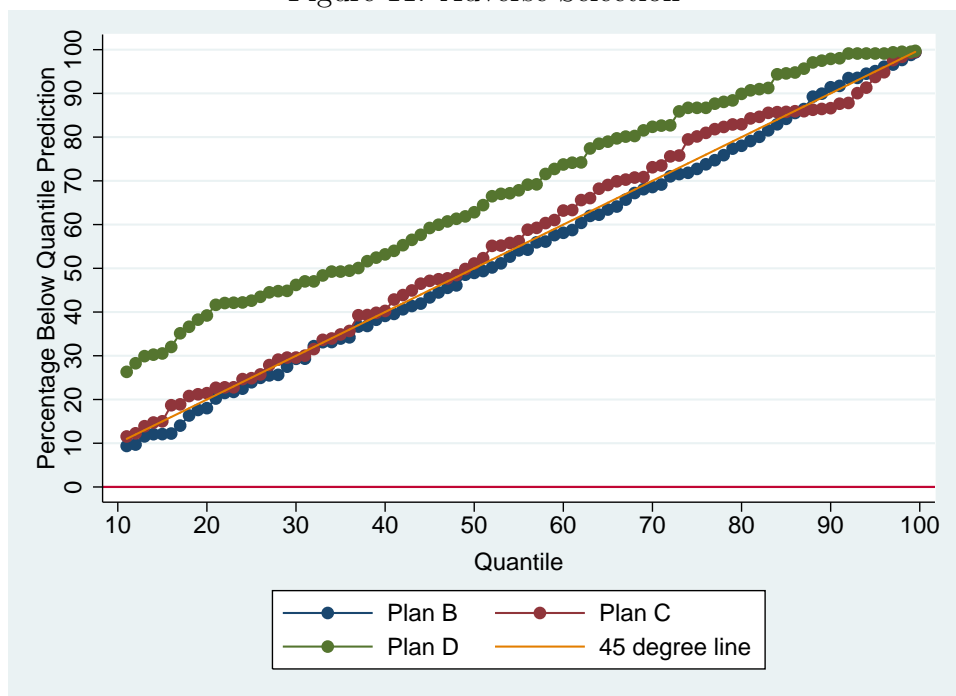
Notes: We use the plan elasticities presented in Figures 5, 7, 9 to estimate the empirical probability that an enrollee in the plan is below the estimate quantile function for that plan. Confidence intervals generated using clustered subsampling. The 45 degree line represents a plan with no systematic selection.

Figure 12: Adverse Selection
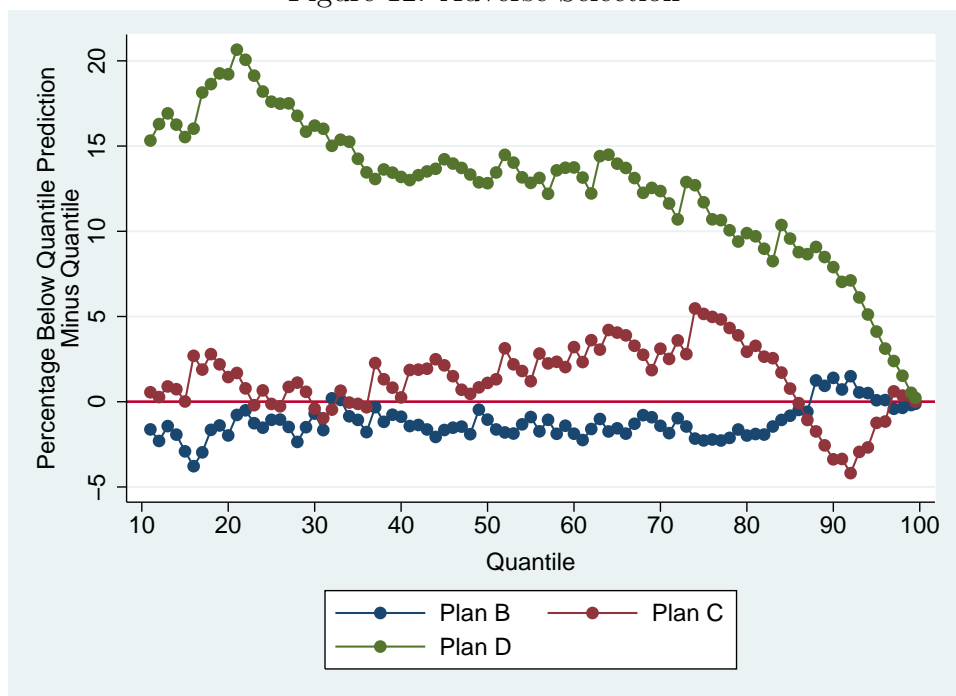
Notes: We use the plan elasticities presented in Figures 5, 7, 9 to estimate the empirical probability that an enrollee in the plan is below the estimate quantile function for that plan. We graph this probability minus the quantile. The 0-line represents a plan with no systematic selection. Confidence intervals generated using clustered subsampling.

44

Tables

	Plan (Year)	Plan A (2005)	Plan B (2006)	Plan C (2006)	Plan D (2006)
Deductible	Individual	$0	$200	$400	$800
	Family	$0	$400	$800	$1,600
Stop loss	Individual	None	$1,000	$2,000	$4,000
	Family	None	$2,000	$4,000	$8,000
Coinsurance		20%	10%	20%	20%
Plan Type		PPO	PPO	PPO	PPO

Table 1: Health Insurance Plans

Table 2: Summary Statistics

2006 Plan	Plan B	Plan C	Plan D
2005 (Plan A) Medical Expenditures	$5,161	$2,720	$1,268
	($10,746)	($5,651)	($2,628)
2006 Medical Expenditures	$5,084	$2,960	$1,310
	($12,535)	($7,537)	($3,064)
End-of-Year Marginal Price = 0.1	64.08%	N\A	N\A
End-of-Year Marginal Price = 0.2	N\A	50.08%	26.46%
End-of-Year Marginal Price = 0	18.69%	7.83%	0.95%
Family Size	1.69	1.72	1.55
	(0.46)	(0.45)	(0.50)
Male	0.46	0.52	0.60
	(0.50)	(0.50)	(0.49)
Age	51.82	49.49	40.40
	(9.63)	(11.24)	(13.29)
Age of Employee	51.47	48.27	38.46
	(10.11)	(12.17)	(12.99)
N	5,403	8,238	1,474

Notes: "Age" and "Age of Employee" are the same for individual policyholders but may differ for two-person families.

Table 3: Comparison of Estimators

	IV-QR with covariates	IV-QR without covariates	IV-GQR
Assumption	$U\|Z, X \sim U(0,1)$	$U^*\|Z \sim U(0,1)$	$U^*\|Z, X \sim U^*\|X$
Structural Quantile Function	$d'\tilde{\beta}(\tilde{\tau}) + x'\tilde{\delta}(\tilde{\tau})$	$d'\beta(\tau)$	$d'\beta(\tau)$
Interpretation for τ^{th} quantile	τ^{th} quantile of U	τ^{th} quantile of U^*	τ^{th} quantile of U^*

Table 4: First Stage Estimates

| | Actual Plan Choice | |
Instruments	Plan B	Plan C
Predicted Pr(Plan B) $\times \mathbf{1}(2006)$	0.937***	0.180***
	(0.044)	(0.070)
Predicted Pr(Plan C) $\times \mathbf{1}(2006)$	-0.035	1.187***
	(0.107)	(0.142)
Partial F-Statistic	768.04	68.50

*** Significant at 1 percent level; ** Significant at 5 percent level; * Significant at 10 percent level. Standard errors in parentheses adjusted for clustering at family level. Regressions also include year and cell fixed effects, where cells are based on sex, age, relationship to employee, and family size.

Table 5: Decomposition of Plan Effects

	Plan B	Plan C	Plan D
Per Person Expenditures	$5,127.02	$2,960.70	$1,344.67
	($196.34)	($86.53)	($85.01)
Per Person Expenditures with Random Selection	$3,779.51	$3,070.06	$2,996.89
	($113.51)	($177.37)	($179.27)
Adverse Selection	$1,347.50	-$109.37	-$1,652.22
	($177.41)	($180.61)	($155.69)

Standard errors in parentheses adjusted for clustering at family level. Sub-sampling is used to generate the standard errors. "Adverse Selection" is equal to "Per Person Expenditures" minus "Per Person Expenditures with Random Selection".

Table 6: Comparisons Across Plans

	Plan B relative to Plan D	Plan C relative to Plan D	Plan B relative to Plan C
Per Person Expenditures	$3,782.34	$1,616.02	$2,166.32
	($212.86)	($120.72)	($211.74)
Per Person Expenditures with Random Selection	$782.62	$73.17	$709.45
	($221.60)	($232.94)	($254.60)
Adverse Selection	$2,999.72	$1,542.85	$1,456.87
	($250.30)	($221.19)	($308.97)

Standard errors in parentheses adjusted for clustering at family level. Subsampling is used to generate the standard errors.