NBER WORKING PAPER SERIES

DISENTANGLING MORAL HAZARD AND ADVERSE SELECTION IN PRIVATE HEALTH INSURANCE

David Powell
Dana Goldman

Working Paper 21858
http://www.nber.org/papers/w21858

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2016

This paper was supported by a grant from the Agency for Healthcare Research & Quality (1R03HS023628-01, PI: David Powell) and the National Institute on Aging (P01AG033559). Funding from the Bing Center for Health Economics is also gratefully acknowledged. We received helpful comments from seminar participants at the Annual Health Economics Conference, Annual Health Econometrics Workshop, Conference of the American Society of Health Economists, Midwest Health Economics Conference, RAND, and USC. We are especially grateful to our discussants James Marton, Frank Windmeijer, Mireille Jacobson, and David Frisvold. We also received helpful comments from Abby Alpert, James Burgess, Norma Coe, Peter Huckfeldt, Tim Layton, Chuck Phelps, Julian Reif, and Travis Smith. We are especially grateful to Jean Roth for help with the data and to Dan Feenberg and Mohan Ramanujan for their help with the NBER Unix servers. The content is solely the responsibility of the authors and does not necessarily represent the official views of AHRQ, NIH, or the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w21858.ack

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2016 by David Powell and Dana Goldman. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Disentangling Moral Hazard and Adverse Selection in Private Health Insurance
David Powell and Dana Goldman
NBER Working Paper No. 21858
January 2016
JEL No. I1,I10,I11,I12,I13

ABSTRACT

Moral hazard and adverse selection create inefficiencies in private health insurance markets and understanding the relative importance of each factor is critical for policy. We use claims data from a large firm to isolate moral hazard from plan 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. The nonlinear budget constraints generated by health insurance plans make these assumptions especially poor and we statistically reject their appropriateness. 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 parameterized price. Our empirical strategy exploits the introduction of new plans during the sample period as a shock to plan generosity, and we account for sample attrition over time. We use an instrumental variable quantile estimation technique that provides 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. We estimate that 53% of the additional medical spending observed in the most generous plan in our data relative to the least generous is due to moral hazard. The remainder can be attributed to adverse selection. A policy which resulted in each person enrolling in the least generous plan would cause the annual premium of that plan to rise by $1,000.

David Powell
RAND Corporation
1776 Main St., 4121
Santa Monica, CA 90407
powelld@nber.org

Dana Goldman
Schaeffer Center for Health Policy and Economics
University of Southern California
635 Downey Way
Los Angeles, CA 90089-3333
and NBER
dpgoldma@usc.edu

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. The policy implications are very different, however, depending on the relative magnitudes of each source of distortion, though isolating the independent roles of both moral hazard and adverse selection is rare in the health insurance literature. 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 Patient Protection and Affordable Care Act (ACA) of 2010 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] finds 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.

Adverse selection is another impediment to efficiency in health insurance markets and the ACA's individual mandate was motivated by the efficiency gains of reducing systematic selection into insurance. A large literature documents the difficulties in separating adverse selection and moral hazard (Chiappori and Salanie [2000], Chiappori [2000], Finkelstein and Poterba [2004]). Bajari et al. [2014] estimates a structural model to separate adverse selection and moral hazard in the context of private health insurance by fitting a specified utility function. 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 2005-2007. This employer offered only one insurance plan in 2005. In 2006, they introduced three different PPO plans to replace the 2005 plan. 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. We use the availability of new plans and the differential preferences for plan generosity based solely on observable family characteristics. We condition on individual characteristics to account for the independent effects of the covariates and we further condition on 2005 medical expenditures to control for individual heterogeneity in health care consumption. Identification originates from the introduction of the plan choice in 2006 and the differential but predictable effect that this introduction had on different households. This strategy allows us to estimate the effect of the plan on medical care if selection into the plan were random (i.e., everyone were enrolled in the 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 (i.e., the causal impact of the plan). We estimate these distributions non-parametrically instead of assuming that individuals respond to a specific price in the plan. We map this estimated distribution directly to the non-linear budget constraint created by the health insurance plan and observe whether the medical spending distribution is especially responsive for annual expenditure levels above the deductible or stop loss. This mapping provides 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 study 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 va-

lidity 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 [2015]).

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 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. While this growing literature suggests that individuals are not entirely myopic, it does not imply that it is reasonable to assume that individuals respond *only* to their end-of-year price. Alpert [2015] 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.

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]) typically do not use variation across plans for identification, require strong structural assumptions, and assume perfect foresight. More recently, Cronin [2015] uses structural restrictions to separate myopic behavior from responsiveness to end-of-year prices. Abaluck et al. [2015] uses the nonlinearities in Medicare Part D to study prescription drug utilization and consumer responsiveness to spot and expected marginal prices

While we are generally interested in how people respond to the price of medical care, 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 four important contributions. First, extending a new quantile estimation technique introduced in Powell [2015], we generate 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 90^{th} quantile, the estimates provide the elasticity at the 90^{th} quantile of the distribution for a fixed age. People at the 90^{th} 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 of the treatment variables on the outcome distribution.

Second, our method allows us to be agnostic about how benefit design impacts 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.

Third, we directly account for attrition in our sample. Families opt out of health insurance for a variety of reasons, such as job changes and changes in preferences for health care. In the literature, it is common to select on individuals that the researchers observe for the entire time period. Attrition in the data can often be quite severe and the sources of attrition are potentially not random. In our empirical strategy, we predict both plan choice and attrition. We have a separate selection instrument which statistically predicts attrition, permitting us to condition on a selection adjustment term, comparing people with similar probabilities of remaining in the sample for the entire period. This sample selection adjustment is straightforward to implement and an important methodological contribution

for consistent estimation.

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], Handel [2013]), and we are able to provide plan-specific estimates of the magnitude of adverse selection and its relative importance compared to moral hazard.

In this paper, we estimate price elasticities of medical care which range between -0.9 and 0 throughout the distribution. However, we also can statistically reject that this price elasticity is an appropriate parameterization of the plans. We also estimate plan elasticities and find that the most generous plan encourages additional spending throughout the medical care distribution. The average observed expenditures in the most generous plan are $3,969 more than the per person costs in the least generous plan. We estimate that if selection were random, that the most generous plan would lead to $2,117 in more spending than the least generous plan, implying that 53% of the differential can be attributed to moral hazard. We also estimate adverse selection without restrictive structural assumptions. We find that if everyone in the sample were enrolled in the least generous plan that the premium for that plan would increase by over $1,000.

In the next section, we briefly discuss the importance of estimating the roles of adverse selection and moral hazard. We also consider 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

2.1 Moral Hazard and Adverse Selection

An influential theoretical literature links asymmetric information in insurance markets to inefficient outcomes. Rothschild and Stiglitz [1976] models selection into plans with different risk types. Pauly [1968] discusses the role of moral hazard in health insurance and mechanisms to reduce medical care consumption such as coinsurance. Optimal policy depends on the relative important of adverse selection compared to moral hazard in explaining the correlation between plan generosity and medical care costs. The policy implications for moral hazard are different than those required to confront adverse selection. Adverse selection typically requires risk-pooling, while distortions driven by moral hazard would motivate additional cost-sharing. These issues are discussed in further detail in Cutler and Zeckhauser [2000].

Addressing the distortions induced by either moral hazard or adverse selection often exacerbates the inefficiencies created by the other factor. For example, the ACA's individual mandate encourages healthier individuals to purchase insurance, pooling risk across a more heterogenous population. At the same time, the mandate plus the insurance expansions increase the number of insured individuals, driving down the price of care to consumers. Understanding the magnitude of this tradeoff is a first-order concern for health care policy. The importance of isolating the role of these two factors has been noted in insurance markets more generally and methods to empirically identify each have been introduced (see Abbring et al. [2003]).

We interpret all additional spending causally associated with plan generosity as "moral hazard." Nyman [1999] discusses the value of health insurance in providing access to especially expensive forms of care. The plans that we study are relatively similar and there is likely little scope for differential levels of access to expensive medical care. The difference in the maximum out-of-pocket annual payment between the least and most generous plans in our data is $2,750, which is small relative to the expensive types of care referenced in Nyman [1999].

2.2 Nonlinear Budget Sets

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. As discussed above, it is common to parameterize behavioral responses to benefit design when estimating responsiveness to health insurance. We highlight three reasons why these parameterizations are likely too restrictive.

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 basic setup, it is possible that there is not a unique optimum due to the non-convexity of the budget constraint. 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, nonlinear budgets sets generate variation in spot prices throughout the year. Perfect foresight is a strong assumption, especially in the context of medical care consumption, which is a function of unforeseen health shocks. Purely myopic behavior is also unlikely. It is difficult to model how the different prices interact and how that relationship changes

throughout the year.

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 possible mechanisms. We take a necessary step back relative to the literature to understand the role of benefit design in impacting annual medical care consumption. Furthermore, it is difficult to parameterize plans in such a way as to independently isolate the above factors. For example, using variation in spot prices (or variation in the time exposed to a specific spot price) within the year is typically tied to variation in the location of the kinks, making it difficult to distinguish myopia from the behavior that we would expect given Figure 1. Conceptually, myopia and behavioral responses (with perfect foresight) to non-convex budget constraints cannot be independently isolated without additional assumptions.

Understanding the overall impact of benefit design 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 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.

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 at which point they 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 for the plans that we study. 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. In 2007, the deductibles and stop loss points for Plans B and C changed. 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. Our goal in Figure 2 is to simply illustrate the shapes of the budget constraints for the first $18,000 of annual medical expenditures. While we do not observe premiums in our data, we estimate premium variation across plans using spending differences. This informs the starting point for each budget constraint. There are significant differences in the kink points where the slopes of the respective budget constraints change.

3.2 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 2005-2007 and we restrict our attention to employees for whom we observe insurance choice and spending in 2005. We model medical expenditures for those with a full year of reported expenditures in 2007. Because this is a selected sample, we adjust for attrition. The benefit of observing multiple years is that we can compare the distributions of expenditures in 2005 and 2007 to account for individual-level (and household-level) heterogeneity. We observe all individuals enrolled in a plan even if they do not consume any medical care.

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 so we can control for 2005 medical expenditures as an accurate representation of 2005 demand for medical care (as each person is equally treated by plan generosity).[1]

Summary statistics are presented in Table 2. We present the summary statistics by 2007 plan. We do this for our sample in Panel A (family sizes of 1 or 2) and, for comparison purposes, the full sample in Panel B. 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. Furthermore, the "No Plan" group (the attriters) do not appear to be randomly-selected. Medical expenditures are skewed and the mean potentially masks important distributional differences.

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

3.3 Identification Strategy

We use 2005 medical care to account for individual heterogeneity in underlying health care usage. Our outcome variable is 2007 medical care expenditures. We study 2007 instead of 2006 for the possibility that the change in plan options caused individuals to shift care that

[1]While the lack of choice in 2005 is convenient, the identification strategy would work similarly given multiple plans in 2005 as long as adjusted for the treatment effects of the 2005 plans as well.

they otherwise would have consumed in 2005 for coverage under their 2006 plan (or vice versa).[2] In Section 5.8, we present results using 2006 medical expenditures. The results are similar.

3.3.1 Variation in Plan Generosity

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. Our data provide demographic information such as family size, age, sex, and relationship to employee. 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 255 people.

We use the changes in plan options for identification. Each person in our sample was enrolled in Plan A in 2005. We can assume that many of these families would have preferred 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. To generate the instruments, we estimate

$$P(\text{Plan}_{i,2007} = k) = \Phi(T_i'\Pi_k + \tilde{T}_i'\Lambda_k) \tag{1}$$

where T_i is a set of indicators based on cells for individual i and $\tilde{T}_i$ is a set of indicators based on cells for individuals i's spouse or dependent (these indicators are all equal to 0 for single plan holders). The instruments are the predicted probabilities for $k = \{C, D\}$.[3] The use of a probit model is inessential to the empirical strategy.[4] These instruments are only exogenous conditional on covariates and we condition on covariates (detailed below) throughout our analyses.

The motivation for our empirical strategy is to compare changes in the medical care distribution between 2005 and 2007 and how these distributions differentially changed

[2]Given the benefit design of the plans, both directions are potentially beneficial.

[3]We estimate the effects of Plan C, Plan D, and not enrolling (as described in Section 3.4) relative to Plan B.

[4]Our results are not meaningfully changed if we use an ordered probit model or a set of linear probability models.

for families based on the plans that they were likely to choose. Each family, regardless of preferences for more or less generous health care coverage, is constrained to enroll in the same plan in 2005. In 2006, plans with varying levels of generosity become available and households sorted based on these preference. We use this shock to plan availability for identification.

We condition on 2005 medical expenditures to account for individual-level heterogeneity in medical care consumption propensities. Since each person was in the identical plan in 2005, medical care expenditures reflect variation in health and preference for medical care since the treatment (plan generosity) is the same for each person. In our main analysis, we condition on $\ln M_{i,2005}$, the log of 2005 medical expenditures for person i (set to 0 if $M_{i,2005} = 0$), and a dummy variable equal to 1 for individuals with no 2005 medical expenditures $M_{i,2005} = 0$. In robustness checks, we condition on 2005 medical expenditures more flexibly through the inclusion of indicators for each percentile of 2005 expenditures and find little difference in the final estimates.

Preferences for medical care may change over time so we cannot use actual enrollment in the 2007 plan. Instead, we predict plan enrollment based on family characteristics. Individual characteristics may also predict changes in health care preferences so we condition on individual covariates in our analyses so that identification originates from variation in spousal (or dependent) characteristics (which predict differential preferences for health care coverage associated with those characteristics) and the introduction of new plans. The assumption is that spousal characteristics do not predict *changes* in *own* medical care consumption conditional on the person's own characteristics, except through differential choices in plan generosity. We also condition on the 2005 medical expenditures of the individual's spouse/dependent to account for differences across households. Variation in the instruments conditional on covariates then originates from predictable differences in preferences for health insurance generosity, similar in spirit to the heterogeneity in plan choice discussed in Geruso [2013].

3.4 Attrition

A common concern with the use of claims data is attrition, which may bias estimates if this attrition is systematic. However, selection bias concerns extend throughout the literature on medical care utilization, including influential randomized experiments (see Nyman [2008], Newhouse et al. [2008] for discussion of attrition in the RAND HIE and Finkelstein et al.

[2012] for the Oregon Health Insurance Experiment). The RAND Health Insurance Experiment, which consisted of plans with very generous coverage still had attrition rates as high as 37%. In firm-level claims data, attrition results from employees dropping coverage, leaving the firm, switching to a spouse's plan at another employer, and so on. If we define attrition as individuals enrolled in 2005 but enrolled for less than 365 days in 2007, the attrition rate in the MarketScan data (selecting on firms in the data in both 2005 and 2007) is 58.3%. Our sample has a 58.7% attrition rate.

We select on individuals that are enrolled for 365 days in 2005. Some individuals disenroll in 2005, but this disenrollment should be orthogonal to the introduction of the new plans. The plans begin on January 1 and there is little incentive to switch to a different source of insurance (or no insurance) before that date. Attrition in the last part of 2005 is rare and unlikely to be systematic. Furthermore, we have tested the robustness of our result to treating the first 11 months of 2005 as the full initial year (i.e., selecting on individuals enrolled for the first 11 months, treating individuals that leave the sample in December as attriters, and conditioning only on medical care in the first 11 months[5]) and the results are similar.

Despite the high rates of attrition in the MarketScan data and claims data more generally, research using these data sets often ignore the potential problems caused by selection. The concern in our context is that some individuals may attrit instead of, for example, enrolling in the least generous plan. The observed sample of enrollees in the least generous plan, then, would not be random. Conditioning on 2005 medical expenditures and covariates alleviates this concern. However, there may still be attrition on unobservables.

We account for selection explicitly. We consider anyone in our sample for all of 2005 but enrolled for less than 365 days of 2007 as "attriters." This definition is necessary because we will study the distribution of medical expenditures in 2007, which we only observe for those enrolled for all of 2007. We predict selection in our sample (non-attrition) in the same manner as we predict plan choice. We estimate equation (1) where the outcome variable is an indicator equal to 1 if the individual is enrolled for the full 2007 year. In other words, there are four possible choices for an individual enrolled in 2005: enroll in Plan B, enroll in Plan C, enroll in Plan D, not enroll (which also includes dropping out at some point before the end of 2007). We can predict probabilities for all of these choices (excluding

[5]Conditioning on only the first 11 months of medical care is also a test of robustness to systematic changes in the timing of medical care if individuals differentially utilized medical care in 2005 that they otherwise would have consumed in 2006.

one) using our covariates. Attrition is separately identified using our empirical strategy and the conditions for a valid selection instrument are equivalent here to those needed for our plan choice instruments: the estimated probability of non-attrition must (1) predict non-attrition and (2) not affect medical care expenditures except through selection. We can show that (1) is true empirically. We assume that (2) holds for the same reasons that we assume our instruments are plausibly exogenous conditional on 2005 medical expenditures and covariates.

3.5 Sample Selection

We select our sample on families with two or fewer members and a full year of enrollment in 2005. 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.

We study the medical care expenditures of the employees only. Our analysis sample includes 21,429 people. A high fraction of these 2005 enrollees attrit at some point before the end of 2007. Many are not observed in the data because they left the firm or dropped insurance at some point in 2006 or 2007. We also consider individuals that change plans after January 1, 2007 as attriters since we cannot map their behavior to a single budget constraint. Moreover, some policies gain members due to spouses or dependents joining the policy or due to births. We label employees which originally had policies with two of fewer members but added members by the end of 2007 as attriters since, as described above, these employees are subject to the family deductible and out-of-pocket maximum. Our selection adjustment term is estimated with these attriters and should account for different sources of attrition. When estimating the quantile functions of interest, our sample is the set of employees enrolled in the same 2007 plan for the entire year. We adjusting for the probability of remaining in the sample until the end of 2007 to account for non-random attrition.

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 and, 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 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 quantile treatment effect (QTE) framework introduced in Powell [2015]. 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 4.2 and 4.3. We discuss the IV-GQR estimator more generally first.

4.1 Quantile Estimation

4.1.1 IV-GQR

This paper uses IV-GQR (Powell [2015]), 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 for our context in proceeding sections but discuss the

estimator more generally 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 (untreated by plan generosity) tendency to consume medical care due to health, preferences for medical care, and other factors. 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 medical care 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 (or need) to condition on covariates as well. Adding these covariates in a traditional quantile framework changes the interpretation because the $\tau = 0.9$ estimates now refers 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.

Let $U^* \sim U(0,1)$ be a rank variable which represents proneness to consume medical care (normalized to be distributed uniformly). Powell [2015] models proneness for the outcome variable as an unknown and unspecified function of "observed proneness" (X) and "unobserved proneness" (U): $U^* = f(X,U)$. The specification of interest can be written as

$$Y = D'\beta(U^*), \quad U^* \sim U(0,1). \tag{2}$$

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

$$S_Y(\tau|d) = d'\beta(\tau). \tag{3}$$

The SQF defines the τ^{th} quantile of the outcome distribution given the policy variables if each person in the data were subject to the policy variables $D = d$. It is common and frequently necessary to condition on additional covariates. IV-QR requires those covariates to be included in the structural model, altering the SQF. The parameters are no longer assumed to vary by proneness, only the unobserved component of the disturbance term. A primary motivation of employing quantile techniques is that they allow for a nonseparable

disturbance term. Adding covariates to the quantile function separates this term into different components, undermining the original motivation. Instead of treating the covariates in the same way as the policy variables, IV-GQR lets the covariates provide information about 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, jointly estimating the probability that the outcome is less than the quantile function.

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 (3)) 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 2005 medical expenditures and covariates) distributions, but the parameters refer to the unconditional distribution.[6] Unconditional quantile regression (Firpo et al. [2009]) and distribution regression (Chernozhukov et al. [2013]) were introduced with similar motivations as IV-GQR. Neither was developed in an IV framework or to permit sample selection adjustments. Furthermore, neither estimates equation (3).

The IV-GQR estimator simultaneously uses two moment conditions. We write the quantile function as $D'\beta(\tau)$:

$$E\left\{ Z\left[\mathbf{1}(Y \leq D'\beta(\tau)) - \hat{\tau}_X \right] \right\} = 0, \tag{4}$$

$$E[\mathbf{1}(Y \leq D'\beta(\tau)) - \tau] = 0. \tag{5}$$

where $\hat{\tau}_X$ is an estimate of $P(Y \leq D'\beta(\tau)|X)$, which we define by

$$\hat{\tau}_X = P\left(Y \leq D'\beta(\tau)|X\right) = \Phi\left(X'\alpha\right) \tag{6}$$

In words, IV-GQR uses the covariates 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 (4) is equivalent to IV-QR when $\hat{\tau}_X$ is replaced by τ. Put differently, when there are no covariates, IV-

[6]We use "unconditional" to mean unconditional on the covariates (2005 medical expenditures and indicators based on cells). The resulting distribution depends on the treatment variables (price or plan choice).

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. Condition (5) ensures that the estimates refer to the τ^{th} quantile of the unconditional (on covariates) distribution. The use of the probit model (instead of a semi-parametric estimator) in equation (6) is for computational convenience. Powell [2015] discusses how incorrect distributional assumptions for this equation will not necessarily bias the estimates of the treatment effects since these errors should be orthogonal to the instruments (this is verified in simulations). We use GMM to estimate the parameters of interest.

4.1.2 Extension to Include Sample Selection Adjustment

Because of attrition, we do not observe the full distribution of medical expenditures. Instead, we observe medical care only when $S_i = 1$, where S_i represents selection (i.e., not-attriting). We model selection into the sample as a function of observables and a non-additive selection term (ϵ_i):

$$S_i = F(W_i'\delta, \epsilon_i) \tag{7}$$

where $W_i'\delta \equiv X_i'\phi_1 + Z_i'\phi_2 + \phi_3 p_i$. p_i represents our selection instrument. It is important that we have a variable that affects selection above and beyond the other instruments and covariates, which can be shown by rejecting the null hypothesis $\phi_3 = 0$.

Buchinsky [1998, 2001] discusses sample selection adjustments for quantile regression. However, these papers include an *additive* sample selection term which Huber and Melly [2015] shows assumes a homogenous treatment effect. As before, including an additive control undermines a primary motivation for using a quantile framework. It is straightforward to extend IV-GQR to include the sample selection adjustment as a covariate and allow it to the inform the non-additive disturbance term. This relaxes the IV-GQR assumption to account for the estimated selection term:

$$U_i^*|Z_i, X_i, W_i'\hat{\delta} \sim U_i^*|X_i, W_i'\hat{\delta}$$

The sample selection adjustment term is simply included as an additional covariate in equation (4), accounting for selection while preserving the nonadditive disturbance term. We also modify (5):

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

We define the τ^{th} quantile estimates as the estimates for the τ^{th} quantile of the observed sample ($S = 1$). There is little loss in our context for defining the quantiles in this manner. Alternatively, we could impose an identification-at-infinity assumption and generate estimates for the τ^{th} quantile of the full sample as if we observed the medical expenditure distribution for the full sample.[7] We choose not to do this for several reasons. First, the only difference between the two sets of estimates is which quantile that the estimates refer to (e.g., the 60^{th} quantile of the observed sample may refer to the 50^{th} quantile of the full sample, but the estimated parameters would be the same). Since we will map the estimates to specific medical expenditure dollar values, the values of the quantiles themselves are relatively uninteresting. Second, it is computationally much easier to estimate the quantiles for the observed sample. Third, we are hesitant to impose the identification-at-infinity assumption. Finally, the adverse selection literature (see Geruso [2013], Handel [2013] for two examples) typically defines adverse selection as differences in selection across plans. This metric refers to the observed sample and does not include people unobserved in the data. While understanding systematic selection for the attriters (relative to non-attriters) may be independently interesting, the assumptions required to infer it are restrictive.[8]

4.2 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 (P). We are interested in estimating the Structural Quantile Function:

$$S_{\ln M}(\tau|P) = \phi(\tau) + \delta(\tau)\left[\ln P \times \mathbf{1}\left(P > 0\right)\right] + \gamma(\tau)\mathbf{1}\left(P = 0\right). \tag{8}$$

In this equation, $\delta(\tau)$ represents the price elasticity for the τ^{th} 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.

[7] The relevant condition would be $E\left[P(Y \leq D'\beta(\tau))\,|X, P(S = 1|W)\right] = \tau$.

[8] Since medical expenditures for the attriters are unobserved, it should not be surprising that it would require additional assumptions to infer their medical care distribution.

4.3 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. The SQF is

$$S_{\ln M}(\tau|\text{Plan}) = \phi(\tau) + \sum_{k\in\{B,C,D\}} \beta_k(\tau)\left[\mathbf{1}\left(\text{Plan}_i = k\right)\right]. \tag{9}$$

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. 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. Conditioning on 2005 medical expenditures, which we believe is necessary for our identification strategy, exacerbates the problems associated with conditional quantile estimation. An individual at the top of the distribution conditional on 2005 medical expenditures may be at the lower end of the unconditional distribution. IV-GQR provides the relevant estimates.

4.4 Estimation

We implement the IV-GQR estimator with a non-additive sample selection term to estimate equations (8) and (9). Focusing on the plan elasticity model, our model is

$$\ln M = \alpha(U^*) + \sum_k \beta_k(U^*) \left[\mathbf{1}\left(\text{Plan} = k\right)\right], \quad U^* \sim U(0,1) \tag{10}$$

$$Y^* = \max(\ln M, C), \tag{11}$$

$$Y = Y^* \quad \text{if} \quad S = 1, \tag{12}$$

$$S = F(W'\delta, \epsilon), \tag{13}$$

$$U^*|Z, X, W'\hat{\delta} \sim U^*|X, W'\hat{\delta} \tag{14}$$

$$\mathbf{1}\left(\text{Plan} = k\right) = \psi_k(Z, X, V) \quad \text{for all } k. \tag{15}$$

We make no assumptions on the functional form $\psi_k(\cdot)$ 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 expenditures. 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 $M > 0$).[9] 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 covariates (X) include T_i, $f(M_i^{2005})$, $\tilde{f}(\tilde{M}_i^{2005})$, and $W_i'\hat{\delta}$, where $\tilde{M}_i^{2005}$ refers to the 2005 medical expenditures of individual i's spouse or dependent.

We estimate the index in equation (13) using the monotone rank estimator introduced in Cavanagh and Sherman [1998], which does not require distributional assumptions. We include a 5-piece spline in the estimated index as covariates to flexibly account for attrition, comparing individuals with similar (ex-ante) probabilities of not attriting.

We use subsampling (Politis and Romano [1994]) for inference and implement the entire procedure for each subsample to account for the inclusion of an estimated prediction in the second step.

[9]Censoring is only problematic 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 some observations have high values for that variable, then the quantile function evaluated for that observation's covariates is likely censored), even at high quantiles. IV-GQR only includes the treatment variables - which take a limited set of values in our context - in the quantile function and 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.

4.5 Reported Parameters

We will present our results with graphs that show the parameters over the entire distribution. When applicable, our graphs will include the points where the distribution has passed the plan deductible and 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 estimates above the deductible refer to individuals that passed the deductible by the end of the year.

4.5.1 Price Elasticities

For the price elasticity estimates, we report the estimates for $\delta(\tau)$ and $\gamma(\tau)$. $\delta(\tau)$ is the price elasticity of medical care for quantile τ. $\gamma(\tau)$ is the causal impact of a marginal price of zero.

4.5.2 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 4.5.1) 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

$$\tilde{\beta}_k(\tau) = \begin{cases} \widehat{\phi(\tau)} \quad \text{if} \quad \exp\left[\widehat{\phi(\tau)}\right] < \text{Plan } k\text{'s Deductible} \\ \widehat{\phi(\tau)} + \widehat{\delta(\tau)}\left[\ln(\text{Plan } k\text{'s Coinsurance Rate})\right] \quad \text{if} \quad \exp\left[\widehat{\phi(\tau)}\right] \geq \text{Plan } k\text{'s Deductible} \\ \qquad \text{and} \quad \exp\left[\widehat{\phi(\tau)} + \widehat{\delta(\tau)}\left[\ln(\text{Plan } k\text{'s Coinsurance Rate})\right]\right] < \text{Plan } k\text{'s Stop loss} \\ \widehat{\phi(\tau)} + \widehat{\gamma(\tau)} \\ \qquad \text{if} \quad \exp\left[\widehat{\phi(\tau)} + \widehat{\delta(\tau)}\left[\ln(\text{Plan } k\text{'s Coinsurance Rate})\right]\right] \geq \text{Plan } k\text{'s Stop loss} \end{cases} \tag{16}$$

This is 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.

4.5.3 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 (9)). These estimates refer to the medical expenditures if the entire sample were exogenously enrolled in the plan, shutting down adverse selection. Consequently, we can compare the expenditure distribution of those actually enrolled in the plan. If the fraction of enrollees in the plan that have medical expenditures below $\widehat{\phi(\tau)} + \widehat{\beta_k(\tau)}$ 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

$$\widehat{\psi_k(\tau)} = \frac{1}{N_k} \sum_{i \in \mathcal{K}} \mathbf{1}(Y_i \leq \widehat{\phi(\tau)} + \widehat{\beta_k(\tau)}). \tag{17}$$

This equation represents the sample equivalent of the probability that an enrollee in plan k is below the τ^{th} 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 and Selection Equation

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 2007. In 2005, all families were constrained to choose Plan A. Identification originates from the availability of Plans B, C, and D in 2007 and the differential preferences (based on spousal observable characteristics) for these plans. We predict these probabilities using the covariates.

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 C in 2007 and the probability of choosing Plan D in 2007. Plan B is the excluded category. We include fixed effects for each demographic cell as well as controls for 2005 medical expenditures. 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.

In Table 5, we report estimates for the selection equation. We present estimates using a probit estimator and a monotone rank estimator. We use the estimates from the latter to construct the selection adjustment term. We find that the selection instrument (the predicted probability of not attriting) is positively associated with remaining in the sample for all of 2007.

5.2 Price Elasticity Estimates

In this section, we provide estimates of equation (8). We present the results graphically. The price elasticity term $\delta(\tau)$ is presented with confidence intervals in Figure 3a. We simultaneously estimate the effect of free marginal medical care. These results are presented in Figure 3b. 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 5.3.[10]

The elasticity estimates are relatively constant throughout the distribution. We estimates an elasticity between 0 and -0.9 for most of the sample. In general, the elasticities are similar to those found and reported by the RAND Health Insurance Experiment. An elasticity of -0.4 implies that a coinsurance rate of 0.2 would increase medical care consumption by 90%. The estimates in Figure 3b suggest that medical expenditures are very responsive to a marginal price of 0. The estimates are between 2 and 4 up to quantile 70 but decreases at upper ends of the distribution. An estimate of 3 implies that a marginal price of 0 increases medical care by 1,909%, relative to facing the full price of care. The results in Figures 3a and 3b suggest that individuals in the top quarter of annual medical expenditures behave as if they are inelastic to the price of medical care.

5.3 Plan Elasticity Estimates

This section presents our main results. We estimate the SQF in equation (9) and then present the differences in the SQFs to show how the plans generate different distributions of medical care. In Figure 4, 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, mapping that distribution to the kinks in the budget sets 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.

We estimate little difference at the very bottom of the distribution, but the Plan B distribution becomes very responsive at quantile 15, peaking with an elasticity of 1.5. This high responsiveness occurs even before either distribution is subject to the deductible. The elasticity decreases until it reaches 0.56 (at about $540 of annual medical expenditures for Plan B) before increasing rapidly prior to Plan B's stop loss point. The out-of-pocket

[10]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.

maximum for Plan B has little effect on the distribution. The Plan D distribution does rise relative to the Plan B distribution after the Part D deductible is reached, suggesting that individuals are reacting to the lower price of care. The elasticity of Plan B relative to Plan D steadily decreases to about 0.25 at $3,248 of medical expenditures for Plan A ($2,519 for Plan D). The Plan D maximum has little effect on the distribution. We estimate large elasticities even at the top of the medical care distribution. Overall, Figure 4 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. There are several possible reasons for this such as differences in the prices of episodes of care for expensive treatments and reductions in spot prices early in the year.

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 5. The resulting distributions are highly unrealistic and look very different from the less parametric results found in Figure 4. 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 6. Here, we observe similar patterns as before. In general, the elasticities are smaller, consistent with the fact that Plan C is more generous than Plan D. The estimated coefficients after the Part B deductible are between 0.2 and 0.4, implying a 22%-49% increase in spending due to enrollment in Plan B relative to Plan C. The relative distributions appear unaffected by the Plan C deductible or the Plan B maximum. The estimates for the two plans do converge, but we still estimate large elasticities near the top of the distribution.

The counterfactual distributions again illustrate that assuming that individuals respond only to the end-of-year price leads to very different conclusions. Figure 7 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 A.1 and A.2 present 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 (9)) 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) test and simulate the distribution of this test statistic using subsampling. The CMS test for Plan B rejects the equality of the two distributions at the 5% level. 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 $P(Y_i \leq \widehat{\phi(\tau)} + \widehat{\beta_k(\tau)}) = \tau$. 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 to the estimated distribution for information about the magnitude of adverse selection.

We estimate our metrics and present them in Figure 8. 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 20^{th} quantile of the SQF for Plan D, which is equal to \$182.32. Instead, we observe that almost 28% of the enrollees have smaller expenditures than \$182.32. 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 30^{th} quantile of the medical care distribution for Plan B would be \$380.65. Only 20.6% 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 observe adverse selection close to the bottom of the medical care distribution but favorable selection

for most of the distribution.

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 B and C, implying that there is systematic selection. We can reject non-systematic selection at the 10% level for Plan D.

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 of equation (9), we 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:

$$\hat{E}\left[\text{Per-Person Medical Expenditures in Plan B with Random Selection}\right] = \int_{\tau} \left[\widehat{\phi(\tau)} + \widehat{\beta_B(\tau)}\right] d\tau \tag{18}$$

We label these "Per Person Expenditures with Random Selection" in Table 6 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.[11]

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

[11] We should also highlight that the standard errors in Table 6 represent the standard errors for the mean estimates and are not comparable to the standard deviations found in Table 2.

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 $1,000.

Table 7 repeats the results in Table 6 but provides complementary metrics by using comparisons between plans. We estimate that enrollment in Plan B increases per-person medical expenditures by over $2,000 relative to Plan D and over $1,000 relative to Plan C. 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 47% 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 2007 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.[12] Our results suggest that previous 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 $3,704. However, in Table 6, we see that the difference in selection in 2007 expenditures is only $1,853. Similarly, the difference in 2005 medical expenditures between Plan B and Plan C enrollees is $2,636. But, in 2007 expenditures, selection accounts for only $1,531. These differences are economically meaningful and highlight the benefits of estimating adverse selection in the same year that the selection is occurring.

[12] All three 2007 plans provide full coverage above the stop loss point, but individuals may still value the financial risk protection at lower levels of annual expenditures.

5.8 Robustness Checks

Table 8 includes estimates from robustness tests. We present only the moral hazard estimates: the effect of Plan B relative to Plan D (top row) and the effect of Plan B relative to Plan C (bottom row). The corresponding estimates from Table 7 are included in the first column. In our first test, we replicate our previous analysis using the 2006 data. We find similar estimates. The consistency of the estimates suggests that are identification strategy is robust to concerns of intertemporal substitution of medical care between 2005 and 2006. We condition on 2005 medical expenditures to account for individual heterogeneity, though it is possible that some care was shifted to 2005 (conversely, it is possible that care was shifted to 2006) in anticipation of the introduction of new plans. This would imply that our medical expenditure controls are partially treated. This shifting would affect care in 2006 as well. Since the estimates are similar whether 2006 or 2007 is used as the "treated" year, we conclude that any shifting of care was not systematic.

We believe that it is important for our empirical strategy that we account for individual-level heterogeneity in medical care consumption. We control for 2005 medical care to account for initial differences in care and predict changes in care due to considerations such as mean reversion. In the last column of Table 8, we present estimates in which we condition more flexibly on 2005 medical expenditures. We include indicator variables for each percentile of individual medical expenditures in 2005, permitting especially flexible predictions of the distribution of care in 2007. Our estimates are robust to the use of these more flexible controls. The moral hazard estimate for Plan B relative to Plan D increases while the estimate for Plan B relative to Plan C is very similar. If the more parameterized controls used in the main analysis were not adequate, we would expect to over-estimate the role of moral hazard since we would attribute differences in underlying heterogeneity to the causal impact of the more generous plans. This does not appear to be the case.

6 Conclusion

Understanding moral hazard and adverse selection in private health insurance is widely-recognized as critical to policy. 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 of different health insurance plans on the entire distribution of medical care consumption 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 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 $773 of additional per-person costs in the most generous plan, implying that an individual considering this plan would pay over $60 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 $1,000.

We estimate that moral hazard is responsible for 53% of the differences in expenditures between the most and least generous plans. Adverse selection also plays an important role, accounting for the other 47%. In the absence of moral hazard, the difference in average medical expenditures across these plans would be $2,117 instead of $3,969. Finally, we find that using the previous year's medical expenditures as a metric of selection greatly overstates the magnitude of selection.

References

Jason Abaluck, Jonathan Gruber, and Ashley Swanson. Prescription drug use under Medicare Part D: A linear model of nonlinear budget sets. Technical report, National Bureau of Economic Research, 2015.

Jaap H Abbring, Pierre-André Chiappori, James J Heckman, and Jean Pinquet. Adverse selection and moral hazard in insurance: Can dynamic data help to distinguish? *Journal of the European Economic Association*, pages 512–521, 2003.

Abby Alpert. The anticipatory effects of Medicare Part D on drug utilization. 2015.

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.

Patrick Bajari, Christina Dalton, Han Hong, and Ahmed Khwaja. Moral hazard, adverse selection, and health expenditures: A semiparametric analysis. *The RAND Journal of Economics*, 45(4):747–763, 2014.

Moshe Buchinsky. The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *Journal of Applied Econometrics*, 13(1):1–30, 1998.

Moshe Buchinsky. Quantile regression with sample selection: Estimating women's return to education in the US. *Empirical Economics*, 26(1):87–113, 2001.

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.

Christopher Cavanagh and Robert P Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84(2):351–381, 1998.

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.

Christopher J Cronin. Insurance-induced moral hazard: A dynamic model of within-year medical care decision making under uncertainty. 2015.

David M Cutler and Richard J Zeckhauser. The anatomy of health insurance. *Handbook of Health Economics*, 1:563–643, 2000.

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.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, and Katherine Baicker. The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106, 2012.

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

Martin Huber and Blaise Melly. A test of the conditional independence assumption in sample selection models. *Journal of Applied Econometrics*, 2015.

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 Kowalski. Censored quantile instrumental variable estimates of the price elasticity of expenditure on medical care. *Journal of Business & Economic Statistics*, 2015.

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.

Joseph P Newhouse, Robert H Brook, Naihua Duan, Emmett B Keeler, Arleen Leibowitz, Willard G Manning, M Susan Marquis, Carl N Morris, Charles E Phelps, and John E Rolph. Attrition in the RAND health insurance experiment: a response to nyman. *Journal of Health Politics, Policy and Law*, 33(2):295–308, 2008.

John A Nyman. The value of health insurance: the access motive. *Journal of Health Economics*, 18(2):141–152, 1999.

John A Nyman. Health plan switching and attrition bias in the rand health insurance experiment. *Journal of Health Politics, Policy and Law*, 33(2):309–317, 2008.

Mark V Pauly. The economics of moral hazard: comment. *The American Economic Review*, pages 531–537, 1968.

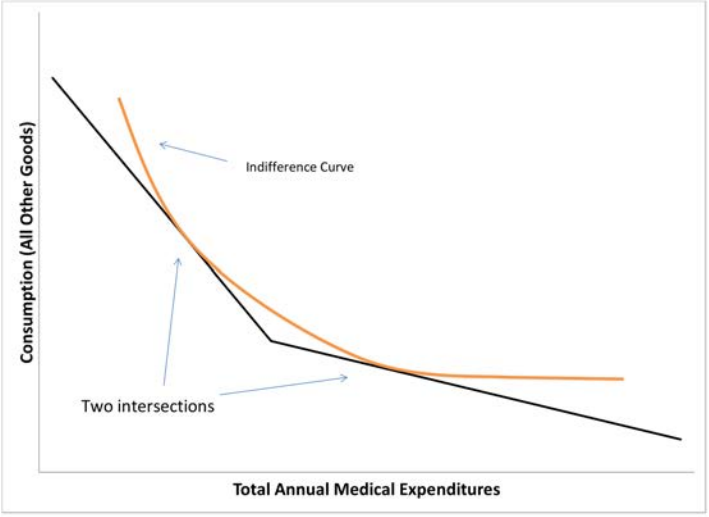
Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):2031–2050, 1994.

David Powell. Estimation of quantile treatment effects in the presence of covariates. Working papers, RAND Corporation Publications Department, 2015.

Michael Rothschild and Joseph Stiglitz. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The Quarterly Journal of Economics*, 90 (4):629–649, 1976.

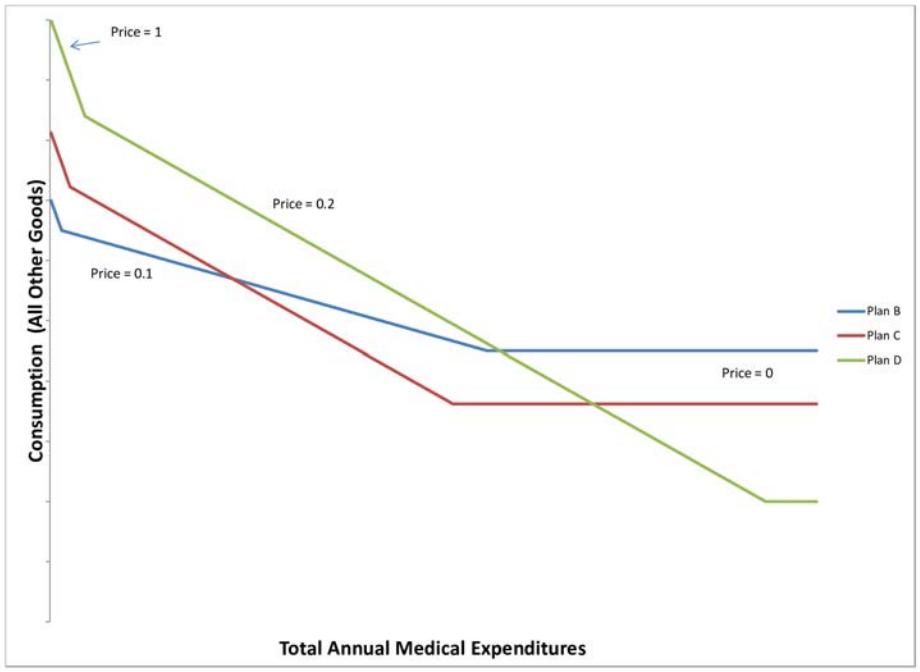
Figures

Figure 1: Indifference Curve and Non-Linear Budget Constraint



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

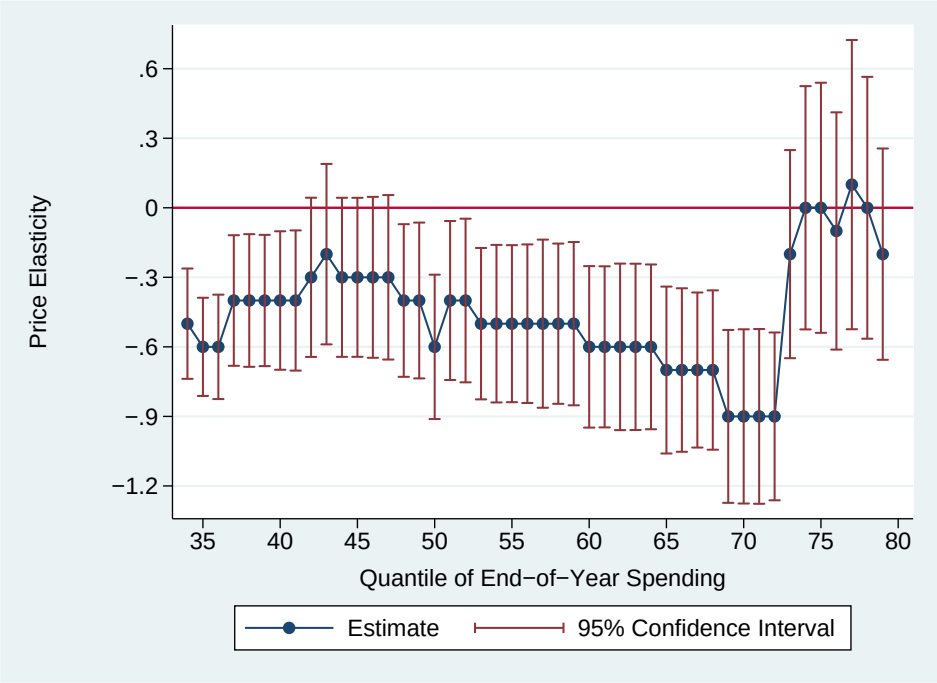
Figure 2: Budget Constraints Generated by Plans in Data



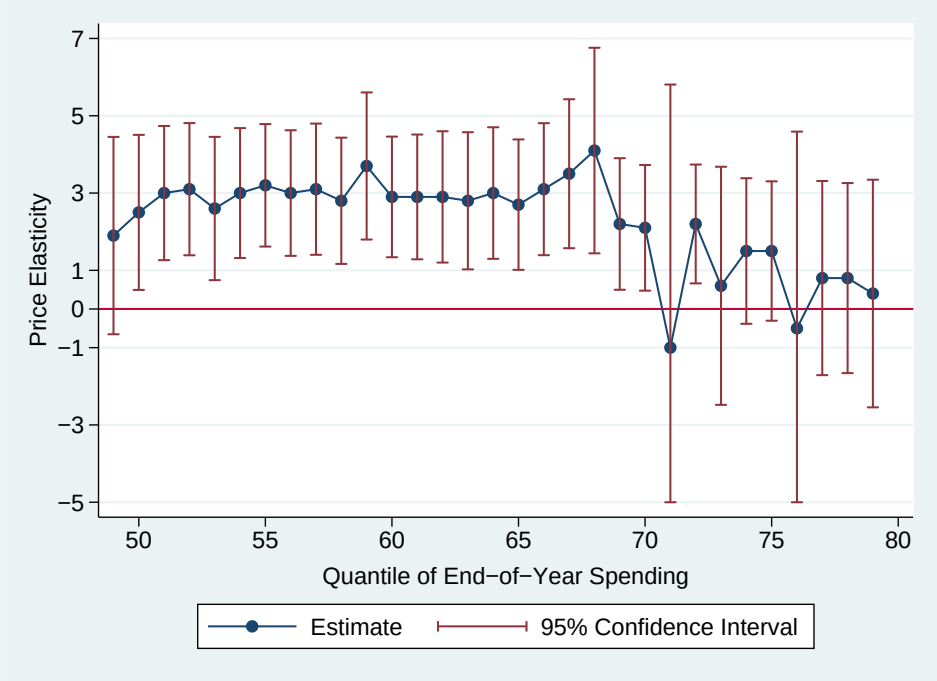
Notes: This figure graphs the nonlinear budget set generated by each 2007 plan. We do not observe premiums, but we estimate premium differences using spending differences across plans. The variation in the starting points of the 3 budget constraints is due to estimated premium variation only.

Figure 3: Parameterized Estimates

(a) Price Elasticity Estimates

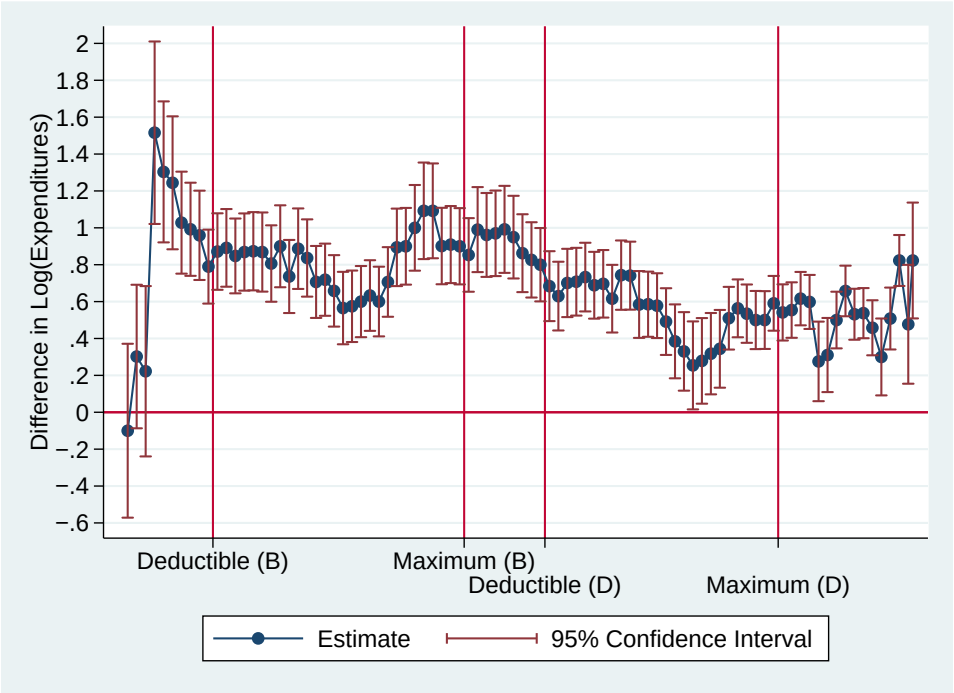


(b) Effect of Price=0 on Medical Expenditures



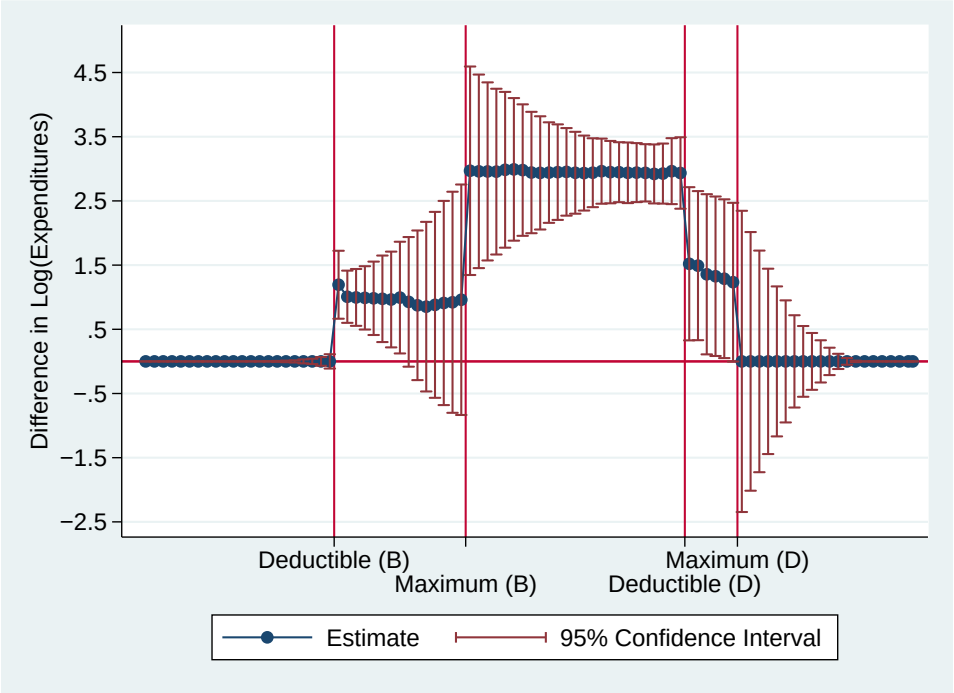
Notes: Using an instrumental variable quantile regression estimator, we estimate the price elasticity and the effects of a marginal price of 0. Confidence intervals generated using subsampling.

Figure 4: Difference in Expenditure Distribution: Plan B vs. Plan D



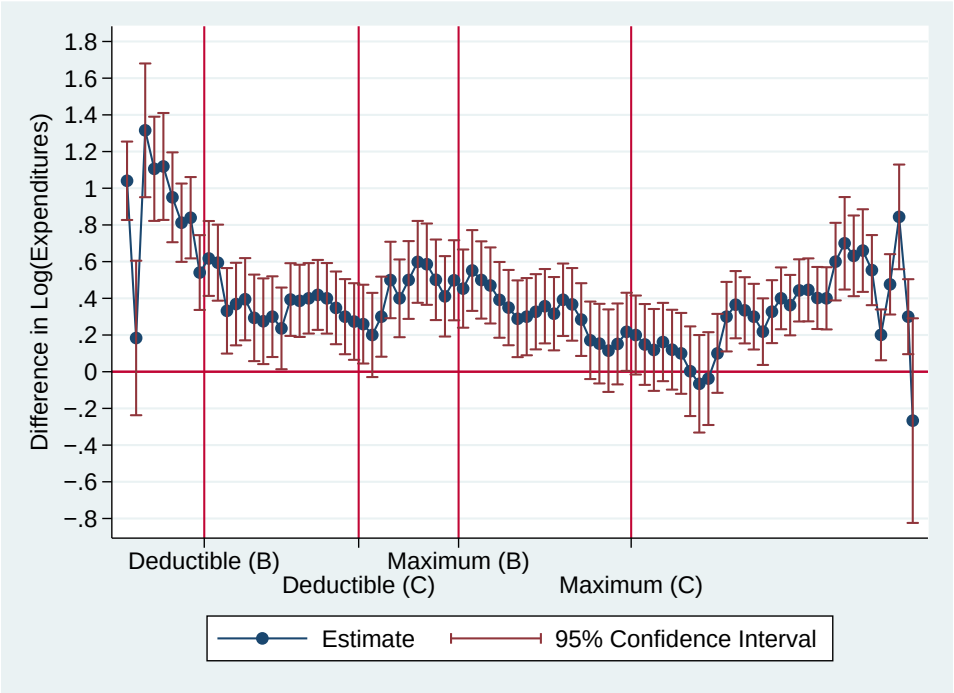
Notes: Using an instrumental variable quantile regression estimator, we estimate the distribution of Plan B and Plan D if enrollment into each plan were random. We graph the difference in these distributions here. Confidence intervals generated using subsampling.

Figure 5: Parameterized Difference in Expenditure Distribution: Plan B vs. Plan D



Notes: We use the price elasticity estimates shown in Figures 3a and 3b 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 subsampling.

Figure 6: Difference in Expenditure Distribution: Plan B vs. Plan C

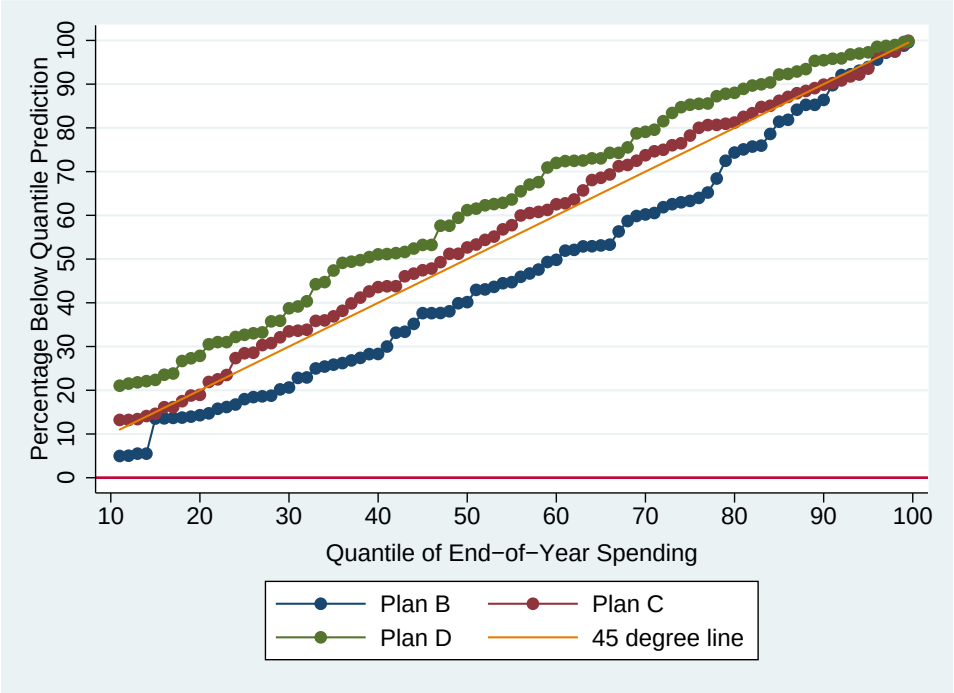


Notes: Using an instrumental variable quantile regression estimator, we estimate the distribution of Plan B and Plan C if enrollment into each plan were random. We graph the difference in these distributions here. Confidence intervals generated using subsampling.

Figure 7: Parameterized Difference in Expenditure Distribution: Plan B vs. Plan C

Notes: We use the price elasticity estimates shown in Figures 3a and 3b 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 subsampling.

Figure 8: Adverse Selection



Notes: We use the plan elasticities presented in Figures 4, 6, A.1 to estimate the empirical probability that an enrollee in the plan is below the estimate quantile function for that plan. Confidence intervals generated using subsampling. The 45 degree line represents a plan with no systematic selection.

Tables

Table 1: Health Insurance Plans

	Plan (Year)	Plan A (2005)	Plan B (2006)	Plan C (2006)	Plan D (2006)	Plan B (2007)	Plan C (2007)	Plan D (2007)
Deductible	Individual	\$0	\$200	\$400	\$800	\$250	\$450	\$800
	Family	\$0	\$400	\$800	\$1,600	\$500	\$900	\$1,600
Stop loss	Individual	None	\$1,000	\$2,000	\$4,000	\$1,250	\$2,250	\$4,000
	Family	None	\$2,000	\$4,000	\$8,000	\$2,500	\$4,500	\$8,000
Coinsurance		20%	10%	20%	20%	10%	20%	20%
Plan Type		PPO	PPO	PPO	PPO	PPO	PPO	PPO

Table 2: Summary Statistics

A. Family Size $\leq$ 2				
	Plan B	Plan C	Plan D	No Plan
2005 Medical Expenditures	\$5,157.89	\$2,521.67	\$1,453.58	\$3,821.29
	(\$12562.64)	(\$5462.58)	(\$3059.12)	(\$12226.04)
2007 Medical Expenditures	\$5,441.79	\$2,883.20	\$1,559.35	–
	(\$13291.73)	(\$7827.91)	(\$5021.00)	–
Family Size	1.51	1.57	1.41	1.66
	(0.50)	(0.50)	(0.49)	(0.47)
Age	52.23	48.69	39.88	52.88
	(8.98)	(11.33)	(13.05)	(9.84)
End-of-Year Marginal Price = 0.1	67.52%	0.00%	0.00%	–
End-of-Year Marginal Price = 0.2	0.00%	45.72%	24.32%	–
End-of-Year Marginal Price = 0	9.54%	4.14%	0.63%	–
N	2,777	4,639	1,435	12,578
B. Full Sample				
	Plan B	Plan C	Plan D	No Plan
2005 Medical Expenditures	\$5,004.00	\$2,458.86	\$1,520.46	\$3,447.75
	(\$13336.92)	(\$7255.11)	(\$3579.90)	(\$11001.93)
2007 Medical Expenditures	\$5,130.47	\$2,788.15	\$1,466.76	–
	(\$12932.82)	(\$9323.08)	(\$4320.47)	–
Family Size	1.91	2.30	2.68	2.27
	(1.07)	(1.27)	(1.54)	(1.16)
Age	50.85	46.46	39.96	50.44
	(9.36)	(10.86)	(10.55)	(10.38)
End-of-Year Marginal Price = 0.1	68.26%	0.00%	0.00%	–
End-of-Year Marginal Price = 0.2	0.00%	49.29%	30.65%	–
End-of-Year Marginal Price = 0	9.12%	4.15%	0.56%	–
N	3,346	6,729	2,679	17,760

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^*

Differences in the SQF imply differences in the quantile treatment effects.
$U^* = f(X, U)$.

Table 4: First Stage Estimates

Instruments	Actual Plan Choice	
	Plan B	Plan C
Predicted Pr(Plan B)	0.938***	-0.399***
	(0.092)	(0.079)
Predicted Pr(Plan C)	-0.804***	1.301***
	(0.137)	(0.119)
Partial F-Statistic	45.65	53.08

*** Significant at 1 percent level; ** Significant at 5 percent level; * Significant at 10 percent level. Regressions also include cell fixed effects, where cells are based on sex, age, relationship to employee, and family size. They also include logged 2005 medical expenditures and a dummy variable equal to 1 if 2005 medical expenditures equal to 0. They also condition on 2005 medical expenditures of the spouse or dependent.

Table 5: Selection Equation

	Pr(Not Attrit)	Pr(Not Attrit)
Predicted Pr(Not Attrit)	0.954***	0.677***
	(0.037)	(0.028)
Predicted Pr(Plan C)	-0.001	0.314***
	(0.054)	(0.046)
Predicted Pr(Plan D)	-0.026	0.225***
	(0.062)	(0.056)
Estimator	Probit	Monotone Rank

*** Significant at 1 percent level; ** Significant at 5 percent level; * Significant at 10 percent level. Standard errors estimated using subsampling. Regressions also include cell fixed effects, where cells are based on sex, age, relationship to employee, and family size. They also include logged 2005 medical expenditures and a dummy variable equal to 1 if 2005 medical expenditures equal to 0. They also condition on 2005 medical expenditures of the spouse or dependent. Coefficients are scaled so that the norm is equal to 1.

Table 6: Decomposition of Plan Effects

	Plan B	Plan C	Plan D
Per Person Expenditures	\$5,507.19	\$2,928.35	\$1,537.83
	(\$292.41)	(\$111.67)	(\$149.89)
Per Person Expenditures with Random Selection	\$4,733.64	\$3,686.22	\$2,616.87
	(\$303.70)	(\$191.36)	(\$232.89)
Adverse Selection	\$773.55	-\$757.87	-\$1,079.04
	(\$338.74)	(\$172.17)	(\$221.49)

Subsampling is used to generate the standard errors. "Adverse Selection" is equal to "Per Person Expenditures" minus "Per Person Expenditures with Random Selection".

Table 7: Comparisons Across Plans

	Plan B relative to Plan D	Plan C relative to Plan D	Plan B relative to Plan C
Per Person Expenditures	\$3,969.36	\$1,390.52	\$2,578.84
	(\$328.56)	(\$191.66)	(\$308.22)
Per Person Expenditures with Random Selection	\$2,116.77	\$1,069.36	\$1,047.42
	(\$304.38)	(\$344.77)	(\$394.93)
Adverse Selection	\$1,852.59	\$321.17	\$1,531.42
	(\$382.86)	(\$331.89)	(\$461.68)

Subsampling is used to generate the standard errors.

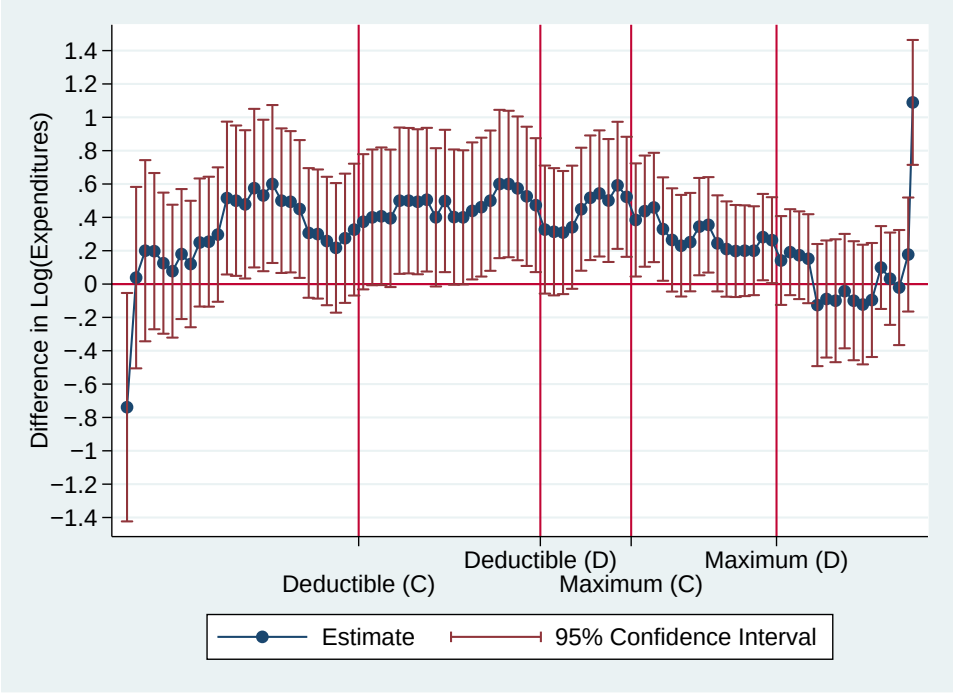
Table 8: Robustness Checks: Moral Hazard Estimates

	Main Results	2006 Data	Less Parametric Controls
Plan B relative to Plan D	\$2,116.77	\$2,217.64	\$2,769.91
Plan B relative to Plan C	\$1,069.36	\$1,051.73	985.88

These estimates correspond to columns 1 and 3 of Table 7. We include the same estimates from Table 7. In the second column, we use the 2006 data and reproduce our analyses. In the final column, we include indicators for each percentile of 2005 medical expenditures as covariates.

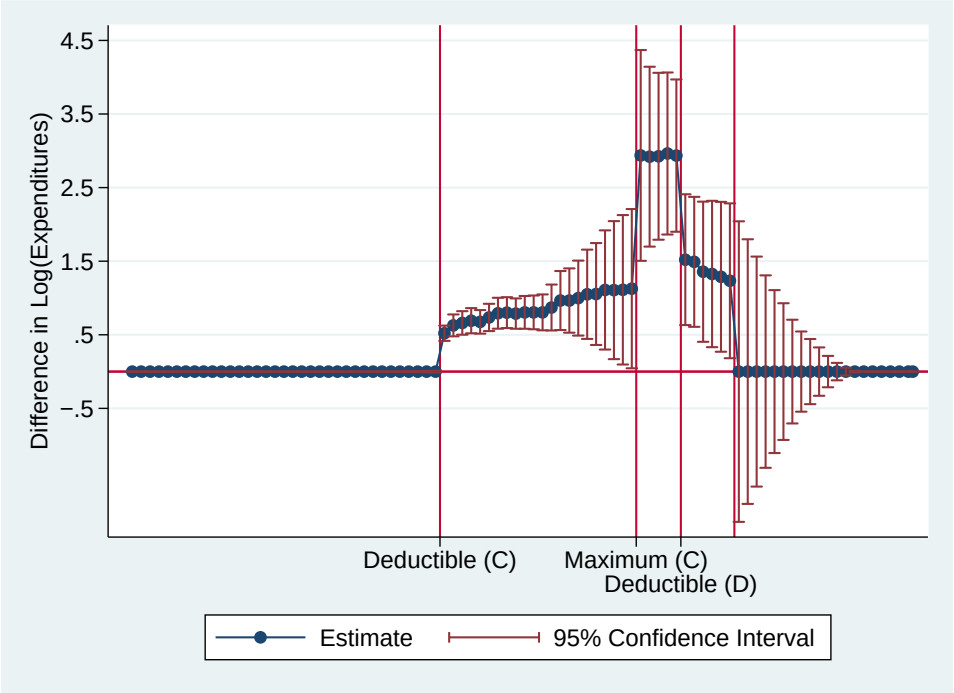
A Appendix

Figure A.1: 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 A.2: Parameterized Difference in Expenditure Distribution: Plan C vs. Plan D



Notes: We use the price elasticity estimates shown in Figures 3a and 3b 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.