International Journal of Industrial Organization 37 (2014) 178–191

Contents lists available at ScienceDirect

International Journal of Industrial Organization

journal homepage: www.elsevier.com/locate/ijio

Estimating demand elasticities using nonlinear pricing☆

Christina M. Dalton *

Department of Economics, 204B Kirby Hall, Wake Forest University, Winston-Salem, NC 27109, United States

ARTICLE INFO

Article history:
Received 14 January 2013
Received in revised form 21 June 2014
Accepted 25 August 2014
Available online 16 September 2014

JEL classification:
D40
I11
C14

Keywords:
Elasticity
Nonlinear
Health insurance
Moral hazard

ABSTRACT

Nonlinear pricing is prevalent in industries such as health care, public utilities, and telecommunications. However, this pricing scheme introduces bias into estimating elasticities for welfare analysis or policy changes. I develop a local elasticity estimation method that uses nonlinear price schedules to isolate consumers' expenditure choices from selection and simultaneity biases. This method improves over previous approaches by using commonly-available observational data and requiring only a single general monotonicity assumption. Using claims-level data on health insurance with two nonlinearities, I am able to measure two separate elasticities, and find that elasticity declines from −0.26 to −0.09 by the second nonlinearity. These estimates are then used to calculate moral hazard deadweight loss. This method enables estimation of many policies with nonlinear pricing which previous tools could not address.

1. Introduction

Demand elasticities are important to policy makers for designing cost-sharing and calculating welfare in sectors such as health insurance, public utilities, and telecommunications. However, pricing is commonly nonlinear in these sectors, for example in deductibles in health insurance, tiered pricing in public utilities, and contracts with usage allowances in telecommunications.[1] Nonlinear pricing contributes to efficient plan design, but complicates estimation of elasticities for several reasons. First, the price a consumer faces is a function of quantity; consumers must pass a certain level of spending to reach a new price level. Second, selection bias occurs when an unobservable factor, such as health status or preferences for high versus low data use, pushes a consumer above or below the nonlinearity. Using observable variables such as age to proxy may not reduce bias, since unobservable health status is likely correlated with age. It is difficult to get rid of this selection bias without experimental data or an exogenous shock, both of which are empirically rare.

In this paper, I present a method to calculate elasticity in the presence of nonlinear pricing in consumer contracts. This method uses the nonlinearity itself to control for bias by taking advantage of the discontinuous change in price across the nonlinearity, while controlling for the underlying distribution of individual unobserved characteristics. This method has very general data requirements and uses only one minimally restrictive assumption: that the expenditure of interest must be increasing in the unobserved preference characteristic. I then apply this method to a private health insurance claims-level dataset with two nonlinearities. In addition to providing an updated health expenditure demand elasticity, my results also are novel because I am able to estimate elasticities at different points on the same demand curve. Identification uses the nonparametric estimation framework of Matzkin (2003). This method uses the same key insight as Bajari et al. (2010), but here I focus on individual consumer contracts rather than provider contracts. Consumers have less precise control over health expenditures given health status than Bajari et al. (2010) find using provider charges by hospitals over a diversity of expense categories. Besides the novel setting of consumer contracts, applying this method to health insurance contracts estimates health expenditure elasticities which are used to design contracts and make welfare predictions of insurance expansions. This paper is able to measure elasticities in two separate regions, which is informative since demand for health care likely changes along its typically skewed spending distribution.

The goal of the method is to generate local elasticities within a contract with nonlinear pricing. The method is aimed at policy applications such as understanding consumer behavior in certain regions, or how

☆ Particular thanks to Patrick Bajari and Robert Town. I also thank W. David Bradford, Meghan Busse, Meg Henkel, Thomas Holmes, Ginger Jin, Willard Manning, Amil Petrin, Connan Snider, Kevin Wiseman, and Thomas Youle for helpful comments.

* Tel.: +1 336 758 4495.
E-mail address: marshcl@wfu.edu.

[1] See for example, Reiss and White (2002), Herriges and King (1994), Maddock et al. (1992) in electricity, Szabo (2010) and Diakite et al. (2009) in water and development, and Grubb and Osborne (2012), Reiss and White (2006), Seim and Viard (2011), and Huang (2008) in cell phone markets.

http://dx.doi.org/10.1016/j.ijindorg.2014.08.007
0167-7187/

changing pricing schedules might affect the distribution of spending, given a particular consumer contract design.

The "gold standard" of elasticity estimation is experimental data. The best example in the health industry is the RAND Health Insurance Experiment (HIE), which began in 1971 and was conducted over 15 years (Manning et al., 1987; Newhouse, 1993). The RAND HIE avoided selection bias by randomizing patients into health plans' pricing schedules. While excellent for reducing selection bias, experimental data is extremely costly in terms of both time and money and is difficult to replicate. In addition, the results from the HIE best apply to the same population type and insurance framework of the HIE. This estimation method can be used on more specific populations of interest to policy makers or on new insurance structures. Since the RAND HIE, exogenous shocks or natural experiments have been used to control for simultaneity and selection bias. Cherkin et al. (1989) use the introduction of office visit copayments for government employees to create a quasi-experimental price change with which to measure elasticity. Selby et al. (1996) use a similar technique taking advantage of a copayment introduction for emergency room visits in a large HMO. In measuring price response more generally, Doyle and Almond (2011) find a substantial increase in mother's length of stay due to better insurance coverage around a policy treatment for children born just before and just after midnight. These natural experiments are difficult for policy makers to use regularly, however, because they rely on unique exogenous changes.

Eichner (1998) and Kowalski (2010) create a natural experiment in the presence of a deductible when an unexpected injury exogenously pushes other non-injured family members into a different pricing zone. Using a two-period utility model, Duarte (2012) also uses an unforeseen accident instrument on Chilean data to reveal how elasticities vary by income and demographics. However, unexpected injury in a deductible structure is hard to replicate in pharmaceutical or public utilities data, for example. The method presented here is accessible to policy makers outside of the health plan family deductible, a useful tool given the prevalence of nonlinear pricing in many other sectors.

Previous methods also estimate one elasticity over the whole range of expenditures. In health expenditures especially, distributions are commonly skewed, with a large proportion of consumers spending small amounts and a long tail of high spending consumers. Tiered pricing structures are often created precisely because different groups of consumers exist. Telecommunications users who end up near usage allowance limits are using bandwidth differently than low bandwidth users, i.e. using email versus video streaming. Estimating one elasticity over an entire range may mask heterogeneity of elasticity values along the distribution. An advantage of this method to policy makers is that it provides a local estimate of elasticity around current pricing points—those very areas that policy makers and insurance administrators may be modifying.

The main intuition of this estimation uses the kink in the price schedule at the nonlinearity. Selection bias exists because agents on either side of the nonlinearity face different prices, but are also different on an unobservable dimension such as health status or preferences for bandwidth use. In this paper's setting of a deductible, patients who surpass the deductible face a lower price for care, but also likely had more health shocks. However, the marginal price of an additional unit of care remains constant on each side, but changes suddenly at the nonlinearity. Identification is off the fact that marginal price is constant within the estimation regions, but the distribution of health status changes along the estimation window. Using the differences in the density of final spending before and after the nonlinearity allows us to isolate the change in spending due only to prices.

Identification is based on Matzkin (2003). The only condition that must hold is that final expenditure is strictly increasing in the individual unobserved characteristics that induce expenditure. For example, if an individual has a higher preference for bandwidth use, his final expenditure on bandwidth usage will be higher than an individual with a lower preference for use. For health insurance, this unobserved characteristic measure will be able to capture a more general ranking of health than diagnosis codes or self-reported health status. The unobserved characteristics are essentially a latent error term. Given this condition and using Matzkin (2003) I am able to proxy the distribution of unobservable characteristics using the percentiles of final expenditures.

Given both final expenditures and the estimated relative values of the unobserved characteristics, the method uses local linear regression to measure how expenditure increases for an increase in the unobserved characteristic. I calculate this slope on each side of the nonlinearity. The final elasticity estimate is the difference between the two slopes as they approach the nonlinearity and the threshold enrollee, thus controlling for selection and simultaneity bias while isolating the response due solely to price. I then plug this price response into an elasticity formula which includes the price level at the nonlinearity to calculate final elasticity.

I apply this method to a detailed claims-level dataset for an employer-sponsored Consumer Driven Health Plan (CDHP). This plan was chosen because it has two nonlinear pricing points. Although baseline implementation of this method only requires individual-level final expenditure and the pricing structure associated with the expenditures, the greater detail in my data allows me to perform several robustness checks of the method with observable variables. I find elasticity estimates of -0.26 in lower expenditure ranges compared with -0.09 in higher spending ranges. These estimates are slightly above and below the RAND HIE estimate of -0.22, which was not a local estimate, but instead estimated over a broad range of spending. Previous literature uses elasticities as an indicator of moral hazard in insurance. I take my elasticity estimate one step further to measure moral hazard deadweight loss by calculating the counterfactual choices the elasticity predicts. The deadweight loss from full-coverage insurance is approximately 20% of final expenditures less than $1000.

This paper builds on the elasticity estimation literature in health, but also into a more general nonlinear estimation literature. Maximum likelihood approaches such as in Gary and Hausman (1978) and Hausman (1985) require specific distributional assumptions, whereas the method outlined here uses nonparametric estimation and requires only one strict monotonicity assumption. Other tax applications, such as Blomquist and Newey (2002) require substantial variation in prices across sample observations, which is less likely to hold for pricing in the sectors above than for taxes. Recent work by Saez (2010) and Chetty et al. (2013) also look at nonlinearities in the EITC tax code. Saez finds evidence consistent with changing labor hours in response to changes in the tax code, but finds that the most pronounced changes can be attributed to tax evasion. The method here is related, but has the advantage that the main condition of monotonicity links the outcome of interest and unobserved characteristics more flexibly, which allows for the lack of distinct bunching cited by Saez. Aron-Dine et al. (2012) also highlight the highly nonlinear environment of health insurance. The authors examine expenditure response to health insurance price within a year, to address the problem that a patient's price changes along his distribution of expenditure. This work highlights the difficulties of calculating an elasticity using only one price over a large range of values. This question of forward-looking or myopic behavior is not of first-order concern in this paper, however, because this method targets those just below or just above a deductible—individuals with relatively similar probabilities of reaching a post-deductible price. Those individuals well beyond a nonlinearity are not in the scope of this estimation method or local elasticity.

This paper has three contributions. First, I present a new method for measuring elasticities with minimal distributional or modeling assumptions. The method has commonly attainable data requirements and can be applied to consumer contracts. Second, this method is based on a common feature which previously introduced bias in estimation, but can now be used in a variety of sectors. Using nonlinearities means that this method is most useful for local elasticities along expenditure distributions. Finally, I use this method to estimate elasticities for an

employer-sponsored health insurance plan over two different areas of the spending distribution to measure changes in elasticity and then measure the deadweight loss of moral hazard.

The rest of the paper proceeds as follows: Section 2 sets up a general model of expenditure choice. Section 3 lays out the estimation framework, Section 4 describes the data, and Section 5 discusses the elasticity results. Section 6 describes the moral hazard estimation and results. Section 7 concludes.

2. General model of health expenditure

This section lays out a model of a patient's health expenditure choice within an insurance plan with nonlinear pricing. The goal of the model is to generate predictions on the relationship between the underlying distribution of health shocks of a population and the population's health expenditure choices. I will use the model to compare the distribution of final expenditures in the pre-deductible region versus the post-deductible region. The model presented here is similar to the framework in Huang and Rosett (1973), and generates the same reduced form predictions as the approaches in Manning et al. (1987), and Newhouse et al. (1980). In what follows, the patient is choosing his annual expenditure in dollars after having already chosen his insurance plan.[2]

To place this model in the example of a deductible, consider patients visiting a physician over the course of the year in response to various health shocks. A patient may benefit from multiple visits to the physician, with a cost for each visit. As the number of visits increases, the patient crosses the deductible and enters a higher coverage region. As the patient's marginal cost of a visit changes, the patient adjusts the frequency of his physician visits. This response is a combination of the severity and number of the patient's health shocks and the marginal cost to the patient of visiting the physician. Both the effect of health shocks and marginal cost will combine to decide the final end-of-year expenditure. The estimation will compare how these two effects reveal different patterns across patients who faced different marginal costs.

A patient has utility over his health expenditure, h, and composite good consumption, c. The patient's unobserved heterogeneity is his accumulated health shocks, θ.

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

Accumulated health shocks, θ, represent the shocks of varying number and severity the patient experienced over the course of the year. Higher values of θ are greater accumulated health shocks, and the population's end-of-year accumulated health shocks have a cdf, F_θ. This θ is defined broadly, because we will use it only to place an individual in relation to the sample population.

This general θ is useful for the estimation method presented below on questions involving final yearly spending, although it doesn't necessarily map empirically to diagnosis codes. Two patients could arrive at similar values of θ in different ways. However, by defining θ broadly we avoid ad hoc assumptions on quantifying the severity of diseases or attempting to rank health conditions.[3] The θ value will capture any unobserved characteristics about the patient which lead to health expenditures.

The utility function satisfies the following conditions:

$$U(h,\theta,c) = u(h,\theta) + c \tag{1}$$

$$\text{For any given } \theta,\ \exists\ \tilde{h} \text{ such that } \frac{\partial u\left(\tilde{h},\theta\right)}{\partial \tilde{h}} = 0 \tag{2}$$

$$\frac{\partial^2 u(h,\theta)}{\partial^2 h} < 0 \tag{3}$$

$$\frac{\partial u(h,\theta)}{\partial \theta} < 0 \tag{4}$$

$$\frac{\partial^2 u(h,\theta)}{\partial h \partial \theta} > 0. \tag{5}$$

Condition (1) is quasilinearity of composite good consumption in the utility function. Quasilinearity removes any income effects of health care consumption, which matches this application. Previous literature on income elasticity in health expenditures found estimates close to zero (Phelps, 1992) or generally lower for higher-income groups (DiMatteo, 2003). Additionally, this paper's application uses expenditures in the range of $200–$1800 for employed consumers, so income effects are not likely to be economically significant.

Condition (2) incorporates any nonmonetary costs of health care consumption and allows for marginal prices of zero, which are common in many nonlinear pricing applications. This condition sets an expenditure point for each level of θ where marginal utility crosses zero. Nonmonetary costs of health care consumption include the inconvenience cost of doctor visits such as travel time, waiting time, and treatment time (Janssen, 1992). Chiappori et al. (1998) also find that nonmonetary costs are important, leading to more price-sensitivity in physician services as compared to home visit services. This condition also captures that marginal utility might be negative for high levels of health expenditures if a patient has low accumulated health shocks.

Condition (3) states health expenditures exhibit decreasing marginal returns to utility. Condition (4) means that higher levels of health shocks decrease utility. Finally, Condition (5) states that there are complementarities between health expenditure and health shocks. For higher levels of accumulated health shocks, the marginal utility of health expenditure increases.

The patient's budget constraint balances the out-of-pocket costs of health expenditures and composite good consumption with patient income. Out-of-pocket costs are a function of the plan's pricing structure and the accumulated health expenditures, h. Denote the patient's budget constraint as:

$$c + OOP(h) \le y.$$

Annual income for each patient is y. Out-of-pocket expenses from the insurance plan's nonlinear pricing structure are $OOP(h)$.

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

$$OOP(h) = \begin{cases} h & \text{if } h \le \bar{h} \\ \bar{h} & \text{if } h > \bar{h} \end{cases}. \tag{6}$$

This pricing schedule determines marginal prices for an additional unit of health expenditure. The marginal price structure is:

$$p = \begin{cases} 1 & \text{if } h \le \bar{h} \\ 0 & \text{if } h > \bar{h} \end{cases}. \tag{7}$$

[2] This framework could be modeled alternatively as a joint decision of a patient and his doctor, where optimization maximizes the patient's health. The predicted relationship between health status and expenditure is the same.

[3] If the behavior of a particular population or health condition was of interest, the approach presented here could also be used to compare only yearly observations from that population, with sufficiently large datasets.

The patient optimizes over health expenditure choice h^*. The FOC over each marginal price segment, given a level of accumulated shocks, θ, is:

$$MU^*_\theta = \frac{\partial u(h^*, \theta)}{\partial h^*} = p. \quad (8)$$

Fig. 1 shows a patient's optimization problem with sample marginal utility curves that satisfy the utility conditions stated above. The marginal utility curves are combined with the nonlinear marginal price structure described above. Comparing the curves, the rightmost patient with the highest level of accumulated health shocks, θ'', has a higher marginal utility for the same level of h as compared to a patient with lower accumulated health shocks, θ.

A static model of choice predicts a gap in expenditures at the marginal price change in Fig. 1, and bunching if the marginal price change reverses. I do not empirically observe such a stark behavioral response because several assumptions implicit in such a simple diagram likely do not hold in this application. Saez (2010) presents a similar example in the tax literature that also addresses a lack of observed bunching. A clear gap assumes that patients have full control of spending down to dollar increments. While expenditures have a monotonic relationship with respect to accumulated shocks, such finite control over expenditures is not likely. Also, final expenditures must be continuous in this simple example, which is not generally the case in health care. Although some nonlinear pricing sectors such as electricity and water may have a more continuous quality to their products, the limited ability of the consumer to continuously monitor purchases may result in similar non-continuous expenditures. Furthermore, the size of the gap present in any model of health expenditures should be very small, because the highly inelastic indifference curves form an increasingly small gap in expenditures.

Two key points emerge from the FOC in Eq. (8) and Fig. 1:

1. Optimal health expenditure, h^*, depends on both accumulated health shocks, θ, and marginal price, p.
2. Optimal health expenditure, h^*, is strictly increasing in accumulated health shocks, θ.

The first point means that accumulated health shocks and the nonlinear pricing plan determine expenditure together, and any empirical estimation of h^* should flexibly incorporate both. Second, optimal health expenditure, h^*, strictly increases with higher accumulated health shocks, θ. This strictly monotonic relationship reflects the balance between utility condition (5), complementarities between h and θ, and utility condition (2), decreasing marginal utility of health expenditure.

The nonlinear pricing schedule's presence in Point 1 makes elasticity estimation difficult for several reasons, however. The first source of bias is that h_i and P_i are simultaneously determined by the deductible level of expenditure, $\bar{h}$. Additionally, the underlying unobserved θ_i determines both h_i and p_i. Higher levels of accumulated health shocks induce a higher h_i and its corresponding p_i. Fixing this simultaneous determination problem is nontrivial. Observable patient characteristics used to proxy for unobserved θ_i are likely correlated with the error term. For example, the latent accumulated health shocks related to an 80-year-old patient's expenditures compared with a 20-year-old patient's expenditures are correlated. This paper's method will use estimation that does not require an uncorrelated i.i.d. error term.

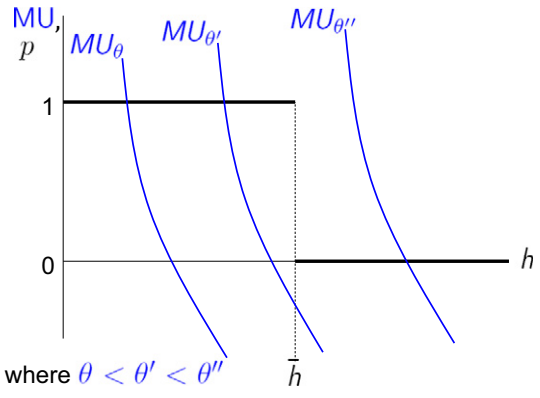

Fig. 1. Example of patients' optimization.

Previous health elasticity approaches use average expenditures of demographically similar populations to control for illness severity (Cherkin et al., 1989; Scitovsky and Snyder, 1972; Scitovsky and McCall, 1977). However, the health shock θ distribution potentially changes between comparison years. For example, demand for physician services includes time-confounding factors such as differences in flu seasons or the availability of new treatments or drugs. More importantly, the most price-sensitive patients have the opportunity to drop out of the sample through dis-enrollment as prices increase. This paper's estimation method compares behaviors within a year, so it relies on within-period variation. This avoids the intertemporal problem of exit from and entry into the insurance plan.

Another approach to the endogeneity problem above is to use instrumental variables for price that are independent of a patient's accumulated health shocks. For example, as in Eichner (1998), Kowalski (2010), where the authors take advantage of when unexpected injuries push a family over the deductible. This approach works well in a setting of general medical expenditures with family deductibles. Duarte (2012) broadens this instrumental variables approach using a wider population in Chile, an important contribution of non-US elasticity estimation. The advantage of the approach presented here is it can be used in a broad range of applications outside of such an empirical setting where "unexpected injury" instruments may be harder to construct, such as nonlinear rate schedules in public utilities and telecommunications, or prescription drugs. This approach can also be used on individual coverage observations, as in this application, or in chronic disease populations which lack an unexpected component.

The model of health expenditure choice above applies to a patient's decision over a defined benefit period. Although a patient's intertemporal decisions across or within benefit periods are interesting as well, the above framework is used for several reasons. First, the motivation for this method is to inform policy on nonlinear pricing schedules, which generally apply to expenditures over a pre-defined benefit period. Long-term elasticities are a different policy question.[4] Second, any effects on the estimates of a patient's ability to postpone treatment into the next benefit period depend on how much this delaying behavior varies across years in the population as a whole. Postponing treatment until the following year is a concern in this framework only if the extent of intertemporal substitution of the population changes yearly in a plan. This characteristic is unlikely to change in consecutive years given local changes in the price schedule. If the ability to postpone health care is relatively constant from year to year, then this simply represents another aspect of the underlying accumulated health shock distribution. The broad health shock measure includes the time-sensitivity of care. Empirically, patients in the data display great persistence in health care spending year over year.

3. Estimation method

The general model of health expenditures reveals two important determinants of health expenditures: marginal prices, p, and accumulated health shocks, θ. Marginal price data is generally easy to obtain. Data on

[4] Aron-Dine et al. (2012) address this understudied question nicely with a natural experiment revealing a patient's intertemporal choices along a year when faced with a deductible.

accumulated health shocks is much more difficult, however. Besides the difficulty in obtaining identifiable data on diagnoses, constructing a health shock variable out of diagnoses is necessarily subject to ad hoc assumptions. Most importantly, any measure of accumulated health shocks will still have a large observed error component because of the endemic difficulty in measuring health.

In the model above, accumulated health shocks are essentially a latent error term. The θ term is any unobserved health characteristic, and health expenditures are increasing in this unobserved, latent error term. Matzkin (2003) presents a framework to address this type of latent error nonparametrically. Identification using the Matzkin (2003) framework is off the fact that marginal price is constant within the estimation regions on each side of the nonlinearity, but the distribution of θ values change along the estimation window.[5]

This framework has several advantages in this setting. Nonparametric specification allows the unobservable random term to be built into the estimator from within the model. The estimator is freed from the assumption of additive error present in any OLS specification, allowing health status to influence expenditures flexibly and nonlinearly. Nonparametric estimation also relaxes the OLS requirement that the error term has a mean zero distribution.

Using the two conditions predicted by the model's optimal decision rule, I construct a flexible, nonparametric relationship to predict yearly health expenditures using the method described in Matzkin (2003). Patient *i*'s choice of yearly health expenditure, h_i, is a function of marginal price, p_i and health shocks, θ_i. The choice of health expenditure is:

$$h_i = G(p_i, \theta_i) \tag{9}$$

where G is a nonparametric function. This nonparametric function maps the space of marginal prices, {0,1}, and the space of health shock values, $\Theta \subset \mathbb{R}$, to the choice of health expenditures. That is, $G : \{0, 1\} \times \Theta \to \mathbb{R}$. As per the FOC outcome of the model above, G is strictly increasing in θ_i.

The known components of Eq. (9) are yearly health expenditures, h_i, and marginal price, p_i. The unknown components of Eq. (9) are the nonparametric function G and the latent health characteristics of the patient, θ_i. These latent health shocks are the error term of the nonparametric function. The goal of the elasticity estimation is to isolate the effect of the function G on the outcome of h_i due solely to changing p_i, while holding θ_i constant.

Fig. 2 displays the intuition behind the function G for the case of a deductible. Fig. 2 represents a narrow expenditure region surrounding the deductible. Consider first the left-hand Panel 2a. The horizontal axis is increasing in the latent component, the health shocks θ_i, while the vertical axis is increasing in health expenditure, h_i, which is observed. Any line on the figure shows how G maps an increase in the latent accumulated health shocks, θ_i, to a corresponding increase in choice of health expenditure, h_i, on the vertical axis. The vertical dashed line at $\bar{\theta}$ denotes the location of the level of health shocks which leads to spending at the level of the deductible. Patients pay a marginal price of one before hitting the deductible, and after the deductible patients pay a marginal price of zero, labeled $p = 1$ and $p = 0$, respectively.

Panel 2a displays the first step of identifying G. When the p argument in Eq. (9) is constant within each region, Matzkin (2003) shows that the function G can be identified within that region. The function $h_i = G(1, \theta_i)$, shown by the solid line, can be estimated based on all the $\theta_i < \bar{\theta}$ where $p = 1$. Above $\bar{\theta}$, the function $h_i = G(0, \theta_i)$, shown by the dashed line, can be calculated based on all the $\theta > \bar{\theta}$ where $p = 0$. These figures show G as a linear function for expositional purposes; final estimation is more flexible.

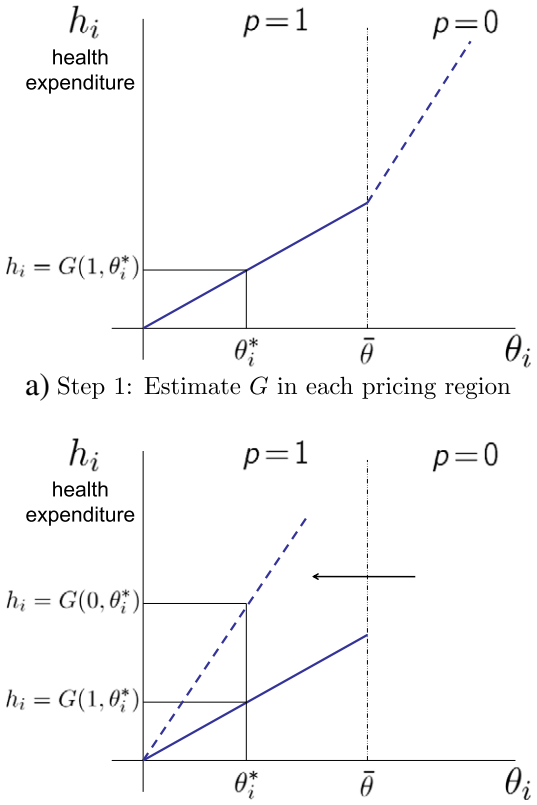

Fig. 2. Simplified exposition of estimator.

Note that the slope of the dashed line where $p = 0$ is more steep than the solid line where marginal $p = 1$. This shows the same increase in accumulated health shock maps to a larger increase in health expenditures when marginal price is zero. To predict health expenditures in the full out-of-pocket region on the diagram, choose a θ_i^* and plug it into the estimated equation $h_i = G(1, \theta_i^*)$.

Panel 2b shows the second step of the estimation—the prediction of health expenditures for different marginal prices. An elasticity calculation requires a patient's health expenditure choice at two different prices, given a fixed level of latent accumulated health shocks. To do this, use the slope of the function G where $p = 0$, which is the dashed line $h_i = G(0, \theta_i)$ in Panel 2a. Transfer this slope into the region where $p = 1$. Panel 2b demonstrates this where the dashed line with the steeper slope begins at the origin and rises above the solid line representing G when $p = 1$. To predict the choice of health expenditure for two different prices, the estimator isolates a given level of health shocks, θ_i^*, and uses the solid line $h_i = G(1, \theta_i^*)$ and the dashed line $h_i = G(0, \theta_i^*)$ to predict the choice of h_i when $p = 1$ and $p = 0$, respectively.

Unknown in the diagrams described so far is the distribution of the latent accumulated health shocks. If we knew the underlying distribution of health shocks, we could use this to estimate the function G. Matzkin (2003) shows that latent θ can be proxied by the percentile function because the function G is strictly increasing in θ_i. That is, even without knowing the exact shape of F_θ distribution, if a particular patient's health expenditures are in the 75th percentile of expenditures, then the latent health shocks of that patient are also in the 75th percentile of health shocks. The percentile function is bijective, surjective, and monotonically increasing. Let F_h be the distribution of observed

[5] Identification takes advantage of constant marginal price within a region. By design, the end-of-year prices in this framework are similar because estimation is over final expenditures near the nonlinearity. As such, this method is best suited for the policy framework described above, that of a local elasticity. This local elasticity may necessarily be an underestimate of elasticity over a large expenditure range, which would reflect greater differences in end-of-year prices. Marginal price is myopic, but the small range also creates similar forward-looking prices. Elasticity estimation over a larger range may not be able to capture potentially greater variances in forward-looking behavior.

individual health expenditures, h_i, and F_θ be the distribution of the individual omitted characteristics θ_i. The shape of the underlying distribution of omitted characteristics, F_θ, can be identified by mapping the percentiles of the health expenditure distribution, F_h. In this way, the unknown θ_i can be inferred by the econometrician, and treated as data.

Now that θ_i can be identified, we can estimate the function G from Eq. (9). In each region where marginal price is constant, this function maps a change in accumulated health shocks, θ_i, to a choice of health expenditure, h_i, given a fixed marginal price set by the price schedule.

Estimation uses a more flexible nonparametric approach than displayed in Fig. 2. Local linear regression allows the functional form of G in each marginal price region to be flexible. Where $\theta_i<\bar{\theta}$ – the left-hand side of the discontinuity – the estimation equation is:

$$\min_{a_L,b_L}\sum_{\theta_i<\bar{\theta}} K\left(\frac{\theta_i-\theta_0}{\kappa}\right)\left(h_i-a^L_{\theta_0}-b^L_{\theta_0}(\theta_i-\theta_0)\right)^2. \tag{10}$$

Analogously, where $\theta_i>\bar{\theta}$, the right-hand side of the discontinuity—the estimation equation is:

$$\min_{a_R,b_R}\sum_{\theta_i<\bar{\theta}} K\left(\frac{\theta_i-\theta_0}{\kappa}\right)\left(h_i-a^R_{\theta_0}-b^R_{\theta_0}(\theta_i-\theta_0)\right)^2 \tag{11}$$

where $a^L_{\theta_0}$, $a^R_{\theta_0}$ are constant coefficients and $b^L_{\theta_0}$, $b^R_{\theta_0}$ are slope coefficients based around the value θ_0, K is a kernel estimator with bandwidth k, and each θ_0 is in a series of points used by the local linear regression over the θ range of estimation.

Local linear regression uses Eqs. (10) and (11) to construct $G(1,\theta)$ and $G(0,\theta)$ over a window of observations on either side of the nonlinearity. The estimation method combines Matzkin (2003) and intuition from a regression discontinuity design approach. The function $G(1,\theta)$ measures the rate of change in health expenditures as θ increases, for a fixed marginal price of 1. The function $G(0,\theta)$ measures the rate of change in expenditures as θ increases, for a fixed marginal price of 0. As these two functions approach the nonlinearity, the limit latent θ value is the same, yet the marginal price component of the functions G is constant. The method compares the difference in the slopes of G as they approach the limit. The estimator controls for the price schedule's simultaneity between price and quantity by isolating the changing θ values in the constant marginal price region.

The estimator is essentially measuring the change in the slope of a cumulative distribution function at the point of the nonlinearity. Unobserved heterogeneity, θ, changes over the entire estimation window, so the estimator recovers behavioral responses using a nonparametric inverse cdf identified via the Matzkin (2003) framework. The function G is similar to a cdf because the unobserved accumulated health shocks can be proxied with a percentile function. Identification occurs through the changes in the slope in each region, where price is fixed, but the cdf relationship is flexible. Cdf interpretation rearranges Fig. 2 so that the horizontal axis would be spending and the vertical axis would be percentiles of spending. The slope of G measures the percentiles of spending for a given level of expenditure. The change in slope simply measures how a small change in θ leads to a change in the probability that the corresponding h_i is a large jump in cumulative probability.

Though borrowing intuition from a regression discontinuity design, one important difference between this approach and RD design is that here the patient's omitted characteristics are the forcing variable that determines a patient's marginal price. Because the choice of h_i is strictly monotone in the omitted characteristics θ_i, the patient's level of θ_i is what forces him to the left or to the right of the discontinuity. Identification in this method is not the same as a regression discontinuity design, which requires omitted characteristics to be the same in each region. This method specifically does not require this assumption. Here, G nonparametrically incorporates changes in unobserved characteristics, and identification uses the difference the relationship generated by the unobserved characteristics, G, between constant marginal price regimes.

The final formula for calculating elasticity uses the local linear regression slope coefficients at the limit because we are interested in the point where patients' omitted θ are the most similar. The local linear regressions Eqs. (11) and (10) from above are rearranged to replicate Panel 2b as follows:

First write h_i for the observations $\theta_i<\bar{\theta}$ at the threshold limit $\theta=\bar{\theta}$

$$\begin{aligned} h_i|(\theta_i<\bar{\theta}) &= a^L_{\bar{\theta}}+b^L_{\bar{\theta}}(\theta_i-\bar{\theta}) \\ &= a^L_{\bar{\theta}}-b^L_{\bar{\theta}}\bar{\theta}+b^L_{\bar{\theta}}\theta_i \\ &= A^L+b^L_{\bar{\theta}}\theta_i \end{aligned} \tag{12}$$

$$\begin{aligned} h_i|(\theta_i>\bar{\theta}) &= a^R_{\bar{\theta}}+b^R_{\bar{\theta}}(\theta_i-\bar{\theta}) \\ &= a^R_{\bar{\theta}}-b^R_{\bar{\theta}}\bar{\theta}+b^R_{\bar{\theta}}\theta_i \\ &= A^R+b^R_{\bar{\theta}}\theta_i \end{aligned} \tag{13}$$

where $a^L_{\bar{\theta}}$, $a^R_{\bar{\theta}}$, $b^L_{\bar{\theta}}$, and $b^R_{\bar{\theta}}$ are the constants and slope coefficients at the limit $\bar{\theta}$.[6] and A_L combines the terms $a^L_{\bar{\theta}}$ and $-b^L_{\bar{\theta}}$ and similarly for A_R2E.

To replicate Fig. 2, write a general equation for h_i starting at the left region intercept, A^L, by using an indicator function equal to one when $p=1$:

$$h_i=A^L+b^R_{\bar{\theta}}\theta_i+\left(b^L_{\bar{\theta}}-b^R_{\bar{\theta}}\right)\theta_i\ 1\{p=1\}. \tag{14}$$

Eq. (14)'s slope coefficient order, $\left(b^L_{\bar{\theta}}-b^R_{\bar{\theta}}\right)$, is for the deductible case, where the $p=0$ region is the right-hand side. Eq. (14) solved for $h_i(p=1)$ and $h_i(p=0)$:

$$h_i(p=1,\theta_i)=A^L_{\bar{\theta}}+b^L_{\bar{\theta}}\theta_i \tag{15}$$

$$h_i(p=0,\theta_i)=A^L_{\bar{\theta}}+b^R_{\bar{\theta}}\theta_i. \tag{16}$$

In this application, I use percentage change elasticity. This is to account for the fact that enrollees are moving from a no coverage region into a full coverage region.[7] Elasticity, η, of moving from full out-of-pocket into full coverage is:

$$\begin{aligned} \eta &= \frac{\%\Delta h_i}{\%\Delta p} \\ &= \left[\frac{h_i(p=0,\theta_i)-h_i(p=1,\theta_i)}{h_i(p=1,\theta_i)}\right]\Big/\left[\frac{0-1}{1}\right] \\ &= \frac{-\left[A^L_{\bar{\theta}}+b^L_{\bar{\theta}}\theta_i-\left(A^L_{\bar{\theta}}+b^R_{\bar{\theta}}\theta_i\right)\right]}{A^L_{\bar{\theta}}+b^R_{\bar{\theta}}\theta_i} \\ &= \frac{-\left(b^R_{\bar{\theta}}-b^L_{\bar{\theta}}\right)\theta_i}{A^L_{\bar{\theta}}+b^R_{\bar{\theta}}\theta_i}. \end{aligned} \tag{17}$$

Which evaluated at $\bar{\theta}$ is equal to:

$$\eta=-\left(b^R_{\bar{\theta}}-b^L_{\bar{\theta}}\right)\frac{\bar{\theta}}{\bar{h}}. \tag{18}$$

[6] Refer to the Appendix A for a construction of the estimator using the parametric form of the general utility model. In this case, the coefficients can be built out of structural parameters.

[7] Given the extreme change in marginal price, percentage change is more informative than a midpoint elasticity, although other applications with smaller price changes could certainly use Eq. (14) in a midpoint formula if so desired. Aron-Dine et al. (2013) use a midpoint formula, for example.

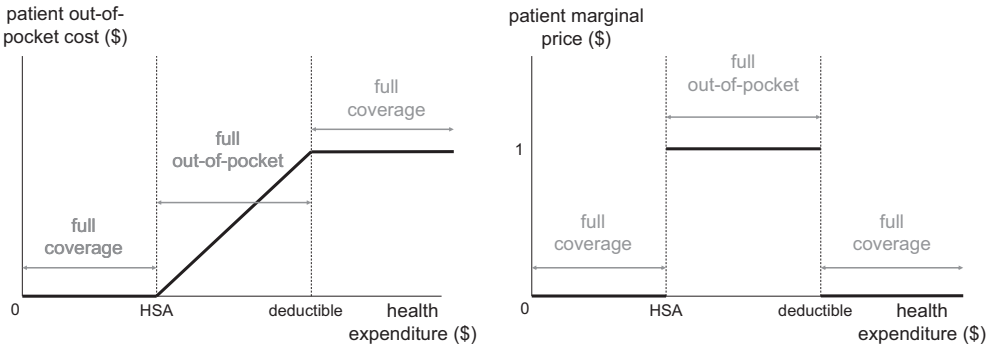

Fig. 3. Nonlinear pricing schedule, consumer driven health plan.

The case of a Health Savings Account (HSA), where patients face opposite marginal prices, reverses the order of b_L and b_R slope coefficients.

4. Data

4.1. Data

The dataset is proprietary claims-level data for an employer with several locations. The employer is self-insured and all individual claims are reported for each of the three years 2002, 2004, and 2005.[8] Each claim entry contains all information necessary for classifying the services received and to remit payments. Each claim has information on the costs incurred by the patient and the amount covered by the employer, as well as information on the treatment facilities, procedure codes, and diagnoses. An advantage of a self-insured employer is that income information is available to identify the socioeconomic status of the population.

In particular, I study the Consumer-Driven Health Plan (CDHP) option available to enrollees, which is a high-deductible plan. Enrollees had the option to enroll in four other plans offered by the employer. The effect this may have on estimates comes through in the broader generalizability of the type of enrollees who chose this plan, but not in the estimates within the plan. I show that this estimation method is not subject to selection bias for within-plan estimates. Characteristics of patients enrolled in this plan determines the broader applicability of these elasticity estimates.

The CDHP plan contains two nonlinearities. The first nonlinearity results from an employer-funded Health Savings Account (HSA), where the employer deposits funds that can be used to purchase health care from the first dollar spent until the patient exhausts the HSA. The second nonlinearity is a deductible. Fig. 3 illustrates the nonlinear structure of this plan.

The threshold levels of both the HSA and the deductible change from year to year. This variation in nonlinearity thresholds helps identify patients' responses to the nonlinear pricing schedules in two ways. If the nonlinearity changes each year, this lends robustness to the estimation method if estimates remain similar over years. Second, the estimator has strong validity for observations just at the nonlinearity, but validity is limited for observations far from the nonlinearity. However, with estimates at proximate intervals, elasticity is estimated over an expanded range of expenditures. Table 1 lists plan nonlinearities during the years 2002, 2004, and 2005.

Table 2 reports plan summary statistics for patients who enrolled under single coverage for the entire 12-month benefit period. The first rows report the yearly means and medians of total expenditure, employer cost, the yearly means of the amount of HSA used, and the amount of deductible fulfilled. Average total expenditure for all three years was $7387. However, health expenditure distributions tend to be skewed, so median total expenditure was lower in all years. Lower expenditure levels of employer cost compared to total expenditures reflect positive out-of-pocket costs to patients. Patient expenditure variables in Table 2 include the amount of the deductible fulfilled and the amount of the HSA used. An average patient's total incurred spending toward his deductible was $644. The average amount used of the HSA on incurred spending was $319.[9]

Patient-level characteristics for single-coverage, full-year enrollees in the insurance plan are reported in the bottom rows of Table 2. Plan enrollment in this category grew from 165 enrollees in 2002 to 349 in 2005. The average age over all three years is 48, and the average salary for the enrollees is $55,934. The plan enrolled 72% women.

5. Elasticity estimation results

I estimate elasticities over patients' yearly expenditures within the estimation windows for each of the three years. Yearly estimates are necessary because the level of the nonlinearities, and thus the threshold $\overline{\theta}$, change every year. Table 3 displays the results for the first plan nonlinearity, the HSA. Table 4 displays results for the second nonlinearity, the deductible. I calculate standard errors using the asymptotic distribution properties developed in Bajari et al. (2010).[10]

The heading of Table 3 reports the elasticity formula from Eq. (18), arranged for the HSA nonlinearity. The first column of Table 3 reports the year, and the second column is the number of observations within the estimation window around the HSA. The third column reports the level of the HSA in that year. The final column reports elasticity estimates.

The elasticity estimates around the HSA nonlinearity are −0.25 in 2005, −0.26 in 2004, and −0.33 in 2002. All are inelastic. The similarity across years is noteworthy because the nonlinearity value of $\overline{h}$ changes across years between $500 and $750. The estimates reported in this

[8] In the years 2002–2005, missing enrollee assignment codes in 2003 prevented using this year in estimation.

[9] The first year this plan was available was 2002, which explains the lower enrollment statistics and associated spending levels. An advantage of several years is that results can be compared across several years if there is concern that the first year was different because of its novelty.

[10] In this method, the difference between the true limit of the expenditure choices and the estimated value of this limit from above and below converges to an independent exponential variable with hazard rates of $f^{-}(h_L)$ from below and $f^{+}(h_R)$ from above. The local linear regression slopes $b^L_{\overline{\theta}}$ and $b^R_{\overline{\theta}}$ are then the inverse of these hazard rates, respectively. The hazard rates are estimated from the data with a one-sided kernel density. The difference between the estimated and true value of the difference in the slopes, $\left(b^L_{\overline{\theta}} - b^R_{\overline{\theta}}\right)$, converges asymptotically to a normal distribution with variance calculated using $f^{-}(h_L)$, $f^{+}(h_R)$, and properties of the chosen kernel.

Table 1
Consumer driven health plan (CDHP) structure.

	First nonlinearity	Second nonlinearity
Year	HSA	Deductible
2002	$500	$1250
2004	$750	$1500
2005	$600	$1500

Table 2
Data summary for full sample.

Year	2002	2004	2005	Overall
Average expenditures				
Total expenditure	$5251	$7130	$8647	$7387
(Median)	$1492	$2024	$2288	$2016
Employer cost	$4852	$6560	$7824	$6746
(Median)	$990	$1551	$1899	$1537
Deductible used	$540	$665	$672	$644
HSA used	$257	$377	$292	$319
Demographics				
Enrollees	165	341	349	855
Percent female	71	72	72	72
Age	46	48	48	48
Salary	$58,783	$51,224	$58,965	$55,935
Single coverage	100%	100%	100%	100%

Includes only single coverage, full year enrollment.

Table 3
Elasticity changing from full out-of-pocket to full coverage.

Estimates at HSA nonlinearity			
Year	N	Measured at	**Elasticity** (η)
2002	39	$500	−0.33[b] (0.15)
2004	55	$750	−0.26[a] (0.07)
2005	61	$600	−0.25[b] (0.12)

Standard errors are in parentheses.
[a] Denotes significance at 1% level.
[b] Denotes significance at 5% level.

Table 4
Elasticity changing from full out-of-pocket to full coverage.

Estimates at deductible nonlinearity			
Year	N	Measured at	**Elasticity** (η)
2002	22	$1250	−0.09 (0.17)
2004	40	$1500	−0.08[a] (0.05)
2005	34	$1500	−0.09 (0.06)

Standard errors are in parentheses.
[a] Denotes significance at 1% level.

paper are for a $300 window and based on an Epanechnikov kernel, but are robust to other specifications.[11]

Table 4 reports the elasticity estimates at the deductible, the plan's second nonlinearity. The first column reports the year, the second column records the number of observations in the estimation window around the deductible, and the third column reports the level of the deductible in that year. The last column reports the elasticity estimates. The lower significance level of these estimates is due to the small sample size at the higher expenditure levels of the deductible. The elasticities at this nonlinearity are highly inelastic. The elasticity is −0.09 in both 2002 and 2005, and −0.08 in 2004.

5.1. Discussion

These estimates place consumers in the inelastic region of demand, confirming common sense about health care being a necessary good. The estimates at the HSA are slightly higher than the RAND Health Insurance Experiment estimates. However, the estimates at the deductible are slightly lower then the RAND HIE. The RAND HIE found elasticities between −0.17 and −0.22, but calculated over the whole spending range instead of local estimates at different spending levels Keeler and Rolph (1988). Aron-Dine et al. (2013) give a nice discussion of different approaches to elasticity calculation in the RAND HIE. Most similar to the framework here, the authors show that HIE participants switching from free care to a 95% coinsurance had an elasticity of approximately −0.23. The advantage of this approach's local estimation is it sidesteps the limitations of calculating a single price for a complex range of expenditures and nonlinearities in plans.

Given the deductible framework in this paper, the most appropriate literature comparisons are those examining the changes from full insurance to none, or vice versa. Boes and Gerfin (2013) compare a managed care plan with temporary full insurance against recently introduced cost-sharing measures. The authors also find elasticity decreasing for higher levels of health expenditures, with an average elasticity of −0.148, after adjusting for selection out of the HMO.

Outside of the RAND HIE, previous literature estimates elasticities for a range of types of health expenditures. Among the estimates that apply specifically to general medical expenditures, Cherkin et al. (1989) found primary care physician visits responded the most to the introduction of a $5 copay. Visits decreased by approximately −10% for what was approximately 15% of a typical visit charge. This response may be slightly larger than the estimates in my data, but the Cherkin et. al data came from an HMO where primary care physician's roles as gatekeepers imply a higher time cost per visit as a baseline comparison. Selby et al. (1996) found a −14% decline in emergency room visits from a much larger copay introduction of $25–$35. The elasticity implied by these numbers is slightly less responsive than those presented above, which follows from emergency room visit versus physician visits.

Claims-level data allows us to look more closely into why these local elasticities are different at different points of the spending distribution. Estimates are more inelastic at the higher deductible nonlinearity compared with the HSA nonlinearity. The first explanation for this may be because a higher level of expenditures means a higher level of illness severity, which is associated with less price sensitivity. To examine this hypothesis, Tables 5 and 6 compare facility types and services in

Table 5
Facility type comparison: HSA vs. deductible.

Percent	Facility type	Expenditures
HSA		
70%	Physician's office	$66,137
6%	Independent lab	$5762
5%	OB/GYN office	$4549
4%	Hospital outpatient	$4055
Deductible		
58%	Physician's office	$79,312
9%	Independent lab	$12,614
8%	Hospital outpatient	$10,881
5%	Clinic	$6154

Expenditures are total within the local estimation window.
Windows are within a $300 window on each side of the nonlinearity.

[11] See the Appendix A for local linear regressions using different window sizes in the ranges of [250, 400] and different bandwidth choices for Gaussian and Uniform kernels.

Table 6
Type of service comparison: HSA vs. deductible.

Percent	Service type	Expenditures
HSA		
22%	Physician care	$20,518
19%	Lab/pathology	$17,536
13%	Vision	$12,398
12%	X-ray diagnostic	$11,252
10%	Routine physical	$9710
Deductible		
19%	Lab/pathology	$25,902
13%	X-ray diagnostic	$18,402
8%	Psychotherapy	$11,377
6%	Vision	$8849
5%	Surgery	$7494

Expenditures are total within the local estimation window.
Windows are within $300 on each side of the nonlinearity.

the HSA local estimation window versus the deductible local estimation window.

Table 5 shows the percentage of expenditures within the local estimation window for a given facility type, comparing the HSA and deductible estimation windows. For both windows, physician offices are the largest category of spending. However, physician offices are a much higher percentage of expenditures in the HSA window, at 70% compared to only 58% around the deductible. More claims in the deductible take place at more intensive facilities compared to the physician office. Hospital outpatient facility use is one of the more striking differences between the two estimation samples. Hospital outpatient facilities were 8% of expenditures in the deductible, which is double the percent found in the HSA sample. This indicates the higher deductible spending levels are correlated with more severe health shocks, which leads to lower price sensitivity.

The services reported by patients' claims also suggest increasing severity toward the deductible, as well as an indication that spending may be less discretionary and more lumpy as the mix of services changes between the HSA region and the deductible region. Table 6 reports the percentage of expenditures in the top 5 services in the estimation windows. The two estimation windows both show significant spending in lab/pathology, X-ray diagnostics, and vision services. However, the HSA region shows a combined total of 32% spending on physician care and routine physical services, neither of which makes the top five at the deductible estimation window. Patients have a greater ability to control the amount of routine physical services, in deciding to complete an annual physical or not. In contrast, the deductible region begins to have more Surgery services (5%), which is a service that may be less discretionary than routine physical.

The broader application of these elasticities depends on the sample population and its generalizability to a larger population. This method's estimates have a high level of internal validity within the sample population, but the external validity of these estimates to populations outside the sample requires further discussion. There are two points of general applicability of this sample population. First, the spending levels are typical of the median spending of a privately insured individual under 65 during the time period. U.S. median health expenditure was $1032 in 2005 (AHRQ, 2005), which lies between the expenditure levels of the HSA estimates ($500–$750) and the deductible estimates ($1250–$1500). Comparing these estimates to previous work also should be restricted to the mix of services found in these HSA and deductible ranges. Secondly, the insurance is employer-sponsored, which is the most common type in the U.S.

There are a few ways in which this population may not translate to a more general U.S. health care population. The patients in this sample are richer than the average U.S. population, thus elasticity for a general population may be lower if income effects hold. Additionally, the patients in the sample population are, on average, healthier than the firm's population as a whole.[12] If the severity of health shocks is lower than a more general privately-insured population, these elasticity estimates may show patients to be more responsive to marginal prices if elasticity decreases with the severity of a health shock. Finally, willingness to participate in a high-deductible plan could have implications for the risk-aversion parameters of patients. In a study on individuals in private insurance, van de Ven and van Praag (1981) find that increasing income and education correlate positively with demand for a deductible.[13] Given these caveats, these estimates are most useful as a local elasticity, but are not atypical of expenditures in U.S. employer-sponsored plans.

5.2. Robustness

The identifying assumption in the estimation method is that the latent error term, θ, is continuous across the nonlinearity but the price changes. In this application, the latent error is accumulated health shocks over the year of the insurance pricing structure. The θ term should be capturing differences in these accumulated health shocks within a $300 window on each side of the spending level of each nonlinearity. The resulting elasticity can be interpreted as a local interaction of health shocks and the pricing structure. Although claims-level detail is not strictly necessary to implement the estimation method, we can use detail on patient claims to verify that these estimates are capturing reactions to accumulated health shocks within a narrow window of final expenditures, and not fixed characteristics such as age and gender.

As discussed in the estimation section, the accumulated health shocks θ are the forcing variable into pre- or post-nonlinearity region. Since age and gender are not health shocks within a $400 window, but instead a fixed endowment, Tables 7 and 8 test if there is selection on age and gender on either side of the nonlinearity.[14]

Table 7 shows the results of both a means test and K-S test for differences in the distribution of ages and gender on each side of the HSA estimation sample. None of the tests can reject the hypothesis that distributions of age and gender are the same on each side of the nonlinearity. For each year, I use a means test to compare the average ages of sample individuals that fall pre-HSA versus post-HSA. Besides just the mean, I also run a K-S test of equality of distributions comparing the age distribution just before versus just after the HSA nonlinearity. Both the means test and the K-S tests cannot reject equality of the age distribution in the pre versus post HSA samples, up to a 15% confidence level. The bottom half of Table 7 performs the same tests on the gender ratios and cannot reject equality for at least the 15% level for 2004 and 2005, and at least the 10% level for 2002.

Table 8 tests for equality of distributions in age and gender at the higher nonlinearity threshold of the deductible. Neither the means tests or the K-S test for equality of age can be rejected for 2002–2005, for at least a 15% confidence level. The only test to come close to showing significant differences pre deductible and post deductible is the

[12] This data is similar to the national statistics on enrollees in Consumer Driven Health Plans. Enrollees in CDHPs tend to be richer, more educated, and healthier than their counterparts in other employer-sponsored plans (Kaiser Family Foundation, 2006). The Kaiser Family Foundation's, 2006 Survey of CDHP enrollees found that 45% of CDHPs enrollees had salaries over 75,000, compared with 30% of the control group of employer-sponsored enrollees in other plans. 64% of CDHP enrollees reported being in excellent or very good health, compared with 52% in the control group.

[13] Einav et al. (forthcoming) develop a model where patients select an insurance plan based on both aggregate risk and the "slope" of the pricing structure. This suggests that there may be selection into high-deductible plans based on the pricing schedule in addition to expected total out-of-pocket costs. Given this model, the patients selected into a CDHP plan may be less price-responsive than a general population. This could bias my final estimates downward compared with a general population.

[14] Age and gender likely enter the patient's problem in the initial choice of plan, as a fixed health status level. In the estimation window, the change in the θ variable maps to a forcing variable within a narrow window of expenditures—the patient's age and gender would enter as a fixed effect.

Table 7
Testing for differences in age and gender in estimation sample: before vs. after HSA.

Year	N	Means test	K-S test	Result
Age				
2002	39	$t = -1.60$ ($p = 0.12$)	$D = 0.31$ ($p = 0.25$)	Cannot reject[a]
2004	55	$t = -1.20$ ($p = 0.23$)	$D = 0.20$ ($p = 0.54$)	Cannot reject[a]
2005	61	$t = 0.17$ ($p = 0.87$)	$D = 0.21$ ($p = 0.45$)	Cannot reject[a]
Gender				
2002	39	$t = 1.51$ ($p = 0.14$)	$D = 0.24$ ($p = 0.54$)	Cannot reject[b]
2004	55	$t = 0.08$ ($p = 0.23$)	$D = 0.01$ ($p = 1.00$)	Cannot reject[a]
2005	61	$t = -0.07$ ($p = 0.95$)	$D = 0.01$ ($p = 1.00$)	Cannot reject[a]

T-statistics for means tests between before and after HSA nonlinearity.
D-statistics for Kolmogorov–Smirnov test of distributional equivalence.
P-values in parentheses under the respective statistics.
[a] Test is not significant at the 15% level or below.
[b] Test is not significant at the 10% level or below.

means test on gender for 2005, which still is not significant at the 5% level.

One possibility for why a patient crossed over the threshold that is not consistent with the model above is that final claims were for a more expensive doctor, which would mean that elasticity estimates were capturing a geographic effect, not a response to health shocks. Therefore, another robustness check compares the location of services before and after the nonlinearity. As such, Table 9 compares service zipcodes on each side of the HSA limit in the estimation sample for the two years when zipcode information was available. In 2004, the top three zipcodes were identical on each side of the nonlinearity. In 2005, two of the top three zipcodes were the same. The non-matched zipcodes on each side were bordering zipcodes in a metropolitan center; one of the zipcodes is for business addresses in the zipcode area. This supports that differences in doctors are not the heterogeneity forcing patients to one side or the other of the nonlinearity.

For the deductible estimation sample, Table 10 shows several common zipcodes appear in on each side of the deductible. The zipcodes xx480 and xx440 are different business addresses in the same metropolitan center, so the most common zipcodes in both years before and after the nonlinearity are also from the same geographical area. Some of the differences in the deductible variety of zipcodes may be attributed

Table 8
Testing for differences in age and gender in estimation sample: before vs. after deductible.

Year	N	Means test	K-S test	Result
Age				
2002	22	$t = 0.12$ ($p = 0.91$)	$D = 0.23$ ($p = 0.95$)	Cannot reject[a]
2004	40	$t = -0.36$ ($p = 0.23$)	$D = 0.20$ ($p = 0.76$)	Cannot reject[a]
2005	34	$t = -0.67$ ($p = 0.51$)	$D = 0.33$ ($p = 0.21$)	Cannot reject[a]
Gender				
2002	22	$t =$ -0.91 ($p = 0.37$)	$D = 0.21$ ($p = 0.98$)	Cannot reject[a]
2004	40	$t = -0.36$ ($p = 0.72$)	$D = 0.05$ ($p = 1.00$)	Cannot reject[a]
2005	34	$t = -1.85^{*}$ ($p = 0.07$)	$D = 0.26$ ($p = 0.49$)	Cannot reject[b]

T-statistics for mean tests between before and after HSA nonlinearity.
D-statistics for Kolmogorov–Smirnov test of distributional equivalence.
P-values in parentheses under the respective statistics.
[a] Test is not significant at the 15% level or below.
[b] Test is not significant at the 5% level or below.

Table 9
Zipcode comparison within the HSA estimation window.

	Pre-HSA	Post-HSA	Pre-HSA	Post-HSA
	2004		2005	
Expenditure	xx480	xx480	xx440	xx480
Locations	xx455	xx455	xx404	xx485
Partial zipcode	xx440	xx440	xx480	xx440

Expenditure location is ranked by total annual expenditures.
Zipcode not available for 2002.

Table 10
Zipcode comparison within the deductible estimation window.

	Pre-deduct.	Post-deduct.	Pre-deduct.	Post-deduct.
	2004		2005	
Expenditure	xx480	xx440	xx440	xx440
Locations	xx440	xx403	xx480	xx485
(Partial zipcode)	xx812	xx435	xx486	xx480

Expenditure location is ranked by total annual expenditures.
Zipcode not available for 2002.

to the increased diversity of services present in claims at the deductible, as shown in the discussion section and Table 6.

Finally, another concern about the patients in the nonlinearity estimation windows is that these patients might simply have run out of time to file insurance claims as the year ended compared to patients outside the estimation window. However, although December is the most common month of last claim both in and out of the estimation sample, less than 40% of the patients in the estimation sample filed their last claim in December. There was a positive number of patients in the sample showing a last claim in every month of the year, with increasing probability of a last claim as the year progressed. This pattern is consistent with the larger population's month of last claim data.[15]

6. Moral hazard estimation

6.1. Setup

Plans with high reimbursement rates are often cited as inducing overconsumption and moral hazard because the patient bears very little of the cost of his health care. This refers to ex-post moral hazard, where patients consume more health care because insurance insulates them from its cost. Demand elasticities are often used as a proxy for ex post moral hazard. I take this further by measuring the welfare impact between observed expenditures in my generous plan with predicted patient expenditures faced with full out-of-pocket costs. Elasticity measures reveal patients' responses to price. This counterfactual gets closer to what policy makers are really interested in, which is the deadweight loss of insurance-induced demand changes.

To measure the amount of deadweight loss, the extent of moral hazard, I compare patient choice in two scenarios. The first scenario is a generous insurance plan, which covers 100% of care. In the second scenario, all care is paid for out-of-pocket for the same range of expenditures.[16] Fig. 4 illustrates the basic concept behind the deadweight loss calculation. When the consumer is fully insured and pays zero percent of his out-of-pocket costs, the consumer's choice of health care, h_{i1}, gives him consumer surplus of $A + B + C$. The insurance company pays for the entire cost of this choice of health care, which amounts to

[15] See Appendix A for a histogram of month of last claim for each population.
[16] In these scenarios, I use only changes in coinsurance. Although premiums are also part of an insurance plan, choosing the changes in premiums between the two scenarios would require several other assumptions on patient responses and how actuarially fair the insurance plan is. The two scenarios going from full coverage to no coverage with no change in premiums create an upper bound on the amount of moral hazard.

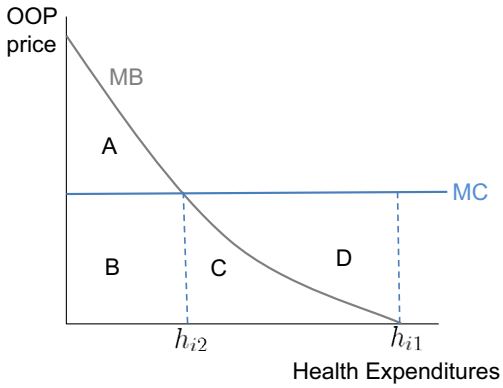

Fig. 4. Patient choice and welfare in counterfactual.

$B + C + D$. Net welfare in the free care scenario is $(A–D)$. I predict the consumer's choice, h_{i2}, if paying 100% of the cost, with consumer surplus equal to $A + B$. In this second scenario the cost of care is the area B. Net welfare in the full out-of-pocket scenario is A. Thus, the area D is the amount of deadweight loss. To get a measure of the deadweight loss D, I take the change in expenditure between full insurance and full out-of-pocket, $C + D$, and subtract out the transfer the patient would need to compensate him for the loss of C in the full out-of-pocket scenario.

For empirical estimation of the reduction in health expenditure choice, I link each value of the estimated θ_i to observed full coverage health expenditure, h_{i1}. I then use the slope coefficients estimated in Section 5 to predict the counterfactual spending of each patient given full out-of-pocket payment, h_{i2}. I use only those patients from the original HSA estimation window for the moral hazard calculation, and I assume a constant elasticity around the nonlinearity. The HSA window is the most reasonable of the two for calculations abstracting from income effects.

To calculate the compensating transfer for each patient, I use the following utility function for the choice of health expenditure. Utility is based on the conditions set on the general utility function in Section 2, where the preference parameters satisfy $\rho \in (0, 1)$ and $\gamma > 0$:

$$U_i(h_i, \theta_i, c_i) = \theta_i h_i^{\rho} + (1-\theta_i)c_i - \gamma\, h_i. \tag{19}$$

The parameters θ_i, h_i, and c have the same interpretation as in the general mode. The γ parameter captures nonmonetary costs of health care, such as time, and satisfies Condition (2) of the utility conditions in Section 2. The utility function is decreasing in illness as long as $h_i^{\rho} < c_i$.[17] Utility is increasing in health care expenditures up to marginal utility of zero, and marginal utility is decreasing in additional health expenditures. The final Condition (5) of the Utility Conditions holds that health care increases utility by a greater amount at higher levels of illness. I use individual salary information for non-health composite good consumption.

The two remaining unknowns in the utility function are: the risk-aversion parameter on health care expenditures, ρ, and the nonmonetary cost of health care, γ. I estimate these parameters from the data using a general method of moments approach which chooses parameters which minimize the difference between observed utility values and the predicted utility value in the counterfactual. The estimated values for $\hat{\rho}$ were approximately 0.23 and for $\hat{\gamma}$ were approximately 0.31.[18]

[17] This condition always holds in my data, because c_i is the salary left after health care expenses, expenditure is less than $1000 in the window, and $\rho < 1$.

[18] See the Appendix A for further discussion of the GMM and tables of the utility parameter estimates.

Table 11
Deadweight loss magnitude.

Year	p25	p50	p75
Level ($)			
2002	9.65	66.26	144.94
2004	46.88	133.88	185.97
2005	9.63	25.88	64.38
Percent of spending (%)			
2002	4.75	20.29	31.68
2004	9.16	21.64	27.19
2005	2.68	6.59	12.65

The compensating transfer C between the two scenarios is the amount of additional non-health care income the patient needs to remain indifferent between his utility consuming h_{i1} and his utility consuming only the lower h_{i2}. The transfer C enters the utility function as additional income in the full out-of-pocket cost scenario. I set the two scenarios' utility functions equal to each other, and solve for C. The compensating transfer is then a function of known variables: estimated θ_i, $\hat{\rho}$, and $\hat{\gamma}$, observed h_{i1}, and predicted h_{i2}.

$$C_i(h_{i1}, h_{i2}, \theta_i, \hat{\rho}, \hat{\gamma}) = \frac{\theta_i}{(1-\theta_i)}\left(h_{i1}^{\hat{\rho}} - h_{i2}^{\hat{\rho}}\right) - \frac{\hat{\gamma}}{(1-\theta_i)}(h_{i1} - h_{i2}). \tag{20}$$

The compensating transfer is decreasing in $\hat{\rho}$. This means that the size of the compensating transfer decreases as a patient becomes less risk adverse in health care expenditures. The compensating transfer is also decreasing in $\hat{\gamma}$, indicating that the amount of compensation for lost health care decreases as the nonmonetary costs of the foregone care increase.

6.2. Moral hazard estimation results

The measure of deadweight loss of moral hazard here is an upper bound, because it compares full coverage with no coverage. The deadweight loss is the portion of the difference in health expenditures where the marginal cost is greater than the patient's marginal utility—the part of the change that the patient does not require back in compensating transfer.[19] Table 11 displays the empirical results of the counterfactual in levels. The magnitude of deadweight loss is highest in 2004, which corresponds to the highest spending window. The HSA cutoff was $750 in 2004 versus $600 and $500 in the other years. The median level of deadweight loss was approximately $66 in 2002, $134 in 2004, and $26 in 2005.

To put the magnitude of the deadweight loss in context, Table 11 also reports the deadweight loss as a percentage of each patient's free-care level of spending. The median percentage of free-care spending level was approximately 20% in both 2002 and 2004. The median percentage was lower in 2004, at approximately 7%. Patients at the beginning of the estimation window had the smallest change in spending, and thus the lowest levels of deadweight loss. The 25th percentile for deadweight loss as a percentage of spending was 2.68% in 2005, 4.75% in 2002, and 9.16% in 2004. The upper end of the distribution was deadweight loss at a value of approximately 30% of free-care spending in 2002 and 2004. The 75th percentile in 2005 was 12.65%.

Previous estimates of moral hazard in health expenditures have focused on pure coinsurance elasticities. (See Newhouse, 1993; Scitovsky and Snyder, 1972; Phelps and Newhouse, 1974; Cherkin et al., 1989;

[19] The deadweight loss in the moral hazard calculation uses the observed charge to the insurer as marginal cost. It is likely that the observed charge to the insurer is different than the true cost of providing the care. This true cost is often difficult to observe even with more detailed provider-level data. If the insurer is making a profit margin above marginal cost on the amount listed in the claims data, then the deadweight loss measure will be an upper bound on moral hazard measured as marginal benefit greater than marginal cost.

Huang and Rosett, 1973) This measure of moral hazard is more nuanced than an elasticity because it takes into account the marginal value of health expenditures the consumer gains, so not all increases in expenditure are moral hazard. To compare this counterfactual deadweight loss to previous estimates, we expect pure elasticity measures of moral hazard be larger than this deadweight loss from generous insurance.

Existing estimates on changing from full insurance to no insurance have been measured both in natural experiments and the RAND Health Insurance Experiment (HIE). Scitovsky and Snyder (1972) used an exogenous change in the coinsurance rate from free care to approximately 25% coinsurance to find that health expenditures decreased by approximately 25%. Similarly, Scheffler (1984) used a pre-post design on the introduction of 40% coinsurance to outpatient care and found that expenditures decreased by approximately 38%. Both of these percent decreases are close to the deadweight loss percentages presented above. Assuming similar patient populations, my estimates should be smaller than the previous studies without marginal utility adjustment. However, my counterfactual also measures a larger price change—going from full insurance to zero insurance. The RAND HIE is a closer match. Keeler and Rolph (1988) found that care almost doubled going from no insurance to full insurance in the RAND HIE. To compare the HIE findings to the deadweight loss above, the HIE found that original expenditures decreased by half when patients had to pay full out-of-pocket costs, and the median deadweight loss of my estimates is 20% of original expenditures. This implies that less than half of the HIE's change in expenditures was inefficient moral hazard.

The deadweight loss estimates above are of interest to policy makers concerned with the transition from first-dollar coverage plans to higher out-of-pocket costs for patients. The results above suggest the upper bound on moral hazard savings are, on average, 20%. These results apply to a population of group-insurance patients with relatively low levels of spending. The results also suggest that the introduction of these high deductible plans had welfare-improving effect for this population in the lower regions of patient spending, though savings seemed to top off at less than approximately 30% of existing spending.

7. Conclusion

This paper addresses two goals. The first goal is to outline a method to measure consumer demand elasticity in the presence of nonlinear pricing by using a nonlinearity to control for unobserved heterogeneity. The second goal is to apply this method in patient-level data and calculate a health expenditure elasticity at two different expenditure points. I then use this elasticity to measure the extent of moral hazard in insurance.

This paper presents an estimation method which isolates consumers of similar unobserved heterogeneity, but who face different prices. The observed distribution of expenditures reveals the underlying distribution of the unobserved heterogeneity. This empirical inverse cdf of expenditures identifies the slope of the relationship between an index of heterogeneity and resulting expenditure choices. The method then calculates this slope on each side of a nonlinear change in marginal prices. The estimation uses a flexible specification of a local linear regression. As the slope of these local linear regressions between expenditures and unobserved heterogeneity approaches the nonlinearity from each side, the underlying unobserved characteristics become similar. Thus, as the limit approaches the nonlinearity point from each side, the difference in the slopes identifies the consumer's response to a change in prices. Applying this method to patient data, the resulting demand elasticity estimates in health insurance data are consistent across all three years of study, at approximately −0.26. The elasticity estimates apply to patient behavior in employer-sponsored insurance over ranges of expenditures less than $1000.

In a counterfactual scenario, this elasticity is used to calculate the compensating transfers of going from full coverage to full out-of-pocket payment. The difference between the compensating transfer and the change in expenditure predicted by elasticities reveals a measure of the deadweight loss of moral hazard. The extent of moral hazard in my data is, on average, 20% of full-coverage expenditures, for expenditures less than approximately $1000. This measure provides insight beyond existing moral hazard measures by netting out positive marginal benefit to the consumer lost by the reduction in health expenditures.

The major contributions of this paper are threefold. First, I create a flexible estimation method that can be used in consumer contract datasets where nonlinear pricing introduces bias. This method presents less restrictive behavioral assumptions and data requirements than previous methods. Second, because nonlinearities are present in many sectors of policy importance, this method offers another tool for estimating elasticities in specific populations or along expenditure distributions where local price changes occur. Finally, I produce an elasticity estimate applicable to patients in employer-sponsored insurance plans, the most common type of insurance coverage in the U.S.

Appendix A

Table A1
Elasticity changing from full out-of-pocket to full coverage.

Estimates at HSA nonlinearity: different kernels				
		Kernel Rule		
Year	Measured at	Epanechnikov	Gaussian	Uniform
2002	$500	−0.33	−0.31	−0.39
2004	$750	−0.26	−0.25	−0.36
2005	$600	−0.25	−0.24	−0.31

Table A2
Elasticity changing from full out-of-pocket to full coverage.

Estimates at deductible nonlinearity: different kernels				
		Kernel Rule		
Year	Measured at	Epanechnikov	Gaussian	Uniform
2002	$1250	−0.09	−0.10	−0.06
2004	$1500	−0.08	−0.08	−0.12
2005	$1500	−0.09	−0.11	−0.11

Table A3
Elasticity changing from full out-of-pocket to full coverage.

Estimates at HSA nonlinearity: Different windows					
		Window size			
Year	Measured at	250	300	350	400
2002	$500	−0.31	−0.33	−0.32	−0.31
2004	$750	−0.29	−0.26	−0.27	−0.26
2005	$600	−0.19	−0.25	−0.28	−0.33

Table A4
Elasticity changing from full out-of-pocket to full coverage.

Estimates at deductible nonlinearity: Different windows					
		Window size			
Year	Measured at	250	300	350	400
2002	$1250	−0.08	−0.09	−0.09	−0.15
2004	$1500	−0.11	−0.08	−0.10	−0.15
2005	$1500	−0.07	−0.09	−0.10	−0.08

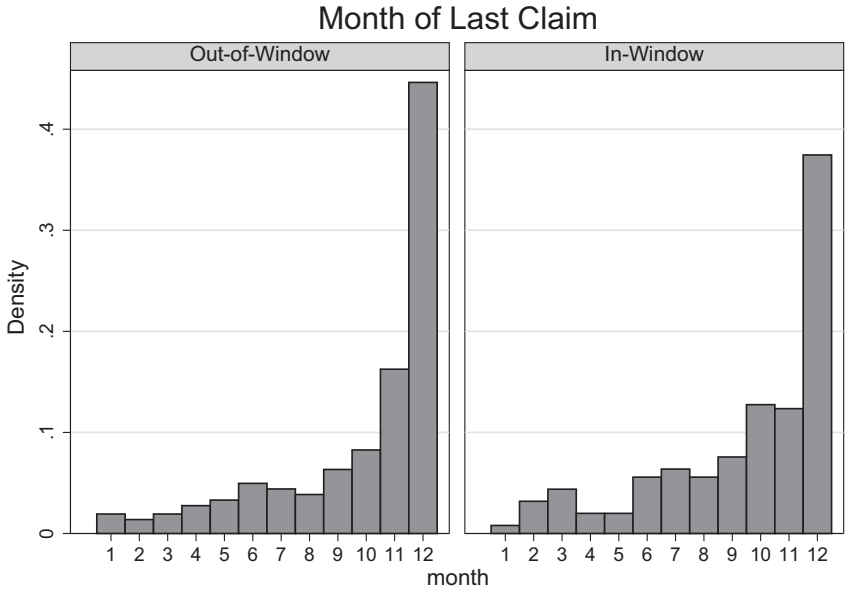

Fig. A1. Month of last claim in-window and out-of-window.

Moral hazard counterfactual

Solving the utility maximization problem listed in Section 6 with full coverage yields an equation for optimal h_{i1} in terms of ρ and γ. I estimate these two parameters by choosing the values that minimize the difference between observed and iterated estimation scenarios. The observed health expenditures are where insurance was full coverage. The iterated estimated values are for $h_{i1}(\rho, \gamma)$ formed by substituting in different values for parameters ρ,γ into the model's utility function that predicts health expenditure choice. The values which minimize the difference between the observed and predicted values of health expenditure choice are optimal $\hat{\rho}, \hat{\gamma}$. Varying number of iterations on a grid search produced very similar estimates for $\hat{\rho}$ and $\hat{\gamma}$. Table A5 lists the parameter estimates for three of these grid searches.

Table A5
Utility parameter estimates.

Grid size	Year	$\hat{\rho}$	$\hat{\gamma}$
30 × 30	2002	0.23	0.28
	2004	0.37	0.31
	2005	0.23	0.31
40 × 40	2002	0.22	0.29
	2004	0.45	0.34
	2005	0.24	0.31
50 × 50	2002	0.20	0.27
	2004	0.30	0.29
	2005	0.23	0.31

Parametric form of general utility model

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

Consider the following general utility function which satisfies utility conditions (1)–(5) from Section 2:

$$\begin{aligned} U(h,\theta,c) &= u(h,\theta)+c \\ &= \gamma\theta+\alpha\theta h^{\beta}-\tau h+c \end{aligned} \tag{A1}$$

Where γ is a parameter on the health shock, α is the parameter on the interaction of the health shock and health expenditures, β is the patient's risk parameter, and τ is the parameter on the time inconvenience from going to the doctor or from utility-reducing levels of health care.

Utility conditions (1), (3), and (5) are clearly satisfied. Condition (2) is satisfied with $h_{\theta}^{max}=\left[\frac{\alpha\beta\theta}{\tau}\right]^{\frac{1}{1-\beta}}$. Condition (4) is satisfied with γ small enough, such that $\gamma<\alpha h^{\beta}$.

The corresponding optimal decision rule is:

$$h^{*}=\left[\frac{\alpha\beta\theta}{1-p+\tau}\right]^{\frac{1}{1-\beta}}. \tag{A2}$$

Recall the reimbursement schedule with a deductible, $\overline{h}$, and resulting MC schedule as described in Section 2:

$$MC=1-p=\begin{cases}1 & \text{if } h\le\overline{h}\\ 0 & \text{if } h>\overline{h}\end{cases}.$$

Given the reimbursement schedule and a health status of θ, the corresponding optimal h^{*} in each marginal cost section is:

$$h^{*}=\begin{cases}\left[\frac{\alpha\beta\theta}{\tau+1}\right]^{\frac{1}{1-\beta}} & \text{if } h^{*}\le\overline{h}\\ \left[\frac{\alpha\beta\theta}{\tau}\right]^{\frac{1}{1-\beta}} & \text{if } h^{*}>\overline{h}\end{cases}. \tag{A3}$$

Notice that the optimal h^{*} that is chosen in the first part of the reimbursement schedule, where all health expenditure must be paid fully out-of-pocket, is smaller than the h^{*} that is chosen in the second part of the reimbursement schedule, where all additional units of expenditure are fully covered by the insurance plan.

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

Eq. (A3) can be transformed into a linear equation using a Taylor approximation around $\bar{\theta}$. The linear approximation is:

$$h = \left[\frac{\alpha\beta\bar{\theta}}{\tau + 1\{\theta \le \bar{\theta}\}}\right]^{\frac{1}{1-\beta}} + \frac{1}{1-\beta}\left[\frac{\alpha\beta\bar{\theta}^{\beta}}{\tau + 1\{\theta \le \bar{\theta}\}}\right]^{\frac{1}{1-\beta}}(\theta - \bar{\theta}) \tag{A4}$$

To construct a local linear regression from Eq. (A4), estimation occurs in the neighborhood of the $\bar{\theta}$ term, with the following coefficients:

$$a(\theta) = \left[\frac{\alpha\beta\bar{\theta}}{\tau + 1\{\theta \le \bar{\theta}\}}\right]^{\frac{1}{1-\beta}}$$

$$b(\theta) = \frac{1}{1-\beta}\left[\frac{\alpha\beta\bar{\theta}^{\beta}}{\tau + 1\{\theta \le \bar{\theta}\}}\right]^{\frac{1}{1-\beta}}.$$

References

AHRQ, 2005. Total health services-mean and median expenses per person with expense and distribution of expenses by source of payment. Briefing, Medical Expenditure Panel Survey Housing Component Data.

Aron-Dine, A., Einav, L., Finkelstein, A., Cullen, M.R., 2012. Moral hazard in health insurance: how important is forward looking behavior? Working Paper 17802. National Bureau of Economic Research, (February).

Aron-Dine, A., Einav, L., Finkelstein, A., 2013. The RAND health insurance experiment, three decades later. J. Econ. Perspect. 27 (1), 197–222 (Winter).

Bajari, P., Hong, H., Park, M., Town, R., 2010. Regression discontinuity designs with an endogenous forcing variable and an application to contracting in health care. NBER Working Paper w17643.

Blomquist, S., Newey, W., 2002. Nonparametric estimation with nonlinear budget sets. Econometrica 70 (6), 2455–2480.

Boes, S., Gerfin, M., 2013. Does full insurance increase the demand for health care? IZA Discussion Papers. Institute for the Study of Labor (IZA).

Cherkin, D., Grothaus, L., Wagner, E., 1989. The effect of office visit copayments on preventative care services in an HMO. Med. Care 27 (7), 669–679.

Chetty, R., Friedman, J.N., Saez, E., 2013. Using differences in knowledge across neighborhoods to uncover the impacts of the EITC on earnings. Am. Econ. Rev. 103 (7), 2683–2721 (December).

Chiappori, P.-A., Durand, F., Geoffard, P.-Y., 1998. Moral hazard and the demand for physician services: first lessons from a French natural experiment. Eur. Econ. Rev. 42.

Diakite, D., Semenov, A., Thomas, A., 2009. A proposal for social pricing of water supply in Cote d'Ivoire. J. Dev. Econ. 88 (2), 258–268.

DiMatteo, L., 2003. The income elasticity of health care spending: a comparison of parametric and nonparametric approaches. Eur. J. Health Econ. 4 (1), 20.

Doyle, J., Almond, D., 2011. After midnight: a regression discontinuity design in length of postpartum hospital stays. Am. Econ. J. Econ. Policy 3 (3), 1–34 (August).

Duarte, F., 2012. Price elasticity of expenditure across health care services. J. Health Econ. 31 (6), 824–841.

Eichner, M., 1998. The demand for medical care: what people pay does matter. American Economic Review Papers and Proceedings of the Hundred and Tenth Annual Meeting of the American Economic Association 88 (2), pp. 117–121 (May).

Einav, L., Finkelstein, A., Ryan, S., Schrimpf, P., Cullen, M., 2013. Selection on moral hazard in health insurance. Am. Econ. Rev. 103 (1), 178–219.

Gary, B., Hausman, J., 1978. The effect of taxation on labor supply: evaluating the Gary negative income tax experiment. J. Polit. Econ. 36, 23–52.

Grubb, M.D., Osborne, M., 2012. Cellular service demand: biased beliefs, learning, and bill shock. American Economic Review Feb 2012.

Hausman, J., 1985. The econometrics of nonlinear budget sets. Econometrica 53, 1255–1282.

Herriges, J., King, K., 1994. Residential demand for electricity under inverted block rates: evidence from a controlled experiment. J. Bus. Econ. Stat. 12 (4), 419–430.

Huang, C.-I., 2008. Estimating demand for cellular phone service under nonlinear pricing. Quant. Mark. Econ. 6 (4), 371–413.

Huang, L., Rosett, R.N., 1973. The effect of health insurance on the demand for medical care. J. Polit. Econ. 81 (2), 281–305 (Part 1).

Janssen, R., 1992. Time prices and the demand for GP services. Soc. Sci. Med. 34 (7), 725–733.

Kaiser Family Foundation, 2006. Employer health benefits: 2006 summary of findings. Annual Report 7528. The Kaiser Family Foundation and Health Research and Education Trust.

Keeler, E., Rolph, J.E., 1988. The demand for episodes of treatment in the health insurance experiment. J. Health Econ. 7, 337–367.

Kowalski, A., 2010. Censored quantile instrumental variable estimates of the price elasticity of expenditure on medical care. Working Paper 15085. National Bureau of Economic Research.

Maddock, R., Castaño, E., Vella, F., 1992. Estimating electricity demand: the cost of linearising the budget constraint. Rev. Econ. Stat. 74 (2), 350–354.

Manning, W.G., Newhouse, J.P., Duan, H., Keeler, E.B., Leibowitz, A., 1987. Health insurance and the demand for medical care: evidence from a randomized experiment. Am. Econ. Rev. 77 (3), 251–277.

Matzkin, R.L., 2003. Nonparametric estimation of nonadditive random functions. Econometrica 71 (5), 1339–1375.

Newhouse, J.P., 1993. Free for All: Lessons from the RAND Health Insurance Experiment. Havard University Press, Cambridge, MA.

Newhouse, J.P., Phelps, C.E., Marquis, M.S., 1980. On having your cake and eating it too: econometric problems in estimating the demand for health services. J. Econ. 13, 365–390.

Phelps, C.E., 1992. Health Economics. HarperCollins, New York.

Phelps, C., Newhouse, J., 1974. Coinsurance, the price of time, and the demand for medical services. Rev. Econ. Stat. 56, 334–342.

Reiss, P., White, M., 2002. Household electricity demand, revisited. Rev. Econ. Stud. 72 (3), 853–883.

Reiss, P., White, M., 2006. Evaluating welfare with nonlinear prices. Working Paper (June).

Saez, E., 2010. Do taxpayers bunch at kink points? Am. Econ. J. Econ. Policy 2 (3), 180–212.

Scheffler, R.M., 1984. The United Mine Worker's Health Plan: an analysis of a cost-sharing program. Med. Care 22 (3), 247–254.

Scitovsky, A.A., McCall, N., 1977. Coinsurance and the demand for physician services: four years later. Soc. Secur. Bull. 40 (5), 1–41.

Scitovsky, A.A., Snyder, N.M., 1972. Effect of coinsurance on use of physician services. Soc. Secur. Bull. 36 (6), 3–19.

Seim, K., Viard, V.B., 2011. The effect of market structure on cellular technology adoption and pricing. Am. Econ. J. Microecon. 3 (2).

Selby, J.V., Fireman, B.H., Swain, B.E., 1996. Effect of a copayment on use of the emergency department in a health maintenance organization. N. Engl. J. Med. 334 (10), 635–642.

Szabo, A., 2010. The value of free water: analyzing South Africa's free basic water policy. Working Paper (September).

van de Ven, W.P., van Praag, B.M., 1981. The demand for deductibles in private health insurance. J. Econ. 17, 229–252.