Estimating Price Sensitivity of Economic Agents Using Discontinuity in Nonlinear Contracts

Patrick Bajari, Amazon.com and NBER

Han Hong, Stanford University

Minjung Park, University of California, Berkeley

Robert Town, University of Pennsylvania and NBER[1]

Abstract

This paper proposes a method to estimate price sensitivity of economic agents using nonlinear contracts. Our proposed estimator exploits discontinuity in nonlinear contracts. As an application, we study contracts between a managed care organization and hospitals for organ transplants. Exploiting "donut holes" in the reimbursement contracts, we show that hospitals submit significantly larger bills when the reimbursement rate is higher, indicating information asymmetries between the insurer and hospitals in this market. Our methodology is applicable to important classes of models such as consumer choice under nonlinear pricing and contracting with nonlinear incentives.

JEL Classifications: C51, C10, I11

[1] Author Affiliations: Patrick Bajari (Amazon.com, patbajari@gmail.com), Han Hong (Stanford University, doubleh@stanford.edu), Minjung Park (University of California, Berkeley, mpark@haas.berkeley.edu), Robert Town (University of Pennsylvania, rtown@wharton.upenn.edu). We thank Aviv Nevo and participants at the Industrial Organization of Health Care Conference at HEC Montreal and the Cowles Structural Microeconomics Summer Conference for helpful comments. All remaining errors are our own.

1 Introduction

Nonlinear pricing is commonly used in a broad array of consumer and business-to-business transactions. In these contexts researchers are often interested in estimating the price responses of participants, but the available data often do not contain the traditional across firm or across time variation to credibly identify them. In this paper, we propose a method to estimate price sensitivity of economic agents using discontinuities in nonlinear contracts. An important issue that arises in such inference is simultaneity: Agents' choices are affected by the marginal price they face, but the marginal price itself is a function of the agents' choice. In this paper, we propose to address such a challenge by using a modified regression discontinuity design (RDD) estimator.

In an RDD, the researcher searches for a forcing variable that shifts the regressor of interest discontinuously at a known cutoff (see Hahn, Todd, and Van der Klaauw, 2001). Although RDD is a commonly used method to estimate treatment effects in a non-experimental setting, RDD identification strategies may fail if the forcing variable can be manipulated by an agent (McCrary, 2008; Lee and Lemieux, 2010; Imbens and Lemieux, 2008; Urquiola and Verhoogen, 2009).

Using the RDD framework to empirically model the choice problem under nonlinear pricing, the simultaneity issue implies that the forcing variable and the outcome variable are identical. For instance, consider our application. Suppose a hospital chooses the optimal level of medical care intensity for a patient given a piecewise linear reimbursement schedule. In order to estimate the responsiveness of the hospital's medical care provision to changes in the reimbursement rates (a long held parameter of interest to health economists), one might exploit discontinuous changes in marginal reimbursement rates to identify this effect. Since marginal reimbursement rates are a function of health care expenditures, health care expenditures would be both the forcing variable and the outcome variable while the reimbursement rate would be the regressor of interest. As the forcing variable can be manipulated by the hospital and is identical to the outcome variable, a standard RDD cannot be applied to identify the price sensitivity of the hospital.

We propose an estimation strategy that can be applied to such a model to recover price sensitivity of economic agents and discuss a set of conditions under which our estimator is consistent. A key idea is that for many choice models, including the one considered in our paper, the optimal solution implies a strictly monotonic relationship between the type of the agent and the agent's choice, except at a point in which different types of agents will behave identically and bunch. We exploit this monotonicity to recast the problem such that the type of the agent is seen as the forcing variable. Our estimator is an RDD-style estimator applied to this reformulated problem, and the key idea behind it is that a discontinuous change in the outcome variable at the cutoff can allow us to infer the incentive effects.

We apply our estimator to understand a fundamental question in health economics – the responsiveness of health care providers to financial incentives. As Arrow (1963) observed, hospitals, physicians and other health care providers possess more information about the appropriateness and necessity of care than the patients or, importantly, their insurer. This fact combined with the likelihood that health care providers are concerned with their own financial well-being implies that first-best contracts may be difficult to implement. Understanding the magnitude of this agency problem is a requisite step to both assessing the welfare consequences of provider agency and designing the optimal contracts in health care settings.

Physicians and hospitals control most of the flow of resources in the health care system and medical care expenditures are a large component of most industrialized countries' GDP – in the US, health care expenditures are currently over 16% of GDP (Congressional Budget Office, 2008). Thus, the welfare gain from better aligning incentives in these contracts with societal objectives is potentially very large. Despite the importance of this issue and the existence of a large theoretical literature (McGuire, 2000), the empirical literature examining the role of the reimbursement contract structure in affecting provider behavior is relatively sparse.[2]

We have collected a unique data set on contracts for organ and tissue transplants between one of the largest US health insurers and all of the hospitals in its network. Organ and tissue transplants are extremely expensive and rare procedures. The infrequency and complexity of the procedures likely lead to information asymmetry between hospitals and insurers, making organ transplants an interesting place to examine provider agency. To the best of our knowledge, no other study in the literature has assembled a panel data set

[2] See Dranove and Wehner (1994) for a discussion of the limitation of the attempts to estimate physician agency. There are important exceptions, however, including: Hodgkin and McGuire (1994), Cutler (1995), Gaynor and Gertler (1995), Gruber and Owings (1996), Yip (1997), Gaynor, Rebitzer and Taylor (2004), Dafny (2005) and Ketcham, Léger and Lucarelli (2011). See Chandra, Cutler and Song (2012) and McClellan (2011) for an excellent review of recent developments in this research stream.

of reimbursement contracts between a major private insurer and hospitals of this scope and detail.

The form of the contracts in our data is fairly simple. As a hospital treats patients, it uses its information system to keep track of all expenses such as drugs and nights in the hospital. Our hospitals have standard "list prices" for each of these expenses. The sum of all of these list prices times the items is referred to as "charges." The contract specifies what fraction of the charges submitted by the hospital for each patient will be reimbursed by the insurer. A key feature of the reimbursement schedules is that the total reimbursement amount for each patient follows a piecewise linear schedule: the marginal reimbursement rate changes discontinuously when certain levels of expenditure are reached. This generates discontinuities in the marginal price received by the hospital for its provision of health care.

We apply our estimator to a discontinuity point in the reimbursement schedule to estimate the sensitivity of health care provision to the reimbursement rate. Our results show that hospitals will submit significantly larger bills if they face a higher reimbursement rate. When the marginal reimbursement rate changes from 0% to 50%, a magnitude of change typically found in our contracts, the marginal increase in hospitals' expenditures for a given increase in patients' illness severity becomes 2 to 14 times larger. These results suggest that hospitals' behavior is strongly influenced by financial incentives.

The methodology characterized here can be applied to other settings in order to address a number of important economic questions. For example, our methodology is directly relevant for examining consumer choice under nonlinear pricing. In this setting, the marginal price of a product is a function of the number of units the consumer purchases. If a researcher is interested in learning how the marginal price affects consumers' purchase decision, the units of purchase would be both the forcing variable and the outcome variable while the marginal price would be the regressor of interest. Under the reasonable (and indeed extremely common) assumption that the marginal utility of consumption is increasing in consumer's type, our proposed method can be applied to this setting. Similarly, our method can be applied to contracting with nonlinear incentives. Nonlinear incentives are often observed in contracts, such as deductibles in insurance products or coverage gaps in Medicare Part D. Under the natural assumption that the marginal benefit of consumption is increasing in the insured's type, our method can be employed in such settings to infer, for instance, how the insured's consumption of health services responds to changes in marginal reimbursement rates.

The literature on estimation of price sensitivity using nonlinear pricing schedules is extensive. Beginning with Burtless and Hausman (1978), many researchers have estimated demand functions when consumers face piecewise linear budget constraints[3] (e.g., Hausman (1979, 1985); Moffitt (1986); Pudney (1989); Reiss and White (2005)). Recently, there has been also some work that estimates workers' sensitivity to incentives using dynamics introduced by nonlinear compensation schemes (Copeland and Monnet (2009); Misra and Nair (2011); Nekipelov (2010)). Most of these papers take a structural approach where they specify utility functions and estimate the parameters using maximum likelihood estimation or generalized method of moments. There is also some work that uses the size of bunching to infer sensitivity of labor supply to marginal tax rates (Saez (2010); Chetty et al. (2011)).[4]

Our paper proposes an alternative approach to the problem of estimating price sensitivity using nonlinear pricing schedules. Since our approach exploits only local variation around a discontinuity point without bunching, while the existing approaches use either the entire schedule or only a discontinuity point with bunching, we view our proposed methodology as complementary to the existing methods. The existing approaches explicitly specify utility functions and estimate structural parameters, while our approach does not rely on particular functional forms of utility functions and only requires the utility functions to satisfy certain properties. As a result, the existing approaches have the advantage of their structural estimates being directly related to the agents' underlying preferences, while our approach has the advantage of relying more on economic restrictions and less on statistical parametric assumptions.

The rest of this paper proceeds as follows. In Section 2, we present a model of hospitals' health care choice. In Section 3, we propose our estimation strategy and discuss its sampling properties. Section 4 describes our data. In Section 5, we present model estimates. Section 6 discusses other settings where our proposed estimator could be employed and Section 7 concludes the paper.

[3] Labor supply decision in the presence of piecewise linear tax schedule is a similar problem.

[4] Note that these approaches do not require variation in prices over time or across agents since nonlinearity in the pricing schedule itself provides natural variation required for identification. The same is true for our proposed methodology. In Section 6, we provide detailed discussions of how our paper relates to these papers.

2 Model

In this section, we will set up a model in order to derive key conditions required for our estimator. For clarity of exposition, we write down a specific model of a hospital's medical care provision decision in an asymmetric information setting to derive these conditions. Later we discuss how our methodology can be applied to other settings such as consumer choice under nonlinear pricing.

2.1 Setup

Consider a health insurer (the principal) that designs compensation contracts for the provider of a medical service (the agent). The insurer's enrolled patients arrive at the hospitals and need potentially costly treatment. Patients differ in their severity of illness that is amenable to medical care which is denoted $\theta \geq 0$, where θ is a random variable with a continuous density function $f(\theta)$ and cdf $F(\theta)$. The health shock, which is determined prior to admission, captures patient heterogeneity in the demand for health care. A central assumption is that patients' heterogeneity is unidimensional, fully captured by θ.[5] The provider then chooses a level of treatment $q \geq 0$. The value of the health outcome to the patient is given by $v(q, \theta)$, which is twice continuously differentiable. The cost of providing treatment at level q is given by $c(q)$. Patients are passive players in this framework.

The agent (the hospital) observes θ and chooses the level of health care q. The principal (the insurer) cannot observe θ but can observe the hospital's choice of q. Hence, the principal cannot directly contract on the optimal level of q, and instead must rely on a compensation scheme to the agent of the general form $r(q)$ in order to implement the desired q.

The cost of treatment is borne by the agent, and $r(q)$ is paid to the agent by the principal. We assume that the agent's net monetary benefits are just $r(q) - c(q)$.[6] Furthermore, we assume that the agent receives a non-pecuniary benefit that is proportional to the patient's payoff. This captures the idea that the agent benefits from successful

[5] In health settings, patients' heterogeneity is likely to be multidimensional. Our analysis assumes that we can well summarize the multidimensional heterogeneity into a single index.

[6] Hospitals typically receive copays from patients as well, but since copays do not depend on q, they would not affect the optimal decision rules of the hospitals.

health outcomes.[7] We also assume quasi-linear utility functions so that there are no income effects. We can write the payoffs of the agent as

$$u^a(q,\theta) = \gamma^a v(q,\theta) - c(q) + r(q).$$

Thus, the agent maximizes $\gamma^a v(q,\theta) - c(q) + r(q)$ and the FOC is (for now, ignoring potential non-differentiability in $r(q)$),

$$\gamma^a \frac{\partial v(q,\theta)}{\partial q} = c'(q) - r'(q). \tag{1}$$

The equality in (1) has a simple economic interpretation: the left hand side is the agent's *marginal benefit* from treatment while the right hand side is her *net marginal cost* (total marginal costs less marginal reimbursement).

2.2 Assumptions

In what follows, we shall assume that θ is uniformly distributed on [0,1]. This assumption is WLOG for a general utility function since we are merely rewriting preferences as a function of the percentile of θ. We shall assume that the payoffs obey the following conditions:

$$\begin{aligned}
\frac{\partial v(q,\theta)}{\partial q} &> 0 && (2)\\
\frac{\partial^2 v(q,\theta)}{\partial^2 q} &< 0 && (3)\\
\frac{\partial v(q,\theta)}{\partial \theta} &< 0 && (4)\\
\frac{\partial^2 v(q,\theta)}{\partial \theta \partial q} &> 0 && (5)\\
\frac{\partial c(q)}{\partial q} &> 0 && (6)\\
\frac{\partial^2 c(q)}{\partial^2 q} &\geq 0 && (7)
\end{aligned}$$

[7] For example, the hospital will value positive patient outcomes if for no other reason than concerns over attracting future patients or deflecting scrutiny by regulators.

Assumptions (2) and (3) state that the value of the health outcome to the patient is increasing and strictly concave in q. Assumption (4) implies that health shocks adversely affect utility. Assumption (5) implies that the value of the health outcome to the patient exhibits strictly increasing differences in (q, θ): the marginal utility of health care increases as agents receive more adverse health shocks. According to assumptions (6) and (7), the cost of providing treatment is an increasing and (weakly) convex function in q.

This structure captures the intuitive idea that (*i*) extra treatments lead to a better health outcome, and the marginal benefit of extra treatments becomes lower as the level of treatment increases; (*ii*) a more severe condition has a higher marginal benefit of extra treatments; and (*iii*) providing more treatment costs more money, and marginal treatments are (weakly) more expensive. As a result, when a patient's condition is more severe she should be offered more treatment.

When the agent consumes q dollars of health care to treat a patient, the agent is reimbursed $r(q)$ by the principal. Since the reimbursement schedule applies to each patient separately, there is no linkage across patients and the agent makes a separate decision for each patient. As we discussed in the introduction, we are interested in situations where the constraint set faced by the agent displays kinks. Reflecting the typical reimbursement schedules used by the health insurer in our data, we shall assume that $r(q)$ satisfies:

$$r(0) = 0 \tag{8}$$

$$r'(q) = \delta_1 \text{ for } 0 < q < q_1 \tag{9}$$

$$r'(q) = 0 \text{ for } q_1 \leq q \leq q_2 \tag{10}$$

$$r'(q) = \delta_2 \text{ for } q > q_2. \tag{11}$$

This assumption implies that the amount of reimbursement for each patient is piecewise linear. For expenditures between 0 and q_1, the hospital is reimbursed δ_1 for every dollar spent to treat the patient. Once expenditures exceed q_1, the hospital hits what is called the donut hole and is forced to bear all of its health care expenses at the margin. Finally, for expenditures above q_2, the hospital is reimbursed δ_2 for every dollar spent. Figure I illustrates a reimbursement scheme implied by assumptions (8)–(11). The region $[q_1, q_2]$ is often referred to as the "donut hole." Donut holes are observed in other health care settings as well, most notably Medicare Part D and high deductible health plans with an attached health savings account.

[Figure I about here]

In our empirical application, our main interest lies in understanding hospitals' behavioral responses to the reimbursement structure, not in understanding what the optimal reimbursement scheme should look like. Although the question of if and why the observed contract differs from the optimal one is a very interesting topic,[8] we abstract away from the optimal contract design problem faced by the principal and just condition on the existence of donut holes to learn about the impact of financial incentives on hospital behavior. We note that in reality we might observe an incentive scheme that departs from the optimal one for various reasons, such as institutional constraints or complexity in implementing the optimal contract.[9]

2.3 Optimal Decision Rule

Under the assumptions written above, the optimal decision rule of an agent who treats a pool of patients exhibits the following features:

1. There will be bunching at q_1.
2. There will be a gap near q_2 and the size of the gap crucially depends on the shape of $u^a(q, \theta)$.
3. The optimal choice of q is strictly increasing in θ except for bunching at q_1.

Figure II illustrates these observations. In drawing the figure, we assume that $0 < \delta_2 < \delta_1 < 1$, which is what we typically observe in the data. The marginal benefit curve for a given level of θ is decreasing in q, and is given by $\gamma^a \frac{\partial v(q,\theta)}{\partial q}$. The lower is γ^a, the flatter are the marginal benefit curves. A higher θ is associated with a marginal benefit curve that is more to the right. The net marginal cost curve is just $c'(q) - r'(q)$. For this

[8] For instance, researchers have argued that optimal health contracts should not have donut holes as they pose excessive risk and there are better ways of dealing with moral hazard (Rosenthal, 2004).

[9] In case of Medicare Part D, a donut hole was introduced due to limited government budget available for the program.

figure, we assume that $c'(q)$ is constant, which is not crucial for any of our results but simplifies the graphical analysis.

[Figure II about here]

Imagine a level of θ that corresponds to an optimal choice below q_1. As θ increases, the optimal choice will also increase until some level θ_1 at which it will be exactly q_1. Given the kink in the incentive scheme, there is a jump in the net marginal cost curve, causing bunching at q_1 for levels higher than θ_1. At some point, however, high enough levels of θ above θ_1 will cause the marginal benefit curve to shift enough so that optimal choices will exceed q_1 and be on the part of the net marginal cost curve that is $c'(q)$ (i.e., $r'(q) = 0$). The choice of q then continues to rise monotonically with θ until we hit a gap in choices just around q_2, where the net marginal cost drops. To see why we have a gap, consider the level θ^* that is depicted in Figure II. For this level of severity the agent is indifferent between choosing two levels of health care, one strictly below q_2 (say q_L) and another strictly above (say q_H).[10] By the monotonicity of $q(\theta)$ which follows from the assumption of increasing differences in (q, θ), there will not be any choices of treatment that correspond to expenditures within the interval (q_L, q_H). Finally, for all $\theta > \theta^*$, $q(\theta)$ is strictly increasing.

In Figure III, an agent with a higher level of γ^a is depicted. The marginal benefit curves of this agent will be shifted up and right compared to those in Figure II, and they become steeper (a consequence of $v(q, \theta)$ being multiplied by γ^a). This implies that all choices will be shifted to the right (higher levels of q for any given θ), and the steepness of the marginal benefit curves implies that the gap will be small. In this particular example, the size of the gap $[q_L, q_H]$ is almost negligible. The graphical analysis of Figures II and III offers a complete treatment of what the agent's behavior would be in face of a kinked incentive scheme as described in Figure I.

[Figure III about here]

Throughout our discussion, we have assumed that the agent cannot "cheat" and fraudulently announce costs that were not incurred. If this can happen, then we might observe

[10] q_L and q_H are functions of economic primitives – the reimbursement rates, patients' utility function, how much the hospital values patients' health outcomes (γ^a) and the cost function.

patterns that are not implied by the optimal decision rule. Although such a fraudulent reporting is not impossible, we think it is uncommon among the hospitals in our data because they are large, established hospitals that are subject to regular audits.

The above decision rules, in particular decision rules 2 and 3 (the presence of a discontinuity point without bunching and strict monotonicity of choice in type except at bunching), are necessary for development of our estimator in the next section. Although we derived the optimal decision rules from the particular model on hospital health care provision, they in fact hold under quite a general class of models.[11] For instance, in models of consumer demand under nonlinear pricing schedules, θ will represent people's willingness to consume goods, q will represent the units of purchase by a consumer, and a piecewise linear pricing schedule will generate discontinuity points like q_1 and q_2. To generate a discontinuity point where bunching does not occur (q_2 in the above figure), we need the marginal price to have a sudden drop at least once in the pricing schedule, which is indeed common in practice, e.g. volume discounts. We will have strict monotonicity of choice q in type θ except at bunching under the assumption of single-crossing property (similar to the assumption of increasing differences in the above model), which is an extremely common assumption in consumer choice models.

We will also get similar decision rules in models of contracting with nonlinear incentives such as insurance products with deductibles. In those settings, θ will represent the insured's risk type, q will represent the amount of claims filed by the insured, and the deductibles will generate discontinuity points like q_2 (deductibles imply a sudden drop in the marginal costs faced by the insured once a certain threshold is reached). Again, we will have strict monotonicity of choice in type under the assumption of single-crossing property.

3 Estimation

In this section we propose an estimator that will yield consistent estimates of the agent's behavioral responses exploiting discontinuity in nonlinear schedules. We first discuss the

[11] In particular, a principal-agent relationship is not necessary in deriving the optimal decision rules. Our model on hospitals' health care provision happened to have such a feature, but we will get similar optimal decision rules in models that do not have principal-agent relationships, e.g., models where consumers know their own types and choose the optimal level of consumption under a nonlinear budget constraint.

key intuition behind our approach and then outline our estimation procedures.

3.1 Using Discontinuous Changes for Identification

At the two discontinuity points q_1 and q_2, the marginal reimbursement rate faced by the hospital changes discontinuously. These discontinuities seem to present a natural setting for an RDD. This canonical choice model, however, differs significantly from typical RDD settings because q is both the forcing variable (the level of q determines the marginal reimbursement rate) and the dependent variable (our goal is to estimate how the level of q responds to the marginal reimbursement rate).

In this paper, we propose a method that can be applied to such a setting. A key step in our approach is to transform the problem so that we make the type of the patient θ a forcing variable. From the earlier discussion, and more generally the monotone comparative statics literature of Topkis (1978) and Milgrom and Shannon (1994), we know that the assumption of strictly increasing differences in (θ, q) implies that the optimal health care provision q is a strictly increasing function of patient type θ, with the exception of where there is bunching at q_1. As a result, the percentiles of q will identify θ. That is, if we see a patient with the 5^{th} percentile of health expenditure within a hospital, that patient will have the 5^{th} percentile of health shock within that hospital. This means that for all practical purposes, the types are observable to the econometrician. Since q is only weakly increasing in θ around the first discontinuity point due to the presence of bunching, the econometrician cannot infer θ from the cdf of q in the region. Hence, our estimation procedure can be applied to a discontinuity point without bunching, but not a discontinuity point with bunching.

Once we reformulate the problem so that the patient type θ is viewed as a forcing variable (which is exogenously endowed and cannot be manipulated), a shift in the patient type θ determines whether the hospital's choice of q for that patient will be on the left hand side or right hand side of the second discontinuity point. This then generates an exogenous change in the marginal price faced by the hospital, allowing for identification of the hospital's response to incentives. Our approach boils down to estimating a variant of regression discontinuity models in the empirical quantile function of the hospital's choice q.

Figure IV illustrates the idea behind our approach. Among patients who come to the hospital with a realization of health shock, there will be a value of θ at which the hospital is indifferent between choosing q_L ($< q_2$) and q_H ($> q_2$). Let θ^* denote the level of severity which leads to such an indifference. Then for all patients whose θ is greater than θ^*, the hospital will choose q larger than q_H and will face a marginal reimbursement rate of δ_2. For patients whose θ is smaller than θ^*, the hospital will choose q smaller than q_L and will face a marginal reimbursement rate of 0. Thus, the hospital's supply of health care services will be more responsive to an increase in θ on the right hand side of θ^* than on the left hand side of θ^* as long as the hospital is price sensitive. Therefore, by comparing how quickly q rises with an increase in θ for values of θ just below and just above θ^*, we can infer how the total claims filed by the hospital depend on the reimbursement structure. We let ϕ_{SLOPE} denote the change in the slope of the quantile function, $q'(\theta)$, at θ^*. Note that the slope of the quantile function $q'(\theta)$ is equal to the inverse of density. We will interpret a positive value of ϕ_{SLOPE} as evidence that hospitals' health care provision is influenced by financial incentives embodied in the reimbursement schedule. Furthermore, the same logic implies that ϕ_{SLOPE} would be greater for a larger jump in the marginal reimbursement rate at q_2, all else equal. The figure also shows a possibility of gap $\phi_{GAP} = q_H - q_L$ at θ^*. As discussed earlier, for a high value of γ^a the gap could be very small, while for a low value of γ^a the gap could be large.

We note that a difference in densities between the two sides of θ^* could come from two sources. First there could be a gap at the threshold, leading to different quantities q_L and q_H on the two sides of the threshold. Second, the reimbursement rates differ between the two sides. The densities on the two sides could differ due to a gap in q (we measure densities at different quantities q_L and q_H on the two sides) and/or due to the different reimbursement rates. The marginal rate of substitution between q and θ in $v(q, \theta)$ is likely to be different at q_H and q_L when the gap is present. However, a gap, if any, is generated in the first place because of changes in reimbursement rates, so different densities on two sides due to a gap could be also interpreted as a result of incentive change. In this paper, we do not explicitly distinguish between these two sources – how much of the difference in density is due to a gap (in some sense indirect effect of different reimbursement rates) vs. direct effect of different reimbursement rates – because doing so would require more specific assumptions on the shape of $v(q, \theta)$ and $c(q)$.

[Figure IV about here]

It is worthwhile to note the key underlying assumptions in our approach. A key identifying assumption is that the slope of the quantile function would be the same at θ^* from both sides if there were no change in the marginal reimbursement rate at θ^*. In other words, the density of q would be continuous at θ^* in the absence of a discrete change in incentives at θ^*. This assumption allows us to attribute any discrete change in the slope of the quantile function at θ^* to a discrete change in the financial incentive. A similar type of continuity assumption is found in the conventional RDD (Hahn, Todd, and Van der Klaauw, 2001) and in McCrary (2008) who proposes a test for manipulation of a forcing variable by examining discontinuity in the density function of the forcing variable at the cutoff point.

In order to correctly map q to θ, we also require that there be no error in agents' choice of $q(\theta)$. If $q(\theta)$ contains error in it, we cannot infer type θ from the observed q.[12] In reality, $q(\theta)$ is very likely to have error in it since hospitals cannot perfectly control the level of treatment: there is lumpiness in treatments, there could be some unforeseen events that make it more costly to treat a less sick patient, etc.[13] Since such error would lead to measurement error in inferred θ, we expect that our estimator might suffer from a downward bias. Intuitively, when θ is measured with error, the difference between the slopes of the quantile function on the LHS and RHS of the threshold will be smoothed out, leading to a downward bias. In the extreme case, if q is determined entirely randomly without any relation to patient sickness and the hospitals have no control over q, we would not observe any difference in the slopes of the quantile function between the small neighborhoods on the two sides of θ^*.

3.2 Estimation

We consider several different estimation approaches that deal with different levels of hospital heterogeneities. The first method applies to individual hospital data. The second method makes use of a global parametric assumption to pool information from data across all hospitals.[14] The first method is more robust since it doesn't rely on parametric as-

[12] We will face a similar challenge if there exist other sources of heterogeneity (other than θ).

[13] We note that such a problem is likely to be less severe in typical consumer choice settings where consumers directly choose the optimal level of consumption under a nonlinear budget constraint (those settings also tend to have less uncertainty).

[14] In typical consumer choice settings, the first method applies to individual market data (and there are many consumers in each market, as there are many patients in each hospital in our application). The

sumptions, but will require a large amount of data per hospital. The second method depends on validity of the parametric assumptions, but doesn't require as much data per hospital. Which method is more appropriate will depend on specific applications.

3.2.1 Individual Hospital Estimates

Suppose that there are $i = 1, ..., n$ individuals treated in the hospital under consideration. Let q_i denote the health expenditure of individual i. Let $\hat{F}(\cdot)$ denote the empirical distribution of the observed q's for the hospital. We propose estimators for ϕ_{GAP} and ϕ_{SLOPE} at the upper regression discontinuity point of q_2 and derive the asymptotic distribution of the estimators.

The incentive scheme is such that for θ approaching a cutoff value θ^* from the left, $q(\theta)$ approaches a limit value q_L. As soon as θ moves to the right of θ^*, $q(\theta)$ takes a discrete jump at the point of θ^* by an amount $\phi_{GAP} > 0$ to q_H.

By normalization, θ is estimated as the empirical CDF of the observed q. Hence θ^* is estimated by

$$\hat{\theta}^* = \hat{F}(q_2) = \frac{1}{n}\sum_{i=1}^{n} 1(q_i \leq q_2),$$

where $\hat{F}(\cdot)$ is the empirical distribution of the observed q's. Given that we define $\theta^* = F(q_2)$ where $F(\cdot)$ is the true distribution function of q, the asymptotic distribution of $\hat{\theta}^*$ is immediate:

$$\sqrt{n}(\hat{\theta}^* - \theta^*) \xrightarrow{d} N\left(0, F(q_2)(1 - F(q_2))\right).$$

We are interested in estimating the magnitude of the discontinuity ϕ_{GAP}. This is estimated by

$$\hat{\phi}_{GAP} = \hat{q}_H - \hat{q}_L = \min\{q_i : q_i > q_2\} - \max\{q_i : q_i \leq q_2\}.$$

The goal is to derive the asymptotic distribution of $\hat{\phi}_{GAP} - \phi_{GAP}$.[15] It suffices to

second method will pool information from multiple markets.

[15] We note that estimation of a gap could be sensitive to outliers. For instance, the estimates may be off if the hospital behaves suboptimally even for a small fraction of patients. This is inherent in estimators based on order statistics. Below we will discuss an alternative estimation method that is more robust against outliers.

show that the joint distribution of $n(\hat{q}_L - q_L)$ and $n(\hat{q}_H - q_H)$ are independent exponential distributions. To see this, note that

$$\begin{aligned} P(n(\hat{q}_L - q_L) &\leq -x, n(\hat{q}_H - q_H) \geq y) \\ &= P(q_i \leq q_L - x/n, q_i \geq q_H + y/n, \forall i) \\ &= (1 - P(q_L - x/n \leq q_i \leq q_H + y/n))^n \\ &= \left(1 - f^- x/n - f^+ y/n + o(1)/n\right)^n \overset{n\to\infty}{\longrightarrow} e^{-f^- x - f^+ y}. \end{aligned}$$

In other words, $n(\hat{q}_L - q_L)$ and $n(\hat{q}_H - q_H)$ converge to two independent (negative and positive) exponential random variables with hazard rates $f^- = f(q_L)$ and $f^+ = f(q_H)$, where we have used f^- and f^+ to denote the (left and right) densities at q_L and q_H. The limiting distribution of $n(\hat{\phi}_{GAP} - \phi_{GAP})$ is therefore the sum of two independent exponential random variables.

Next we turn to the estimation of difference between the slopes of $q(\theta)$ at q_H and q_L, defined as $\phi_{SLOPE} = \lim_{\theta \to \theta^*_+} q'(\theta) - \lim_{\theta \to \theta^*_-} q'(\theta)$. Note that $\phi_{SLOPE} = \frac{1}{f^+} - \frac{1}{f^-}$. Hence it suffices to obtain consistent nonparametric estimators for f^+ and f^-. This can be done using standard one sided kernel smoothing methods.

Define

$$\hat{f^-} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k(\frac{\hat{q}_L - q_i}{h}) 1(q_i \leq \hat{q}_L),$$

and

$$\hat{f^+} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k(\frac{q_i - \hat{q}_H}{h}) 1(q_i \geq \hat{q}_H).$$

In the above, $k(\cdot)$ is a one-sided density function supported on $(0, \infty)$, and h is a sequence of bandwidth parameters used in typical kernel smoothing. It is straightforward to show that as long as $nh \longrightarrow \infty$ and $nh^3 \longrightarrow 0$,

$$\sqrt{nh}\left(\hat{f^-} - f^-\right) \xrightarrow{d} N\left(0, f^- \int k(u)^2 du\right),$$

and

$$\sqrt{nh}\left(\hat{f^+} - f^+\right) \xrightarrow{d} N\left(0, f^+ \int k(u)^2 du\right)$$

and that they are asymptotically independent. Therefore

$$\sqrt{nh}\left(\hat{\phi}_{SLOPE}-\phi_{SLOPE}\right)\xrightarrow{d}N\left(0,\frac{1}{f^{-3}}\int k(u)^2du+\frac{1}{f^{+3}}\int k(u)^2du\right).$$

3.2.2 A Parametric Model Using Multiple Hospital Data

Now we consider how to extend the previous method to allow for pooling information from data across multiple hospitals. There are a few alternative estimation approaches one can potentially take. In this section we provide an overview. Our estimators can be extended to incorporate heterogeneous data from all hospitals.

Consider first $\phi_{GAP}=q_H-q_L$. We define $y_i=q_i1(q_i\le q_2)$ and $z_i=q_i1(q_i>q_2)+M1(q_i\le q_2)$. In the homogeneous case, we have defined $\hat{q}_L=\max\{y_i\}$ and $\hat{q}_H=\min\{z_i\}$, where M is a number that is larger than any of the data points.

With cross-hospital data, the observed threshold value q_2 can be hospital dependent, which we will denote as $q_2(t)$, where we have used t to index hospitals. Suppose hospital heterogeneity is captured by covariates x_t, where x_t can include $q_2(t)$ itself. Let I_t be the number of patient observations for hospital t. We specify the following parametric assumption that

$$q_L(t)\equiv q_L(x_t)=g_L(x_t,\beta_L)\quad\text{and}\quad q_H(t)\equiv q_H(x_t)=g_H(x_t,\beta_H).$$

In the above, we can use a flexible series expansion functional form of $g_L(x_t,\beta_L)$ and $g_H(x_t,\beta_H)$ so that they are linear in the parameters β_L and β_H. The structure of this problem fits into the boundary parameter estimation method studied in the literature. Possible estimators include the linear programming approach, the extreme quantile regression approach of Chernozhukov (2005) and nonstandard likelihood estimator (c.f. Donald and Paarsch, 1996; Chernozhukov and Hong, 2004). We describe these alternatives in the following.

The linear (or quadratic, etc.) programming approach estimates the parameters by

$$\begin{aligned}\hat{\beta}_L\quad&=\quad\arg\min_{\beta_L}\sum_{t=1}^{T}I_t^Lg_L(x_t,\beta_L),\quad\text{where}\quad I_t^L=\sum_{i=1}^{I_t}1(y_{it}>0),\\ \text{such that}\quad&y_{it}\le g_L(x_t,\beta_L),\forall i=1,...,I_t^L,t=1,...,T,\end{aligned}$$

and

$$\begin{aligned}\hat{\beta}_H \quad &= \quad \arg\max_{\beta_H} \sum_{t=1}^{T} I_t^H g_H(x_t, \beta_H), \quad \text{where} \quad I_t^H = \sum_{i=1}^{I_t} 1(z_{it} < M),\\ \text{such that} \quad & z_{it} \geq g_H(x_t, \beta_H), \forall i = 1, ..., I_t^H, t = 1, ..., T.\end{aligned}$$

The objective functions $\sum_{t=1}^{T} I_t^L g_L(x_t, \beta_L)$ and $\sum_{t=1}^{T} I_t^H g_H(x_t, \beta_H)$ can be replaced by

$$\sum_{t=1}^{T}\sum_{i=1}^{I_t} 1(y_{it} > 0)(y_{it} - g_L(x_t, \beta_L))^2 \text{ and } \sum_{t=1}^{T}\sum_{i=1}^{I_t} 1(z_{it} < M)(z_{it} - g_H(x_t, \beta_H))^2$$

or other types of penalization functions. The linear programming approach however seems to be the easiest to implement.

Alternatively, β_L and β_H can be estimated by the extreme quantile regression method of Chernozhukov (2005):

$$\hat{\beta}_L = \arg\min_{\beta_L} \sum_{t=1}^{T}\sum_{i=1}^{I_t} 1(y_{it} > 0)\rho_{\tau_L}(y_{it} - g_L(x_t, \beta_L)),$$

where $\rho_\tau(u) = (\tau - 1(u \leq 0))u$ is the *check function* of Koenker and Bassett (1978), such that

$$\tau_L \to 1 \quad \text{as} \quad n_L = \sum_t I_t^L \to \infty.$$

Similarly, $\hat{\beta}_H = \arg\min_{\beta_H} \sum_{t=1}^{T}\sum_{i=1}^{I_t} 1(z_{it} < M)\rho_{\tau_H}(z_{it} - g_H(x_t, \beta_H))$, where

$$\tau_H \to 0 \quad \text{as} \quad n_H = \sum_t I_t^H \to \infty.$$

The quantile regression approach has the advantage of being robust against a certain fraction of outliers in the data. On the other hand, the programming estimators always satisfy the constraints of the relation between y_{it} and $g_L(x_t, \beta_L)$ and between z_{it} and $g_H(x_t, \beta_H)$.

By adopting a parametric functional form on $q_L(x_t)$ and $q_H(x_t)$ we are maintaining a strong specification assumption which can potentially be tested by the data. An implicit assumption of the parametric functional form is that $g_L(x_t, \beta_L^0) \leq q_2(t) \leq g_H(x_t, \beta_H^0)$ for all t at the true parameters β_L^0 and β_H^0. Of course their estimates introduce sampling noise, but we still expect that it should be largely true for most t:

$$g_L(x_t, \hat{\beta}_L) \leq q_2(t) \leq g_H(x_t, \hat{\beta}_H).$$

The approximate validity of this condition can be used as the basis of a model specification test.

Then $\phi_{GAP}(x_t)$ will be estimated consistently by

$$\hat{\phi}_{GAP}(x_t) = g_H(x_t, \hat{\beta}_H) - g_L(x_t, \hat{\beta}_L).$$

Conducting statistical inference on $\hat{\phi}_{GAP}(x_t)$ requires the limiting *joint* distribution of $\hat{\beta}_L$ and $\hat{\beta}_H$. They converge to a nonstandard distribution at a fast $1/n$ rate for $n = \sum_t I_t$.[16] The limiting distribution can be obtained by simulation which we will describe below in the context of the parametric likelihood approach.

If we are interested in the shape of the distribution of q_{it}, specifically the slope parameter $\hat{\phi}_{SLOPE}(x_t)$, in addition to q_L and q_H, we can adopt a maximum likelihood approach. To this end, assume that

$$\epsilon_{it}^L = g_L(x_t, \beta_L) - y_{it} \sim f_L(\epsilon_{it}^L, x_t, \alpha_L) \text{ for } y_{it} \leq g_L(x_t, \beta_L),$$

and

$$\epsilon_{it}^H = z_{it} - g_H(x_t, \beta_H) \sim f_H(\epsilon_{it}^H, x_t, \alpha_H) \text{ for } z_{it} \geq g_H(x_t, \beta_H) \text{ and } z_{it} < M.$$

The maximum likelihood estimator for α_L, α_H and β_L, β_H can then be written as

$$\begin{aligned}(\hat{\alpha}_L, \hat{\beta}_L) \quad &= \quad \arg\max_{\alpha_L, \beta_L} \sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(y_{it} > 0) \log f_L(g_L(x_t, \beta_L) - y_{it}, x_t, \alpha_L) \\ &\qquad \text{such that} \quad y_{it} \leq g_L(x_t, \beta_L), \forall i = 1, ..., I_t, t = 1, ..., T,\end{aligned}$$

[16] Asymptotics is in the number of hospitals.

and

$$\begin{aligned}(\hat{\alpha}_H, \hat{\beta}_H) &= \arg\max_{\alpha_H, \beta_H} \sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(z_{it} < M) \log f_H(z_{it} - g_H(x_t, \beta_H), x_t, \alpha_H) \\ &\quad \text{such that} \quad z_{it} \geq g_H(x_t, \beta_H), \forall i = 1, ..., I_t, t = 1, ..., T.\end{aligned}$$

In fact the linear programming estimator is a special case of the above maximum likelihood estimator when the densities $f_L(\epsilon_{it}^L, x_t, \alpha_L)$ and $f_H(\epsilon_{it}^H, x_t, \alpha_H)$ are exponential distribution with a homogeneous hazard rate parameter: $f(\epsilon) = \lambda e^{-\lambda\epsilon}$. In this case, in addition to obtaining $\hat{\beta}_L$ and $\hat{\beta}_H$ from the linear programming estimators, we also estimate the hazard parameters by

$$1/\hat{\lambda}_L = \frac{\sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(y_{it} > 0) \left(g_L(x_t, \hat{\beta}_L) - y_{it}\right)}{\sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(y_{it} > 0)}$$

and

$$1/\hat{\lambda}_H = \frac{\sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(z_{it} < M) \left(z_{it} - g_H(x_t, \hat{\beta}_H)\right)}{\sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(z_{it} < M)}$$

Even though $\hat{\beta}_L$ and $\hat{\beta}_H$ converge at $1/n$ rate to a nonstandard limit distribution, $\hat{\alpha}_L$ and $\hat{\alpha}_H$ are still root n consistent and asymptotically normal, as long as there are no functional relations between α and β.

To estimate $\phi_{SLOPE}(x_t)$, we can use

$$\hat{\phi}_{SLOPE}(x_t) = \frac{1}{f_H(0, x_t, \hat{\alpha}_H)} - \frac{1}{f_L(0, x_t, \hat{\alpha}_L)}.$$

Since $\hat{\phi}_{SLOPE}(x_t)$ is root n consistent and asymptotically normal, its limiting distribution can be obtained by the delta method combined with the standard sandwich formula, or by simulation or bootstrap, in which $\hat{\beta}_L$ and $\hat{\beta}_H$ can be held fixed because they do not affect the asymptotic distribution.

The joint asymptotic distribution for $\hat{\beta}_L$ and $\hat{\beta}_H$ can be obtained by parametric simulations. Given the assumption that the parametric model is correctly specified, it is possible to simulate from the model using the estimated parameters $\hat{\beta}_L$, $\hat{\beta}_H$, $\hat{\alpha}_L$ and $\hat{\alpha}_H$. The approximate distribution can be obtained from repeated simulations. Instead of recomputing the maximum likelihood estimator at each simulation, it suffices to recompute weighted programming estimators of β_L and β_H at each simulation:

$$\begin{aligned}\tilde{\beta}_L &= \arg\min_{\beta_L} \sum_{t=1}^{T}\sum_{i=1}^{I_t} f_L(0, x_t, \hat{\alpha}_L)\frac{\partial g_L(x_t, \hat{\beta}_L)'}{\partial \beta_L}\beta_L 1(y_{it} > 0) \\ &\quad \text{such that} \quad y_{it} \leq g_L(x_t, \beta_L) \;\; \forall i, t,\end{aligned}$$

and

$$\begin{aligned}\tilde{\beta}_H &= \arg\max_{\beta_H} \sum_{t=1}^{T}\sum_{i=1}^{I_t} f_H(0, x_t, \hat{\alpha}_H)\frac{\partial g_H(x_t, \hat{\beta}_H)'}{\partial \beta_H}\beta_H 1(z_{it} < M) \\ &\quad \text{such that} \quad z_{it} \geq g_H(x_t, \beta_H) \;\; \forall i, t.\end{aligned}$$

with the understanding that now all the "data" x_t, y_{it}, z_{it} and M are specific to each simulation draw.

We can also consider the possibility that $g_L(x_t, \beta_L)$ and $g_H(x_t, \beta_H)$ are correctly specified but $f_L(\epsilon_{it}^L, x_t, \alpha_L)$ and $f_H(\epsilon_{it}^H, x_t, \alpha_H)$ are misspecified. In this case, each of the above methods (linear and quadratic programmings, extreme quantile regression, (pseudo) maximum likelihood estimation) will still deliver consistent estimates of β_L and β_H and hence ϕ_{GAP}. But the estimates for α_L, α_H and hence ϕ_{SLOPE} are clearly inconsistent.

In this case, if we are willing to impose parametric assumptions on ϕ_{GAP} through $g_L(x_t, \beta_L)$ and $g_H(x_t, \beta_H)$, but are not willing to make parametric assumptions on ϕ_{SLOPE}, we can estimate ϕ_{SLOPE} using nonparametric density estimators. We can also use nonparametric density estimators to perform semiparametric simulations for consistent inference about $\hat{\phi}_{SLOPE}$. Suppose x_t is continuously distributed with dimension d. Let

$$\hat{f}^-(x) = \sum_{t=1}^{T}\sum_{i=1}^{I_t}\frac{1}{h}w(x_t, x)k(\frac{g_L(x_t, \hat{\beta}_L) - q_{it}}{h})1(q_{it} \leq g_L(x_t, \hat{\beta}_L)),$$

and

$$\hat{f}^+(x) = \sum_{t=1}^{T}\sum_{i=1}^{I_t}\frac{1}{h}w(x_t, x)k(\frac{q_{it} - g_H(x_t, \hat{\beta}_H)}{h})1(q_{it} \geq g_H(x_t, \hat{\beta}_H)),$$

where

$$w(x_t, x) = k^d(\frac{x_t - x}{h}) / \sum_{t=1}^{T}\sum_{i=1}^{I_t} k^d(\frac{x_t - x}{h}).$$

$k^d(\cdot)$ is a d-dimension two sided symmetric kernel function for $d = dim(x)$. Then we can form the estimate $\hat{\phi}_{SLOPE}(x_t) = 1/\hat{f}^{+}(x_t) - 1/\hat{f}^{-}(x_t)$.

The limiting distribution of the MLE's $\hat{\beta}_L$ and $\hat{\beta}_H$ in this case can be obtained by recomputing the following weighted programming estimators with simulated data:

$$\begin{aligned} \bar{\beta}_L &= \arg\min_{\beta_L} \sum_{t=1}^{T} \sum_{i=1}^{I_t} \hat{f}^{-}(x_t) \frac{\partial g_L(x_t, \hat{\beta}_L)'}{\partial \beta_L} \beta_L 1(y_{it} > 0) \\ &\quad \text{such that} \quad y_{it} \leq g_L(x_t, \beta_L) \ \ \forall i, t, \end{aligned}$$

and

$$\begin{aligned} \bar{\beta}_H &= \arg\max_{\beta_H} \sum_{t=1}^{T} \sum_{i=1}^{I_t} \hat{f}^{+}(x_t) \frac{\partial g_H(x_t, \hat{\beta}_H)'}{\partial \beta_H} \beta_H 1(z_{it} < M) \\ &\quad \text{such that} \quad z_{it} \geq g_H(x_t, \beta_H) \ \ \forall i, t. \end{aligned}$$

As before, the simulated distributions of $n(\bar{\beta}_L - \hat{\beta}_L)$ and $n(\bar{\beta}_H - \hat{\beta}_H)$ should approximate the limit distributions of the maximum likelihood estimates $n(\hat{\beta}_L - \beta_L^0)$ and $n(\hat{\beta}_H - \beta_H^0)$.

4 Application: Data and Setting

Our empirical application is contracting between a private health insurer and hospitals for the procurement of organ and tissue transplants. According to our data source, because of the complexity and patient heterogeneity in transplant care there is significant information asymmetry between the insurer and the hospitals leaving ample opportunity for provider agency. Consistent with our modeling framework, the majority of these contracts are piecewise linear with multiple kink points.

We have acquired detailed hospital contracting data from the largest private insurer of organ and tissue transplants in the US. The insurer contracts with 127 hospitals in the US and we have data on the shape of the reimbursement schedule for all of these contracts. The data span 2004 to 2007. The insurer negotiates different contracts for each organ and therefore our contract information is at the year/hospital/organ level. Typically hospitals renegotiate their contracts with the insurer every three or four years.

In addition to the contract information, we also have detailed, administrative claims-level information for each transplant the insurer covered over this period. Linking these two data sets yields an analytic data set that has (*i*) claim-level information, such as the admission and discharge dates of the patient, the type of transplant received by the patient, the size of the bill submitted by the hospital to the insurer and the reimbursement amount paid by the insurer, as well as (*ii*) hospital-level information, such as the name and location of the hospital and the reimbursement schedule the hospital faces for each type of organ transplant surgery it performs.

The contracts cover the major organ and tissue transplants, the most common being bone marrow transplant (BMT), kidney transplant, liver transplant, heart transplant and lung transplant. Organ and tissue transplants are a rare and exceedingly expensive procedure. In 2007, 27,578 organs were transplanted in the US. The average total billed charges for kidney transplantation in our data, the least expensive and most commonly transplanted organ, exceed $140,000. An organ transplant is an extremely challenging and complex procedure taking anywhere from 3 (kidney) to 14 hours (liver). Organ transplants usually require significant post-operative care (up to 3 weeks of inpatient care) and careful medical management to prevent rejection. The infrequency of the procedures, the complexity of the treatments and the large variation across patients in their response to transplantation make it difficult for the insurer to determine the appropriateness of the care for a given episode. That, in turn, implies that hospitals are in a position to engage in agency behavior in response to the incentives embodied in their contracts.

The insurer in our data is the largest private payer for organ transplants (80% market share among private payers), but is smaller than Medicare.[17] Private insurers pay for approximately 40% of kidney and 50% of all other organ/tissue transplants (OPTN / SRTR, 2007). Typically, the reimbursements made by the payer we study will comprise a significant portion of the transplant revenue of a transplant hospital. There are a total of 270 kidney transplant centers (and much fewer transplant centers for the other organs) in the US. According to our conversations with the payer, they contract with the hospitals that they believe provide the highest quality care while also maintaining a network that has a broad geographic coverage.

The insurer negotiates a separate contract with each individual hospital for each major organ type. As a result, the reimbursement schedule differs substantially across hospitals. For about 75% of hospitals in our original sample, the reimbursement schedule takes a form

[17] In addition, Medicaid is a significant payer for pediatric transplants.

as shown in Figure I with two kinks (the marginal reimbursement rate starts at positive, becomes zero for a certain range and then becomes positive again). The remaining 25% of hospitals have contracts that have only one kink (the marginal reimbursement rate starts at positive and then remains at zero above a certain expenditure level). Under the second type of contract, the maximum amount of reimbursement is capped at a fixed level, while the maximum reimbursement increases with billed charges under the first type of contract. As a result, hospitals are exposed to greater risk under the second type of contract. Even among hospitals that have the first type of contract, there is a large variation in the locations of the first kink (q_1) and the second kink (q_2), the marginal reimbursement rate for each of the segments (δ_1 and δ_2) and the height of the donut hole ($\delta_1 q_1$). These differences in the contract type and contract terms likely reflect variation in bargaining power as well as heterogeneity in the patient pool across hospitals. For instance, we find that larger hospitals (presumably with greater bargaining power) are more likely to have the first type of contract. Also, conditional on having the first type of contract, larger hospitals are likely to have higher marginal reimbursement rates δ_1 and δ_2 (see Ho (2009) for a nice discussion of hospitals' bargaining power and markups).[18] In our analysis, we focus on hospitals whose reimbursement schedules display two kinks.

Our empirical measure of q is "billed charges" that hospitals submit to the insurer, which is the sum of list prices times the quantities of all items.[19] The list prices are set well above marginal and average costs and are generally determined based on expected costs plus a mark-up. While the 'charge master' (the file in which list prices are kept) is fixed for all patients in a hospital in a given period (the charge master is periodically updated), it is well known that the charge master varies significantly across hospitals, in which case we would observe different charges for two patients in two different hospitals even if they received the same level of treatment. In our analysis, ordering of patients will be done separately for each hospital and organ type in order to address this issue.

One might be concerned that sicker patients might incur lower charges because they

[18] Since we focus on relatively large hospitals in Table I, cross-hospital variations in marginal reimbursement rates and the locations of the first and second kinks are smaller in Table I than in the original sample.

[19] To be precise, q measures all charges incurred between admission and 90 days after discharge. This period includes most of the major components related to transplant care, such as organ procurement, transplant operation, inpatient care and necessary follow-ups within 90 days post discharge. Typically, 75% of the total costs associated with transplant care occur during this period. The reimbursement schedules we examine apply to charges incurred during this period only, and there are separate provisions for charges incurred prior to admission or after more than 90 days post discharge. Since there are no items that are not eligible for reimbursement, all expenses incurred by the hospital are included in q.

die soon after the transplant, violating our assumption about monotonicity of charges in sickness θ. However, the data are not consistent with this hypothesis. We explore the reasonableness of our assumption that health status is monotonic in charges by estimating the functional relationship between charges and patient mortality. While our primary data does not contain information on mortality (or other health outcome endpoints), we can turn to hospital discharge data to examine the relationship between charges and in-hospital mortality. We use California hospital discharge data from the Office of Statewide Health Planning and Development and we cull BMT, kidney and lung transplant patients from that data for our analysis (N = 2,980). We estimate a simple logit model of the likelihood of death as a function of a polynomial of charges with hospital fixed effects for privately insured patients.[20] For all organs, the parameter estimates imply a strong monotonic and statistically significant relationship between charges and mortality. That is, the estimates imply that charges are increasing in severity of illness.

One practical issue we encounter is that the number of patients who receive a certain type of organ transplant within a hospital is typically small. To deal with this issue, we pool observations across years for a given hospital and organ type (as long as the reimbursement structure does not change over time) since it seems plausible to expect that a given hospital's price sensitivity does not change during the short sample period. To further reduce the potential bias arising from the small number of patients, we restrict our attention to (hospital, organ) pairs that have enough observations.[21] Table I presents summary statistics for our estimation sample.

[Table I about here]

From the table, it is clear that there is a huge variation in charges. A simple regression shows that about 15%-25% of variation in charges is explained by hospital dummies for each of the organ types. This could be due to differences across hospitals in patient pool, list prices or innate resource use intensity. Ideally, we would closely examine the various components of the charges—the costs of organ procurement, hospitalization, tests, drugs, etc. for each patient. Our data essentially is the information that the insurer receives from the hospital and such detail is not transmitted to the insurer and is generally not

[20] The inpatient mortality rates for BMT, kidney and lung transplants are .008, .067 and .067, respectively.

[21] In our result tables below we report how many patients each hospital has.

available. It is important to note that the charge master is fixed for a given hospital year and does not vary by patient. Thus, for a given hospital changes in the charges reflect changes in quantity.

The lack of information on detailed components of charges also prevents us from empirically examining what hospitals do in practice to adjust their level of care q in the face of financial incentives. However, an executive at the insurer told us that hospitals can and do manipulate charges to increase reimbursements. Note that the q measure in our application includes post-operative care for some period of time. During this time period, hospitals have significant discretion over q. Hospitals can discharge patients earlier or later, depending on how sick the patient is, and also potentially depending on the reimbursement structure. Hospitals have case managers who are keenly aware of the reimbursement structure for expensive patients like transplants and monitor how long patients have stayed, the associated costs, etc. Transplant surgeons we interviewed highlighted that there is significant variation in resource use that is attributable to testing and many of these tests are, in fact, discretionary with modest expected benefit. A hospital staff also mentioned another interesting example of an action hospitals take that affects the charges for a given patient. Hospitals often contract with nearby hotels and step-down facilities and place transplant patients in the advanced stages of their recovery in them instead of keeping inpatient setting. The staff also noted that the utilization of those facilities is often discretionary and financial incentives can affect their use.

Returning to the insurer data, in Figure V we plot the empirical quantile function of q for one hospital's liver transplants to illustrate the source of variation we use in our estimation. The plot only uses data whose q is above the first discontinuity point since our identification comes from data around the second discontinuity point. The horizontal line reflects the second discontinuity point in marginal reimbursement rates, q_2. For this hospital, the marginal reimbursement rate jumps from 0% to 65% at q_2. From the figure we see that the slope of the quantile function is much steeper right above q_2 than right below q_2, which is consistent with the prediction we saw in Figure IV. The plot also suggests that there is no clear sign of gap, an issue we will return to below.

[Figure V about here]

5 Application: Results

In order to estimate the responsiveness of health care provision to changes in the reimbursement structure, we apply our proposed estimators to the second discontinuity point q_2. In the first set of results, we apply maximum likelihood estimation to data pooled across multiple hospitals. The small number of observations per hospital in our data makes this approach more appealing than nonparametric estimation applied to each hospital separately. The maximum likelihood estimation will yield $\hat{\alpha}_L, \hat{\alpha}_H, \hat{\beta}_L$ and $\hat{\beta}_H$, and these allow us to obtain the size of gap, $\hat{\phi}_{GAP}(x_t) = g_H(x_t, \hat{\beta}_H) - g_L(x_t, \hat{\beta}_L)$, and the change in the slope of the quantile function, $\hat{\phi}_{SLOPE}(x_t) = \frac{1}{f_H(0,x_t,\hat{\alpha}_H)} - \frac{1}{f_L(0,x_t,\hat{\alpha}_L)}$, at the discontinuity point for each hospital characterized by x_t. We use exponential distribution for densities f_L and f_H with hazard rate parameter $\lambda_L(x_t, \alpha_L)$ and $\lambda_H(x_t, \alpha_H)$, respectively. All contract variables that potentially differ across hospitals, such as the locations of the first and second kinks (q_1 and q_2) and the marginal reimbursement rates (δ_1 and δ_2), are included in x_t. We also include higher-order polynomials of these variables in x_t to flexibly capture the distribution of q for multiple hospitals. To compute standard errors, we use parametric bootstrap using 500 simulations. We apply MLE to each organ type separately.

In Tables II-IV, we report maximum likelihood estimates of ϕ_{GAP} and ϕ_{SLOPE} for each hospital in the data, along with hospital characteristics.[22] We report the number of patients treated in each hospital as well. Since our maximum likelihood estimation is applied to pooled observations across hospitals within a given organ type, the relevant number of observations used in estimation is much higher than that indicated by each individual hospital. Table II reports estimates for BMT, Table III for kidney transplants, and Table IV for liver transplants.

[Tables II, III and IV about here]

From the results in Tables II-IV, we see that $\hat{\phi}_{SLOPE}$ is positive and statistically significant for almost all cases. This suggests that for a given increase in the severity of patient health shock, hospitals tend to increase their health care spending by a larger amount when they face a positive marginal reimbursement rate than when the marginal

[22] We know hospital names as well but we are not allowed to divulge them.

reimbursement rate is zero. To interpret the magnitude of the coefficients, take the results for hospital 1 in Table II. The hospital increases its bone marrow transplant spending by \$252.1 for one percentile increase in patient illness severity when it is on the LHS of the kink (marginal reimbursement rate = 0%), while it increases its spending by \$494.5 for one percentile increase in illness severity when it is on the RHS of the kink (marginal reimbursement rate = 50%). This amounts to approximately two times larger sensitivity of the hospital's health care spending to BMT patients' health condition due to the hike in the reimbursement rate. Similarly, take the results for hospital 1 in Table III.[23] The hospital increases its kidney transplant spending by \$113 for one percentile increase in illness severity when it is on the LHS of the kink (marginal reimbursement rate = 0%), while it increases its spending by \$515.3 for one percentile increase in illness severity when it is on the RHS of the kink (marginal reimbursement rate = 55%). This amounts to approximately four and a half times larger sensitivity of the hospital's health care spending to kidney transplant patients' health condition due to the increase in the reimbursement rate. Similar results hold for liver transplants as well.

Overall, the sensitivity of health care spending to patient illness is 2 to 13 times larger on the RHS than on the LHS for bone marrow transplants, 2 to 7 times larger on the RHS than on the LHS for kidney transplants, and 3 to 14 times larger on the RHS than on the LHS for liver transplants. What is also interesting is that $\hat{\phi}_{SLOPE}$ tends to be larger when δ_2 is larger, which is again consistent with the idea that hospitals are sensitive to reimbursement rates in their health care provision decision.[24] For instance, In Table II, $\hat{\phi}_{SLOPE}$ is largest for Hospital 3, which has the largest δ_2, and $\hat{\phi}_{SLOPE}$ is smallest for Hospital 1, which has the smallest δ_2. Similar patterns hold for liver transplants (Table IV) although the picture is less clear for kidney transplants (Table III). These results suggest that financial incentives matter for hospitals' decision on resource use.

Another pattern we observe in Tables II-IV is that $\hat{\phi}_{GAP}$ is always positive, although insignificant most of the time. The fact that $\hat{\phi}_{GAP}$ is always positive alleviates concerns about possible model misspecification. As we discussed in Section 3.2.2, an implicit assumption of the parametric functional form is that $g_L(x_t, \beta_L^0) \leq q_2(t) \leq g_H(x_t, \beta_H^0)$ for all t at the true parameters β_L^0 and β_H^0. Since we find that $g_L(x_t, \hat{\beta}_L) < g_H(x_t, \hat{\beta}_H)$ holds for all hospitals in the data, there is no clear evidence of model misspecification. In

[23] The k^{th} hospital in Table II is different from the k^{th} hospital in Table III or IV.

[24] Since hospitals might differ in various aspects other than δ_2, we simply note this as an interesting pattern that is consistent with our predictions but do not attempt to draw any strong conclusion from it.

terms of sheer magnitudes, the estimates of $\hat{\phi}_{GAP}$ are quite large. For instance, the size of gap for hospital 1 in Table II is $8610. But due to large standard errors, most estimates of $\hat{\phi}_{GAP}$ are statistically indistinguishable from zero. In light of these results, we mainly focus on the interpretation of $\hat{\phi}_{SLOPE}$ in the remainder of this paper.

In order to test the robustness of our results, we perform our analysis at a finer level of aggregation: at the individual hospital level. In this second set of results, we apply kernel estimator as discussed in Section 3.2.1 to estimate ϕ_{SLOPE} separately for each pair of hospital and organ. This approach also allows us to use only local variation around the cutoff for identification of incentive effects. We do not report estimates of ϕ_{GAP}, taking our earlier results into account. In our estimates, half-normal kernels with various choices of bandwidth were used to construct the weights. We use Silverman's plug-in estimates for bandwidths and also try using twice and half the plug-in estimates to test robustness. Table V reports kernel estimates of ϕ_{SLOPE} for each hospital and each organ type.[25]

[Table V about here]

The results in Table V are similar to our earlier results, although the magnitudes and significance differ. The fact that our global estimator (MLE) and local estimator (kernel) lead to similar conclusions is reassuring.

An overall picture that consistently appears in all these results is that hospitals tend to submit much larger bills when marginal reimbursement rates are higher. Although we cannot determine whether this is mainly due to underprovision below the threshold (necessary care is withheld) or overprovision above the threshold (unnecessary care is provided) due to the lack of information on the components of the final charges, the finding that hospitals are highly sensitive to financial incentives in their health care decisions is very interesting. We also emphasize that our estimates are valid only locally around the threshold point, and do not allow us to infer the general price sensitivity of hospitals over a wide range of expenditures. This limitation of local validity is an inherent feature of RDD-type estimators.

If we had infinitely many data points, our approach outlined in Section 3.1 would suggest that we look for a "break" in the slope of the quantile function in an arbitrarily

[25] Since estimation is done separately for each hospital, only those hospitals with sufficient numbers of observations are used in Table V. As a result, the number of hospitals reported in Table V is smaller than those in Tables II-IV. In Table V, we report the number of patients treated by each hospital.

small neighborhood around θ^*. Due to the sparseness of our data, however, it is hard to tell from the raw data whether the density function has a discontinuous change at θ^*. Thus, essentially our estimates simply tell us that the average density on the RHS is smaller than the average density on the LHS within a small window around the cutoff point. Then a question that could potentially arise is whether we can interpret the observed change in the density as a result of the change in the marginal reimbursement rate. This kind of interpretational issue often arises in RDD applications since researchers frequently need to deal with small data.

To address this potential concern, we run the same type of analysis for a "control group." We have a set of hospitals whose contracts have only one kink point (they have the first kink point, but after the first kink point, the marginal reimbursement rate is always zero). We then impose an artificial cutoff point, similar in location to the cutoff point for our estimation sample of hospitals, and perform similar analysis as in Table V. If our earlier results are an artifact of e.g. the right-skewed distribution of expenditures or something else unrelated to financial incentives, we might expect to find similar results for this control group. Estimation results for this control group of hospitals are reported in Table VI.

[Table VI about here]

A comparison of Table VI against Table V indicates that our earlier results were likely reflective of hospitals' true behavioral responses to financial incentives. In Table V we saw that the estimates of ϕ_{SLOPE} were positive and statistically significant for 6 out of 7 hospitals, while the only negative estimate was not statistically significantly different from zero. In contrast, we see that one hospital has a negative estimate and another has a positive estimate among 2 hospitals with a statistically significant estimate of ϕ_{SLOPE} in our control group of 4 hospitals. These results from the "control" group help partially alleviate concerns that our main findings may be an artifact of discontinuities in the density function of spending at q_2 for reasons other than financial incentives, such as a skewed distribution of health shocks (and accordingly expenditures) and expenditure lumpiness due to the presence of frequently implemented medical procedures. Thus, we conclude that our kernel estimates of ϕ_{SLOPE} in Table V mostly reflect true behavioral responses of hospitals to reimbursement structures.

6 Discussion

In this section, we discuss other settings to which our estimator can be applied and also how our proposed methodology is related to the existing methods. The first important class of models where our estimator can be used is consumer choice under nonlinear pricing. Nonlinear pricing is a very common practice in real life. For instance, Wilson in his book on nonlinear pricing (1997) notes “utilities in the power industry have long offered a variety of nonlinear rate schedules, especially block-declining tariffs for commercial and industrial customers.” Block-declining tariffs mean that lower marginal rates apply to successive blocks of usage. Since a sudden drop in marginal cost is equivalent to a sudden jump in marginal reimbursement rate in our model of Section 2, we will not have an issue of bunching at the thresholds in these settings. Under the assumption that the marginal utility of consumption is increasing in consumer’s type (single-crossing property is often assumed in the nonlinear pricing literature), it is clear that our estimator can be applied to these settings.

Another class of models to which our proposed method is applicable is contracting with nonlinear incentives. Many insurance products have deductibles or donut holes, meaning that the marginal reimbursement rate faced by the insured experiences a sudden increase when a certain threshold is reached. For instance, Medicare Part D, a federal program to subsidize the costs of prescription drugs for the elderly in the US, has a coverage gap such that a Medicare beneficiary is fully responsible for the costs of prescription drugs if his expense exceeds the initial coverage limit but falls short of the catastrophic coverage threshold. The presence of this coverage gap implies that marginal reimbursement rates change discontinuously when a person hits the initial coverage limit or the catastrophic coverage threshold. If a researcher is interested in learning how responsive seniors are to marginal reimbursement rates in their usage of prescription drugs, the researcher can apply our proposed method to the second threshold (catastrophic coverage threshold), where marginal reimbursement rate jumps. Relatedly, an empirical application of our method is found in Marsh (2011), where she employs our estimator in the case of health savings accounts in order to infer sensitivity of patients’ health expenditures to price.

We view our work as complementary to the work by Saez (2010) and Chetty et al. (2011), where they use the size of bunching to infer sensitivity of labor supply to marginal tax rates. In case of labor supply, marginal tax rates are higher for successive income brackets, and these sudden increases in marginal “costs” correspond to the first discontinuity point q_1 in our Figure II. We, on the other hand, propose to use the size of gap and a discontinuous change in the density of the outcome variable when there are sudden decreases in marginal costs, or equivalently, sudden increases in marginal returns. Some applications might have both sudden increases and decreases in marginal costs at various thresholds, in which case an examination of bunching, gap and discontinuous change in the density of the outcome variable together could provide a comprehensive understanding of how agents' behavior responds to incentives. In other applications, there might be only sudden increases in marginal costs, in which case bunching would be the only relevant dimension to study. Yet other applications might have only sudden decreases in marginal costs, and our methods can be used in such cases.

In our empirical application, the two dimensions we examined — gap and discontinuous change in density of outcome variable — gave somewhat different answers: we found no evidence of significant gap but the density of the outcome variable has a significant change at the threshold. We think this might be related to optimization error or lumpiness in health expenditures. For instance, Saez (2010) finds surprisingly weak evidence of bunching around the kink points of the income tax schedule and considers agents with optimization error to explain such a finding. In our method, if agents have optimization error, gap is likely to be underestimated as well as the change in the slope of the quantile function, as discussed in Section 3. In general, it is not clear whether the extent of downward bias would be greater for bunching, gap or change in the density of outcome variable in the presence of optimization error. In our empirical application, it might be that the downward bias of gap is larger such that we still find some impact on the density of outcome variable, but no evidence of gap. Although we cannot make a general statement about whether and when this will happen in other settings, our method at the minimum proposes additional dimensions to examine in order to obtain a more complete view on how agents' behavior is influenced by incentives.

7 Conclusion

In this paper, we propose an estimator that exploits discontinuity in a nonlinear pricing schedule to recover price sensitivity of economic agents. Our proposed estimator can be applied to many interesting settings such as consumer choice under nonlinear pricing and contracting with nonlinear incentives. An application of our estimator to contracts in the health care market reveals that hospitals' health care spending is significantly influenced by financial incentives.

The assumptions required for our estimator are unlikely to hold for all settings, and thus it is important for researchers to examine whether the assumptions hold for their problems of interest. A key assumption is the strict monotonicity between the type and the dependent variable. This is likely to be violated if the type is multidimensional or if there is optimization error or measurement error. In future work, we plan to investigate the performance of our estimator under more general conditions and improve our estimator to make it robust against these complications.

References

[1] **Arrow, Kenneth**, 1963, "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review,* 53: 941-973.

[2] **Burtless, Gary and Jerry Hausman**, 1978, "The Effect of Taxation on Labor Supply: Evaluating the Gary Income Maintenance Experiment," *Journal of Political Economy*, 86: 1103-1130.

[3] **Chandra, Amatabh and David Cutler and Zurui Song**, 2012, "Who Ordered That? The Economics of Treatment Choices in Medical Care," in *Handbook of Health Economics, Vol 2,* eds. M. Pauly and T. McGuire and Pedro P. Barros, North-Holland, 397-432.

[4] **Chernozhukov, Victor**, 2005, "Extremal Quantile Regression," *The Annals of Statistics*, 33(2): 806-839.

[5] **Chernozhukov, Victor and Han Hong**, 2004, "Likelihood Estimation and Inference in a Class of Nonregular Econometric Models," *Econometrica*, 72(5): 1445-1480.

[6] **Chetty, Raj, John Friedman, Tore Olsen, and Luigi Pistaferri**, 2011, "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Record," *Quarterly Journal of Economics*, 126, 749–804.

[7] **Copeland, Adam and Cyril Monnet**, 2009, "The Welfare Effects of Incentive Schemes," *Review of Economic Studies,* 76: 93–113.

[8] **Cutler, David**, 1995, "The Incidence of Adverse Medical Outcomes Under Prospective Payment," *Econometrica*, 63(1): 29-50.

[9] **Dafny, Leemore**, 2005, "How Do Hospitals Respond to Price Changes?" *American Economic Review*, 95(5): 1525-1547.

[10] **Donald, Stephen and Harry Paarsch**, 1996, "Identification, Estimation, and Testing in Parametric Empirical Models of Auctions within Independent Private Values Paradigm," *Econometric Theory*, 12: 517-567.

[11] **Dranove, David and Paul Wehner**, 1994, "Physician-Induced Demand for Childbirths," *Journal of Health Economics,* 13(1): 61-73.

[12] **Fuchs, Victor**, 1978, "The Supply of Surgeons and the Demand for Operations," *Journal of Human Resources*, 13: 121-133.

[13] **Gaynor, Martin and Paul Gertler**, 1995, "Moral Hazard and Risk Spreading in Partnerships," *RAND Journal of Economics*, 26(4): 591-613.

[14] **Gaynor, Martin, James Rebitzer and Lowell Taylor**, 2004, "Physician Incentives in Health Maintenance Organizations," *Journal of Political Economy*, 112(4): 915-931.

[15] **Glied, Sherry**, 2000, "Managed Care," in *Handbook of Health Economics,* eds. A. Cuyler and J. Newhouse, North-Holland, 707-753.

[16] **Gruber, Jonathan and Maria Owings**, 1996, "Physician Financial Incentives and Cesarean Section Delivery," *RAND Journal of Economics*, 27(1): 99-123.

[17] **Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw**, 2001, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1): 201-209.

[18] **Hausman, Jerry**, 1979, "The Econometrics of Labor Supply on Convex Budget Sets," *Economics Letters*, 3:171–174.

[19] **Hausman, Jerry**, 1985, "The Econometrics of Nonlinear Budget Sets," *Econometrica*, 53(6): 1255–1282.

[20] **Ho, Katherine**, 2009, "Insurer-Provider Networks in the Medical Care Market," *American Economic Review*, 99(1): 393–430.

[21] **Hodgkin, Dominic and Thomas McGuire**, 1994, "Payment Levels and Hospital Response to Prospective Payment," *Journal of Health Economics*, 13: 1-29.

[22] **Imbens, Guido and Thomas Lemieux**, 2008, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142(2): 615-635.

[23] **Ketcham, Jonathan, Pierre Léger and Claudio Lucarelli**, 2011, "Standardization Under Group Incentives," Working Paper.

[24] **Koenker, Roger and Gilbert Bassett**, 1978, "Regression Quantiles," *Econometrica*, 46: 33-50.

[25] **Lee, David and Thomas Lemieux**, 2010, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48(2): 281-355.

[26] **Marsh, Christina**, 2011, "Estimating Health Expenditure Elasticities Using Nonlinear Reimbursement," Working Paper.

[27] **McClellan, Mark**, 2011, "Reforming Payments to Healthcare Providers: The Key to Slowing Healthcare Cost Growth While Improving Quality?" *Journal of Economic Perspectives*, 25(2): 69-92.

[28] **McCrary, Justin**, 2008, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142(2): 698-714.

[29] **McGuire, Thomas**, 2000, "Physician Agency," in *Handbook of Health Economics,* eds. A. Cuyler and J. Newhouse, North-Holland, 467-536.

[30] **Milgrom, Paul and Chris Shannon**, 1994, "Monotone Comparative Statics," *Econometrica*, 62(1): 157-180.

[31] **Misra, Sanjog and Harikesh Nair**, 2011, "A Structural Model of Sales-Force Compensation Dynamics: Estimation and Field Implementation," *Quantitative Marketing and Economics*, 9(3): 211-225.

[32] **Moffitt, Robert**, 1986, "The Econometrics of Piecewise-Linear Budget Constraints: A Survey and Exposition of the Maximum Likelihood Method," *Journal of Business & Economic Statistics,* 4(3): 317-328.

[33] **Nekipelov, Denis**, 2010, "Empirical Content of a Continuous-Time Principal-Agent Model," Working Paper.

[34] **Organ Procurement and Transplantation Network (OPTN) and Scientific Registry of Transplant Recipients (SRTR)**, 2007, *Annual Data Report*, Rockville, MD: Department of Health and Human Services, Health Resources and Services Administration.

[35] **Pudney, Stephen**, 1989, *Modelling Individual Choice: The Econometrics of Corners, Kinks, and Holes*, Cambridge, UK: Blackwell Publishers.

[36] **Reiss, Peter and Matthew White**, 2005, "Household Electricity Demand, Revisited," *Review of Economic Studies*, 72: 853–883.

[37] **Rosenthal, Meredith, 2004**, "Donut-Hole Economics," *Health Affairs*, 23(6): 129-135.

[38] **Saez, Emmanuel**, 2010, "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 2: 180-212.

[39] **Topkis, Donald**, 1978, "Minimizing a Submodular Function on a Lattice," *Operations Research*, 26: 305-321.

[40] **Urquiola, Miguel and Eric Verhoogen**, 2009, "Class-Size Caps, Sorting, and the Regression-Discontinuity Design," *American Economic Review*, 99(1): 179-215.

[41] **Wilson, Robert**, 1997, *Nonlinear Pricing*, Oxford University Press.

Table I: Summary Statistics

	BMT	Kidney	Liver
Total # Patients	511	265	344
Total # Hospitals	7	6	7
Avg. Charge per Patient (in $1000)	168.56 (114.87)	140.74 (84)	319.63 (246.4)
Avg. Reimbursement per Patient (in $1000)	100.27 (67.98)	76.5 (42.12)	196.12 (145.61)
Avg. # Patients per Hospital	73 (36.82)	44.17 (11.57)	49.14 (30.39)
Avg. q_1 across Hospitals (in $1000)	121.97 (13.65)	80.09 (9.42)	181.57 (17.34)
Avg. q_2 across Hospitals (in $1000)	163.10 (22.44)	128.24 (12.63)	248.37 (28.57)
Avg. δ_1 across Hospitals	0.75 (0.08)	0.82 (0.08)	0.79 (0.07)
Avg. δ_2 across Hospitals	0.56 (0.06)	0.51 (0.05)	0.58 (0.06)
Avg. % Patients with $q < q_1$	37.39 (16.22)	8.92 (5.62)	10.33 (6.58)
Avg. % Patients with $q_1 \leq q \leq q_2$	24.76 (11.45)	42.20 (15.76)	33.98 (15.55)
Avg. % Patients with $q > q_2$	37.85 (20.33)	48.88 (11.02)	55.69 (12.21)

Inside the parentheses are standard deviations.

Table II: Maximum Likelihood Estimates, Bone Marrow Transplant (q measured in $1,000)

	Obs	q_1	q_2	δ_1	δ_2	$\hat{\phi}_{GAP}$	Slope L	Slope R	$\hat{\phi}_{SLOPE}$
H1	28	135.71	190	0.7	0.5	8.61 (78.46)	25.21	49.45	24.24 (14.93)
H2	73	133.33	181.82	0.75	0.55	5.13 (4.05)	22.40	58.00	35.59 (10.38) ***
H3	74	107.33	123.85	0.75	0.65	4.70 (8.42)	11.28	146.29	135.01 (21.35) ***
H4	148	120	150	0.75	0.6	7.10 (2.63) **	15.42	95.99	80.57 (7.46) ***
H5	64	126.67	172.73	0.75	0.55	35.22 (16.23) *	20.31	66.21	45.91 (7.45) ***
H6	71	130.77	170	0.65	0.5	3.44 (3.62)	20.31	66.19	45.87 (14.04) ***
H7	53	100	153.33	0.92	0.6	10.27 (16.63)	15.98	91.44	75.46 (7.71) ***

Inside the parentheses are bootstrapped standard errors.

* Significant at 10% ** Significant at 5% *** Significant at 1%

Table III: Maximum Likelihood Estimates, Kidney Transplant (q measured in $1,000)

	Obs	q_1	q_2	δ_1	δ_2	$\hat{\phi}_{GAP}$	Slope L	Slope R	$\hat{\phi}_{SLOPE}$
H1	29	97.14	123.64	0.7	0.55	0.78 (6.96)	11.30	51.53	40.23 (5.27) ***
H2	37	74.67	124.44	0.75	0.45	6.69 (25.01)	14.25	74.07	59.82 (24.72) **
H3	48	75.29	130.61	0.85	0.49	4.67 (3.98)	14.66	57.15	42.49 (14.41) ***
H4	47	79.41	137.78	0.85	0.49	3.80 (4.13)	16.75	50.41	33.65 (6.37) ***
H5	41	83.38	144.64	0.85	0.49	1.95 (4.97)	19.05	44.66	25.60 (8.13) ***
H6	63	70.65	108.33	0.92	0.6	1.44 (3.32)	7.62	55.86	48.24 (8.76) ***

Inside the parentheses are bootstrapped standard errors.

* Significant at 10% ** Significant at 5% *** Significant at 1%

Table IV: Maximum Likelihood Estimates, Liver Transplant (q measured in $1,000)

	Obs	q_1	q_2	δ_1	δ_2	$\hat{\phi}_{GAP}$	Slope L	Slope R	$\hat{\phi}_{SLOPE}$
H1	95	178.53	206	0.75	0.65	2.39 (5.76)	13.05	181.72	168.67 (22.39) ***
H2	31	166.47	288.78	0.85	0.49	13.99 (17.21)	42.14	141.11	98.97 (25.58) ***
H3	23	160	218.18	0.75	0.55	10.01 (258.73)	14.77	155.15	140.37 (125.67)
H4	38	200	254.55	0.7	0.55	7.77 (12.08)	25.85	155.15	129.30 (17.6) ***
H5	42	198.8	271.09	0.75	0.55	26.02 (13.51) *	33.34	155.15	121.81 (16.72) ***
H6	90	198.72	250	0.78	0.62	6.54 (5.73)	25.19	173.30	148.11 (13.45) ***
H7	25	168.48	250	0.92	0.62	33.30 (29.81)	25.19	173.30	148.11 (13.45) ***

Inside the parentheses are bootstrapped standard errors.

* Significant at 10% ** Significant at 5% *** Significant at 1%

Table V: Kernel Estimates (q measured in $1,000)

	Obs	Bandwidth	Slope L	Slope R	$\hat{\phi}_{SLOPE}$
H2 (BMT)	73	Silverman's Plug-In	34.70 (0.91)	51.87 (1.69)	17.17 (2.6) ***
H4 (BMT)	148	Silverman's Plug-In	21.36 (0.48)	186.29 (9.01)	164.93 (9.48) ***
H3 (Kidney)	48	Silverman's Plug-In	46.90 (3.59)	71.6 (2.03)	24.7 (5.63) ***
H5 (Kidney)	41	Silverman's Plug-In	34.08 (0.79)	72.53 (5.7)	38.45 (6.48) ***
H6 (Kidney)	63	Silverman's Plug-In	16.35 (0.26)	88.06 (2.75)	71.71 (3.0) ***
H2 (Liver)	31	Silverman's Plug-In	130.71 (50.71)	109.98 (10.95)	-20.73 (61.67)
H6 (Liver)	90	Silverman's Plug-In	54.12 (4.81)	140.87 (5.88)	86.75 (10.7) ***
H2 (BMT)	73	1/2 × Silverman's	37.04 (2.21)	42.72 (1.88)	5.67 (4.09)
H4 (BMT)	148	1/2 × Silverman's	19.61 (0.74)	211.55 (26.38)	191.93 (27.11) ***
H3 (Kidney)	48	1/2 × Silverman's	48.39 (7.89)	49.22 (1.32)	0.83 (9.21)
H5 (Kidney)	41	1/2 × Silverman's	29.54 (1.02)	78.6 (14.5)	49.06 (15.52) ***
H6 (Kidney)	63	1/2 × Silverman's	17.13 (0.59)	68.5 (2.59)	51.37 (3.18) ***
H2 (Liver)	31	1/2 × Silverman's	323.97 (1544)	80.65 (8.64)	-243.32 (1553)
H6 (Liver)	90	1/2 × Silverman's	57.35 (11.46)	109.71 (5.56)	52.36 (17.02) ***
H2 (BMT)	73	2 × Silverman's	39.54 (0.67)	71.25 (2.19)	31.7 (2.86) ***
H4 (BMT)	148	2 × Silverman's	27.21 (0.49)	200.7 (5.63)	173.5 (6.12) ***
H3 (Kidney)	48	2 × Silverman's	43.22 (1.41)	116.31 (4.36)	73.09 (5.76) ***
H5 (Kidney)	41	2 × Silverman's	44.5 (0.87)	87.1 (4.93)	42.6 (5.81) ***
H6 (Kidney)	63	2 × Silverman's	17.5 (0.16)	125.74 (4)	108.24 (4.16) ***
H2 (Liver)	31	2 × Silverman's	113.66 (16.67)	170.64 (20.47)	56.99 (37.14)
H6 (Liver)	90	2 × Silverman's	54.81 (2.5)	211.4 (9.94)	156.59 (12.44) ***

Inside the parentheses are standard errors.

Table VI: Kernel Estimates, “Control” Group (q measured in \$1,000)

	Obs	Bandwidth	Slope L	Slope R	$\hat{\phi}_{SLOPE}$
H1 (BMT)	30	Silverman’s Plug-In	65 (6.55)	58.91 (3.27)	-6.09 (9.81)
H1 (Kidney)	36	Silverman’s Plug-In	97.17 (32.18)	26.68 (1.12)	-70.49 (33.3) **
H2 (Kidney)	46	Silverman’s Plug-In	27.92 (0.98)	69.59 (3.56)	41.67 (4.54) ***
H1 (Liver)	60	Silverman’s Plug-In	770.889 (5452)	422.43 (166.13)	-348.46 (5618)
H1 (BMT)	30	1/2 × Silverman’s	73.62 (19.02)	43.34 (2.6)	-30.29 (21.62)
H1 (Kidney)	36	1/2 × Silverman’s	295.37 (1807)	27.94 (2.57)	-267.43 (1810)
H2 (Kidney)	46	1/2 × Silverman’s	29.62 (2.33)	55.24 (3.56)	25.62 (5.89) ***
H1 (BMT)	30	2 × Silverman’s	72.51 (4.54)	83.4 (4.63)	10.89 (9.18)
H1 (Kidney)	36	2 × Silverman’s	55.78 (3.04)	37.16 (1.51)	-18.62 (4.56) ***
H2 (Kidney)	46	2 × Silverman’s	28.62 (0.53)	102 (5.61)	73.38 (6.14) ***
H1 (Liver)	60	2 × Silverman’s	290.81 (146.34)	581.06 (216.18)	290.25 (362.52)

Inside the parentheses are standard errors.

* Significant at 10% ** Significant at 5% *** Significant at 1%

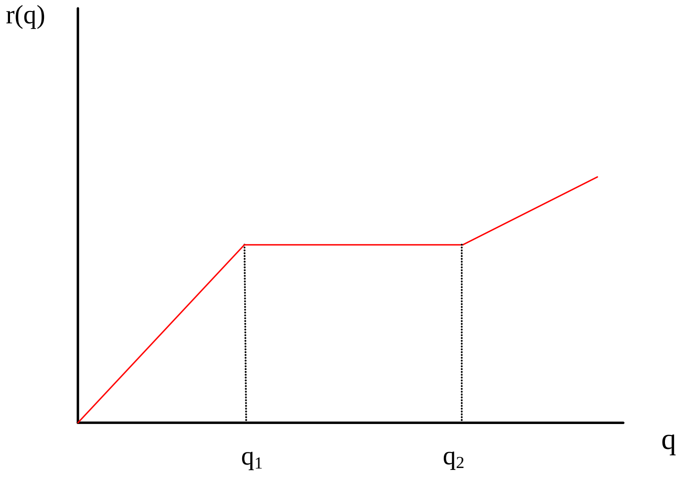

Figure I: A Typical Reimbursement Scheme

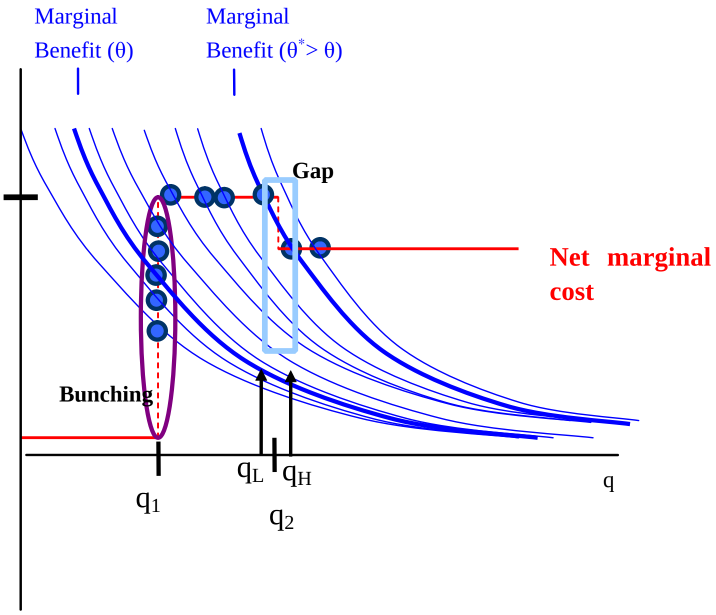

Figure II: Optimal Decision Rule for Low γ^a Agent

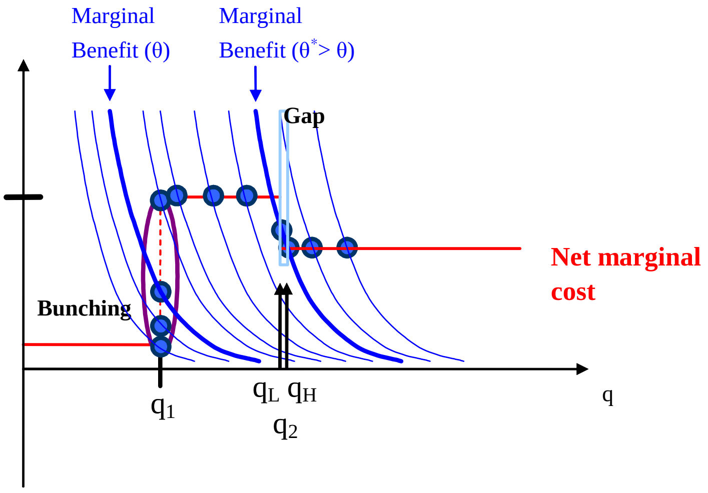

Figure III: Optimal Decision Rule for High γ^a Agent

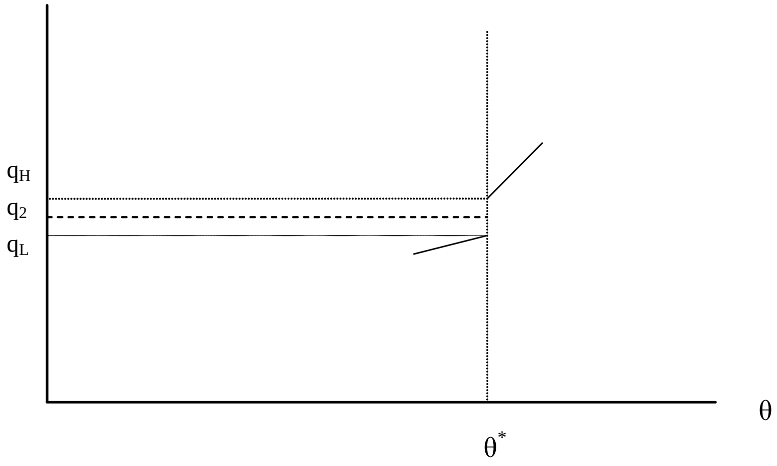

Figure IV: The slope of quantile function $q(\theta)$ changes at θ^*

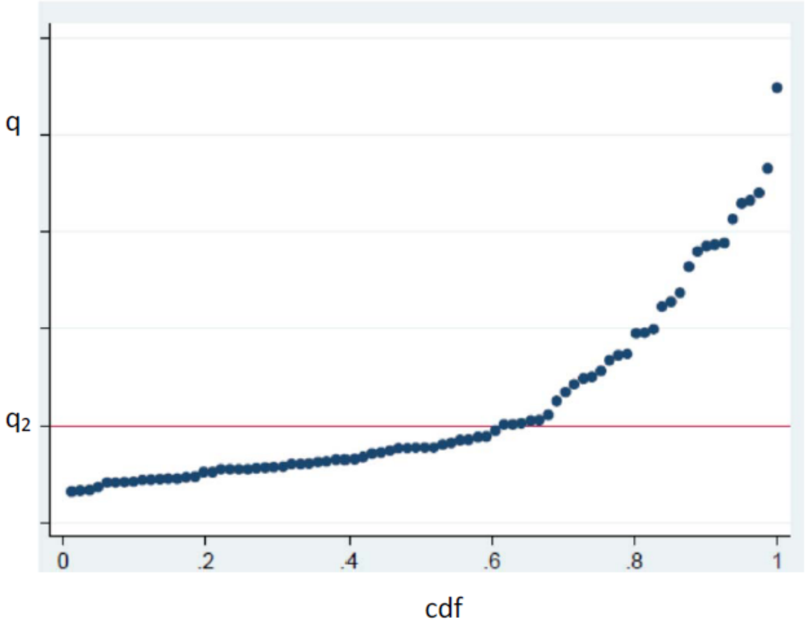

Figure V: Changes in the slope of empirical quantile function