Check for updates

Quantitative Economics 8 (2017), 397–433 1759-7331/20170397

Estimating price sensitivity of economic agents using discontinuity in nonlinear contracts

Patrick Bajari
Department of Economics, University of Washington and NBER

Han Hong
Department of Economics, Stanford University

Minjung Park
Haas School of Business, University of California, Berkeley

Robert Town
Department of Economics, University of Texas at Austin and NBER

This paper proposes a method to estimate price sensitivity of economic agents exploiting 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 the impact of the reimbursement rate on hospitals' provision of health care services varies significantly across patients with different levels of illness severity. Our methodology is applicable to important classes of models such as consumer choice under nonlinear pricing and contracting with nonlinear incentives.

Keywords. Nonlinear contracts, heterogeneity in price sensitivity, health care.
JEL classification. C10, C51, I11.

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

Patrick Bajari: bajari@uw.edu
Han Hong: doubleh@stanford.edu
Minjung Park: minjungp@gmail.com
Robert Town: robert.town@austin.utexas.edu

We are grateful to the editor and anonymous reviewers for their insightful comments and constructive suggestions. We also thank participants at the Industrial Organization of Health Care Conference at HEC Montreal and the Cowles Structural Microeconomics Summer Conference for helpful comments. We gratefully acknowledge the support provided by SIEPR and NSF Research Grants SES 1164589, SES 1325805 and SES 1459975. All remaining errors are our own.

Copyright © 2017 The Authors. Quantitative Economics. The Econometric Society. Licensed under the Creative Commons Attribution-NonCommercial License 4.0. Available at http://www.qeconomics.org.
DOI: 10.3982/QE602

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. For instance, suppose a hospital chooses the optimal level of medical care intensity for a patient given a piecewise linear reimbursement schedule. To estimate the responsiveness of the hospital's medical care provision to the reimbursement rate, one might want to exploit discontinuous changes in marginal reimbursement rates to identify this effect. However, a straightforward approach would not work since marginal reimbursement rates are a function of health care expenditures.

We propose an estimation strategy that can be applied to such a setting to recover price sensitivity of economic agents as well as its heterogeneity, 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 bunching point at which different types of agents will behave identically. We use this monotonicity to recast the problem such that the type of the agent is seen as the forcing variable that shifts the relevant marginal price discontinuously at a known cutoff. We propose two estimators that exploit a discontinuous change in the marginal price at the known cutoff. The first estimator measures the price elasticity of agents' choice by evaluating the size of a gap around the cutoff in the empirical quantile function of agents' choice. The second estimator measures local heterogeneity in price elasticity by evaluating the magnitude of a discontinuous change in the slope of the empirical quantile function at the cutoff.

We apply our estimators 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 both to assessing the welfare consequences of provider agency and to 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' gross domestic product (GDP). In the United States, 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.[1]

[1] 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 (1998), Gaynor, Rebitzer, and Taylor (2004), Dafny (2005), Ketcham, Léger, and Lucarelli (2011), and Ho and Pakes (2014). See Chandra, Cutler, and Song (2012) and McClellan (2011) for an excellent review of recent developments in this research stream.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Nonlinearities in provider and insurer reimbursement contracts are becoming more common as Medicare and private insurers explore ways to encourage increased provider and insurer effort to provide high value care. For example, the Affordable Care Act established bundled payment demonstration in which the hospital would be paid a fixed amount to treat a patient during an episode in which the episode extends for 30 days post inpatient discharge. Under Medicare's value based purchasing project, after the 30 days, care is outside of the window and the hospital would be reimbursed on a fee-for-services basis for outpatient care and on a Diagnosis Related Group basis for inpatient care. Another example is under Medicare Advantage, where insurers receive extra payments and broader enrollment periods if they achieve a five star rating. This star rating system is based on continuous measures of plan performance and thus there is a discontinuity at the five star threshold. Our work proposes an econometric methodology that uses such nonlinearity in contracts to recover the elasticity of response as well as its heterogeneity—important policy parameters for which there are very few credible estimates.

We have collected a unique data set on contracts for organ and tissue transplants between one of the largest U.S. 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.

The form of the contracts in our data is simple. For each patient treated by a hospital, it keeps track of all expenses such as drugs, tests, and nights in the hospital. Our hospitals have standard list prices for each of these items, and 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 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 estimators to a discontinuity point without bunching to estimate the price sensitivity of the hospital's health care provision as well as its heterogeneity. The gap estimator is useful for understanding how effective cost sharing would be, for patients at the discontinuity point, in bending health care costs, while the slope estimator captures how the effects of cost sharing would be locally distributed across patients with marginally different levels of sickness, which is helpful for understanding the welfare effects of cost sharing policies.

Our results suggest that hospitals' behavior is strongly influenced by financial incentives. In particular, we find that in response to a marginal reimbursement rate drop, health care spending goes down more for sicker patients. The estimates indicate that hospitals reduce their health care spending by \$970 more for a slightly sicker patient (specifically, for a 1 percentile increase in illness severity) when the marginal reimbursement rate drops by 50 percentage points (a typical change seen in the data). Thus, the effects of cost sharing would fall more heavily on sicker patients.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 (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 a maximum likelihood estimator (MLE) or a generalized method of moments (GMM). 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)). Further, there is a growing literature that identifies price sensitivity of patients' health care consumption exploiting nonlinear price schedules. Examples include Kowalski (2015), Einav, Finkelstein, and Schrimpf (2015, 2017), and Dalton, Gowrisankaran, and Town (2015).

Our paper proposes an alternative approach to the problem of estimating price sensitivity using nonlinear pricing schedules. Our approach exploits local variation around a discontinuity point without bunching, while the existing approaches use either the entire schedule or a discontinuity point with bunching. Therefore, 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. Finally, our approach does not require variation in prices over time or across agents, similar to the bunching approach, since nonlinearity in the pricing schedule itself provides variation required for identification.

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 provides discussion and we conclude the paper in Section 7.

2. Model

In this section, we set up a model so as 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.

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 hospital and need treatment. Patients differ in their severity of illness that is amenable to medical care, which is denoted by $\theta \geq 0$, a random variable with a continuous density function $h(\theta)$ and cumulative distribution function (cdf) $H(\theta)$. The health shock, which is determined prior to admission, captures patient heterogeneity in the demand

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

for health care. A central assumption is that patients' heterogeneity is unidimensional, fully captured by θ. The provider then chooses a level of treatment $q \geq 0$. The value of the health services 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.[2] 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 of the general form $r(q)$ so as 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 $r(q) - c(q)$. Furthermore, we assume that the agent receives a nonpecuniary benefit that is proportional to the patient's payoff $v(q, \theta)$. This captures the idea that the agent benefits from successful health outcomes.[3] We also assume quasilinear utility functions so that there are no income effects. We can write the payoffs of the agent as

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

Thus, the agent maximizes $\gamma v(q, \theta) - c(q) + r(q)$[4] and the first order condition (FOC) is (for now, ignoring potential nondifferentiability in $r(q)$)

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

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 without loss of generality (WLOG) since we are merely rewriting preferences as a

[2] Our model is static. Since the treatment of transplants typically occurs within a short span of time, dynamics does not seem a first order concern in our setting. However, dynamics has been shown to be a key concern in other health care settings, for example, Einav, Finkelstein, and Schrimpf (2015, 2017).

[3] 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.

[4] We frame the decision of health care choice as being made by the hospital, not patients, because patients' out-of-pocket cost is unrelated to q for the vast majority of privately insured transplant patients in our empirical setting. Under a copayments structure, the patient pays a fixed cost, which does not vary with q. For patients who have a coinsurance design, receiving a transplant will typically mean that the patients will exceed their out-of-pocket maximum for the year. For those patients who do not exceed the out-of-pocket maximum, under a coinsurance structure, they pay a proportionate fraction of $r(q)$. The payments are directly tied to $r(q)$ and not q or $c(q)$. Gowrisakaran, Nevo, and Town (2015) find mean in-patient coinsurance rates of 2%. Thus, the out-of-pocket payments of the patients are a small fraction of $r(q)$ and, for the vast majority of patients, unrelated to q at the margin. As a result, we believe framing the decision of transplant care choice as being made by the hospital that faces the reimbursement schedule $r(q)$ is a very reasonable approximation.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

function of the percentile of θ. We shall assume that the payoffs satisfy the conditions

$$\frac{\partial v(q,\theta)}{\partial q} > 0, \tag{2}$$

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

$$\frac{\partial v(q,\theta)}{\partial \theta} < 0, \tag{4}$$

$$\frac{\partial^2 v(q,\theta)}{\partial \theta\, \partial q} > 0, \tag{5}$$

$$\frac{\partial c(q)}{\partial q} > 0, \tag{6}$$

$$\frac{\partial^2 c(q)}{\partial^2 q} \geq 0. \tag{7}$$

Assumptions (2) and (3) state that the value of the health services 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 services 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 higher patient utility due to a better health outcome[5] 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 is (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

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

$$r'(q) = \delta^1 \quad \text{for } 0 < q < q^1, \tag{9}$$

$$r'(q) = 0 \quad \text{for } q^1 \leq q \leq q^2, \tag{10}$$

$$r'(q) = \delta^2 \quad \text{for } q > q^2. \tag{11}$$

[5]Assumption (2) may not hold beyond an extremely high level of q, for example, flat-of-the-curve medicine or iatrogenic medicine.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

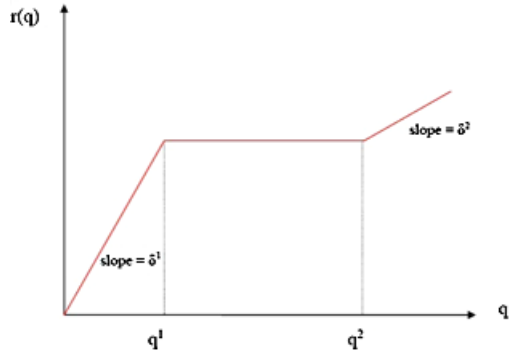

FIGURE 1. A typical reimbursement scheme.

These assumptions imply that the amount of reimbursement for each patient is piecewise linear. For expenditures below q^1, the hospital is reimbursed δ^1 for every dollar spent to treat the patient. Once expenditures exceed q^1, the hospital 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 1 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, a federal program to subsidize the costs of prescription drugs for the elderly in the United States—and high deductible health plans with an attached health savings account.

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,[6] we abstract away from the optimal contract design problem faced by the principal and just condition on the existence of nonlinearity in the contracts 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.[7]

2.3 *Optimal decision rule*

The reimbursement scheme in Figure 1 represents a commonly observed price schedule in health insurance, motivated by the desire to limit moral hazard and cut costs. Under the reimbursement schedule, the choices of the agent can have four regions that vary by the type of the patient: the quantity choice increases with type on the first segment below q^1 in Figure 1, then bunches at q^1. The choice increases further with type until

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

[7]In the case of Medicare Part D, it is often claimed that a donut hole was introduced due to limited government budget available for the program.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

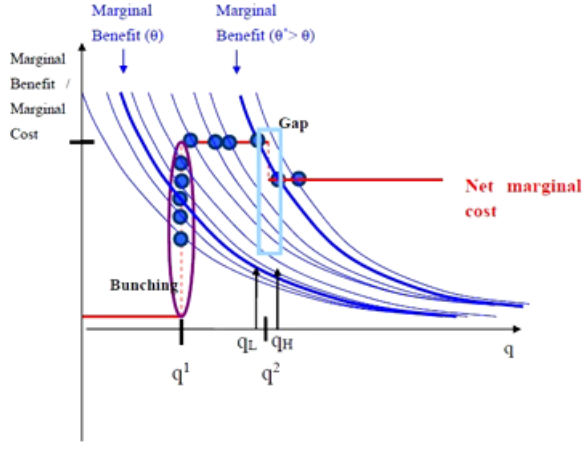

FIGURE 2. Optimal decision rule.

some point q_L below q^2, and jumps to some $q_H > q^2$. Above q_H the quantity choice is again increasing with type. In other words, the optimal decision rule of an agent who treats a pool of patients exhibits the following features:

Rule 1. There will be bunching at q^1.

Rule 2. There will be a gap around q^2.

Rule 3. The optimal choice of q is strictly increasing in θ, except for bunching at q^1.

Figure 2 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 $\gamma \frac{\partial v(q,\theta)}{\partial q}$. The lower is γ, 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 $c'(q) - r'(q)$. For this figure, we assume that $c'(q)$ is constant, which is not crucial for any of our results but simplifies the graphical analysis.

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 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 2. 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).[8] By the monotonicity of $q(\theta)$, which follows from the

[8]The variables 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), and the cost function.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 and will be on the part of the net marginal cost curve that is $c'(q) - \delta^2$. Figure 2 offers a complete treatment of what the agent's behavior would be in face of a kinked incentive scheme as described in Figure 1.

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 the 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 hold under a more general class of models.[9] For instance, in models of consumer demand under nonlinear pricing schedules, θ will represent the agent's willingness to consume goods, q will represent the quantity consumed, and a piecewise linear pricing schedule will generate discontinuity points like q^1 and q^2. To generate a discontinuity point without bunching (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 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 a single-crossing property (similar to the assumption of increasing differences in the above model), which is a very common assumption in both theoretical and empirical literatures.[10]

Similar decision rules apply 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. Again, we will have strict monotonicity of choice in type under the assumption of a single-crossing property.

A more general price schedule might have multiple kinks: some associated with bunching and others associated with gaps. The impact of data generated from contracts with multiple kinks depends on the types of kinks in the contracts. If there are multiple kinks that lead to gaps, for example, an insurance policy with a deductible and a stop-loss provision, it is possible to extend our method to such a setting: We can apply the method to each of the kinks separately, which allows us to recover price sensitivity for patients at different levels of expenditures. When some kinks are associated with a gap and others are associated with bunching, our proposed methodology is valid as long as it is applied to the kinks with a gap. Our approach exploits local variation only, and the presence of kinks in other parts of the schedule does not affect monotonicity in the local region around the focal kink point. If all the kinks are associated with bunching, our approach cannot be applied, since bunching destroys strict monotonicity of choice in agent type.

When the shape of the reimbursement schedule changes, the set of patients for whom the effects are estimated could change, because our estimates are local and the

[9] In particular, a principal–agent relationship is not necessary in deriving the optimal decision rules.

[10] To name just a few, Mussa and Rosen (1978) on product line and a large literature that follows use the assumption. A seminal paper by Spence (1973) on job market signaling in education as well as the large literature on signaling games also rely on this assumption. A vertical demand model (e.g., Bresnahan (1987)) also relies on this type of assumption.

1759733l, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

set of agents that are used for estimation might change. In our setup, the kink associated with a gap is located at a relatively high level of q, which means that our estimates would be relevant for relatively sicker patients. When the empirical setting involves price schedules that feature a gap-associated kink at low q (e.g., a deductible), the estimates would be relevant for relatively healthier patients.

3. Estimation

In this section we propose an estimator that will yield consistent estimates of the agent's behavioral responses exploiting gap-associated discontinuity in nonlinear schedules. We discuss the key intuition behind our approach and outline our estimation procedures.

3.1 *Using discontinuous changes for identification*

At the two 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 a regression discontinuity design (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 choice of q responds to the marginal reimbursement rate).

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 the percentiles of θ. That is, if we see a patient with the 5th percentile of health care expenditure within a hospital, that patient will have the 5th percentile of health shock within that hospital. This means that for all practical purposes, the types are observable to the econometrician.[11] 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 that region. Hence, our estimation procedure can be applied to a discontinuity point associated with gap, but not a discontinuity point associated 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 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.

[11] While θ is naturally interpreted as the level of patient sickness in our setting, the interpretation of θ would be context-specific. For instance, θ can be interpreted as *ability* in labor supply models and *strength of preference for the good* in individual consumption choice models.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

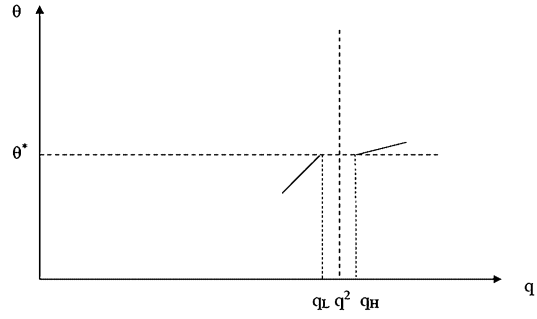

FIGURE 3. Behavioral responses at the kink.

Figure 3 graphically 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 that 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^* (but high enough to put them on the horizontal part of the reimbursement schedule), the hospital will choose q smaller than q_L and will face a marginal reimbursement rate of 0.

Figure 3 suggests two possible measures of the hospital's behavioral response. First, Figure 3 shows a possibility of gap at θ^*:

$$\phi_{\text{GAP}} = q_H - q_L.$$

The gap estimator ϕ_{GAP} can be used for evaluating the response to incentives, since a gap is generated due to the discrete change in the marginal reimbursement rate. To facilitate interpretation, we note that ϕ_{GAP} can be used to recover a discrete version of elasticity with respect to the reimbursement rate δ at θ^*. We can recover arc elasticity of q with respect to δ at θ^* from ϕ_{GAP} using the midpoint method as follows:

$$\text{arc elasticity of } q \text{ wrt } \delta \text{ at } \theta^* = \frac{q_H - q_L}{(q_H + q_L)/2} \Big/ \frac{\delta^2 - 0}{(\delta^2 + 0)/2} = \frac{\phi_{\text{GAP}}}{q_H + q_L}. \tag{12}$$

Second, Figure 3 also suggests the possibility of a change in the slope of the quantile function at θ^*, which we denote by ϕ_{SLOPE}.[12] To see how ϕ_{SLOPE} can capture incentive effects, note that

$$\phi_{\text{SLOPE}} = \frac{\partial q(\theta; \delta^2)}{\partial \theta}\Big|_{\theta=\theta^*} - \frac{\partial q(\theta; 0)}{\partial \theta}\Big|_{\theta=\theta^*},$$

[12]Note that in our setting the optimal choice of q as a function of θ, $q(\theta)$, is equivalent to the quantile function since type θ is assumed to be uniformly distributed on [0, 1]. If the true type follows a different distribution (e.g., normal distribution), we can still interpret θ as the percentile of the true type and $q'(\theta)$ would still be interpreted as change in q for an infinitesimal increase in (the percentile of) type. Thus assuming a uniform distribution for the type is simply an index normalization.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $q(\theta; \delta)$ denotes the optimal choice for a patient with type θ under a marginal reimbursement rate of δ. The slope estimator represents the change in the slope of the quantile function $\frac{\partial q}{\partial \theta}$ in response to a discrete change in marginal reimbursement rate from 0 to δ[2]. A continuous version of it, in the case of an infinitesimal change in the reimbursement rate δ, would be $\frac{\partial}{\partial \delta}$ $(\frac{\partial q}{\partial \theta})$. By changing the order of differentiation, we can interpret $\frac{\partial}{\partial \theta}$ $(\frac{\partial q}{\partial \delta})$ as capturing how the hospital's price sensitivity in health care provision $\frac{\partial q}{\partial \delta}$ changes with patient type θ, that is, heterogeneity in price sensitivity across marginally different patients. While the price change we exploit is discrete rather than continuous, the same intuition applies and we interpret the slope estimator as a discrete version of *how the hospital's quantity response to a change in the reimbursement rate differs across patients with marginally different levels of sickness*.

In other words, in our health care context the slope estimator tells us how the effects of cost sharing would be distributed across marginally different patients: A positive slope estimate means that when the marginal reimbursement rate goes down, health care spending goes down more for (marginally) sicker patients, which indicates that the effects of cost sharing would fall more heavily on sicker patients. Conversely, a negative slope estimate implies that hospitals would decrease their health care spending more heavily for (marginally) healthier patients in response to a drop in the reimbursement rate. A zero slope estimate implies that hospitals would decrease their health care spending equally for sicker and healthier patients in response to a drop in the reimbursement rate. Since theory does not predict the sign of ϕ_{SLOPE} one way or the other, it is an empirical question.

In general, ϕ_{SLOPE} is determined by three different factors. 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 curvatures of $\gamma v(q, \theta) - c(q)$ could differ between the two sides due to the changes in the reimbursement rates (when the reimbursement rate changes, the relationship between optimal q and θ changes, leading to changes in curvatures). Third, the curvatures of $\gamma v(q, \theta) - c(q)$ could differ between the two sides due to the inherent shape of the $v(q, \theta)$ and $c(q)$ functions. Since a gap, if any, is generated in the first place because of changes in reimbursement rates, the first two factors represent incentive effects, while the third factor does not. In this paper, we do not explicitly distinguish among these factors, because doing so would require more specific assumptions on the shape of $v(q, \theta)$ and $c(q)$ that are beyond the available data. Instead, so as to rule out the possibility that our estimate of ϕ_{SLOPE} mainly captures the effect of the inherent shape of $v(q, \theta)$ and $c(q)$, we run a placebo test using a control group where only the third factor is present. A lack of significance for the estimate of ϕ_{SLOPE} in the control group would suggest that our estimate of ϕ_{SLOPE} in the estimation sample reflects incentive effects.

It is worth discussing the relationship between the slope estimator ϕ_{SLOPE} and the gap estimator ϕ_{GAP}. A positive slope estimate is not a necessary implication of a positive gap estimate. If hospitals increase health care spending by the same amount for all patients in response to an increase in the reimbursement rate, we would have a positive gap estimate along with a zero slope estimate. Most likely, one would expect to find a significant gap estimate when the slope estimate is significant, since the model does not

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

predict heterogeneity in price sensitivity across patients when there is zero response for the focal patient.

The slope estimator and gap estimator capture two different dimensions of hospital behavior. The gap estimator captures price sensitivity of hospital's health care provision (for a patient with sickness level θ^*), and can be used to understand how effective cost sharing would be in bending health care costs. Clearly this is a very important measure and has been examined extensively in the prior literature. The slope estimator captures how the effects of cost sharing would be (locally) distributed across patients with (marginally) different levels of sickness, which in our view, is also an important measure for understanding the welfare effects of cost sharing policies (the biggest source of patient heterogeneity in health care settings is obviously the degree of illness).

Given the interpretations of the gap and slope estimators, it is straightforward to see how they can be compared with other approaches. The gap estimator measures price elasticity of health care spending, a commonly studied object in the literature. Examples include Hodgkin and McGuire (1994), Cutler (1995), Gaynor and Gertler (1995), Gruber and Owings (1996), Gaynor, Rebitzer, and Taylor (2004), Dafny (2005), and Ho and Pakes (2014).

The slope estimator measures heterogeneity in price elasticity of health care spending, and this is also an often studied object in the literature. For example, Kowalski (2016) examines heterogeneity in price responsiveness across the conditional expenditure distribution using quantile regression, and Einav, Finkelstein, and Schrimpf (2015) examine how changes in nonlinear contracts affect prescription drug spending of individuals at different points in the spending distribution. Dafny (2005) finds that responses to price changes were more aggressive among for-profit hospitals than not-for-profit hospitals. Lindrooth, Bazzoli, and Clement (2007) find that changes in treatment intensity in response to a reimbursement cut vary by Medicare share of the hospital, the level of reimbursement generosity for the diagnosis, and patient's illness severity. Ho and Pakes (2014) found some evidence that the sensitivity of physicians' care choices to financial incentives differs across patients with different illness severities.

A key identifying assumption in our approach is the smooth differentiability of the payoff function $\gamma v(q, \theta) - c(q)$ so 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 $q(\theta)$ 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 literature (Hahn, Todd, and Van der Klaauw (2001), McCrary (2008)).

One possible threat to this assumption is that q might not be continuous due to lumpiness in health care spending. At the margin the most lumpy decision is likely the patient's length of hospital stay, which can add \$6000 to charges for one extra day. However, most other medical decisions are much less lumpy. Testing is the best example. Each additional test is not expensive but many possible tests can be ordered on any given day. The assumption of continuous q is valid as long as the quantity choice is continuous at the margin, and we believe that the presence of discretionary and inexpensive components like testing makes this assumption reasonable.

1759733l, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Another key assumption is that there is a single unobservable that represents patient type. This is a very common assumption in the health economics literature, for example, used in Kowalski (2015) and Cardon and Hendel (2001) in their analyses of adverse selection and moral hazard. We do not view this as an overly restrictive assumption given that our patient populations are relatively homogeneous—all are severely ill and are undergoing the same major surgical procedure—and we are examining an outcome that is naturally unidimensional: total expenditures.[13]

A related requirement is that there be no error in the observed q. If q contains error, we cannot infer type θ from the observed q. In reality, this assumption could be violated either because hospitals cannot perfectly control the level of treatment—unforeseen events may make it more costly to treat a less sick patient—or because there is measurement error. Unfortunately, our estimator needs to assume away such possibilities.[14]

3.2 *Estimation methods*

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.[15] The first method is more robust since it does not rely on parametric assumptions, but will require a large amount of data per hospital. The second method depends on the validity of parametric assumptions, but does not require as much data per hospital. The first method is preferable but might not be feasible in certain applications, in which case the second approach can be used to improve finite sample inference.

3.2.1 *Individual hospital estimates* Consider a particular hospital. Suppose that there are $i = 1, \ldots, n$ individuals treated in the hospital under consideration. Let q_i denote the health care expenditure on individual i. Let $\hat{F}(\cdot)$ denote the empirical distribution of the observed qs for the hospital. The variation of q_i for this given hospital allows us to develop estimators for ϕ_{GAP} and ϕ_{SLOPE} at the upper discontinuity point of q^2. We also describe 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

[13]Generally speaking, while linear models can accommodate the linear addition of multiple error terms, allowing for multiple sources of unobserved heterogeneity in the presence of nonlinearity and a nonparametric response function is much more difficult. Many influential papers in the literature make use of scalar unobservables in nonlinear and nonparametric models (e.g., Matzkin (2003), Chernozhukov and Hansen (2005)).

[14]The assumption of no error in q may be less severe in typical consumer choice settings, such as residential electricity consumption or spending decisions on prescription drugs. Labor supply decisions for self-employed individuals are also likely to satisfy the condition, as those individuals have a high degree of discretion over their choice of work hours (and accordingly income).

[15]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 second method will pool information from multiple markets.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

discrete jump at the point of θ^* by an amount $\phi_{\text{GAP}} > 0$ to q_H. We are interested in estimating the magnitude of ϕ_{GAP}. This is estimated by $\hat{\phi}_{\text{GAP}} = \hat{q}_H - \hat{q}_L = \min\{q_i : q_i > q^2\} - \max\{q_i : q_i \le q^2\}$.

To derive the asymptotic distribution of $\hat{\phi}_{\text{GAP}} - \phi_{\text{GAP}}$, it suffices to show that the joint asymptotic distributions of $n(\hat{q}_L - q_L)$ and $n(\hat{q}_H - q_H)$ are independent exponential distributions. To see this, note that

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

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 of the distribution of q at q_L and q_H. The limiting distribution of $n(\hat{\phi}_{\text{GAP}} - \phi_{\text{GAP}})$ is therefore the sum of two independent exponential random variables.

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

Define

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

and

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

In the above equalities, $k(\cdot)$ 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 $nh^3 \longrightarrow 0$,

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

and

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

and that they are asymptotically independent. Therefore,

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

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The asymptotic variance above can be consistently estimated by replacing f^+ and f^- with the kernel estimates of $\hat{f}^+$ and $\hat{f}^-$.

3.2.2 *A parametric model using multiple hospital data* Now we consider how to extend the previous method to allow for pooling heterogeneous data across multiple hospitals. There are a few alternative estimation approaches one can potentially take. In this section we provide an overview.

Consider first $\phi_{\text{GAP}} = q_H - q_L$. We define $y_i = q_i 1(q_i \leq q^2)$ and $z_i = q_i 1(q_i > q^2) + M1(q_i \leq q^2)$, where M is a number that is larger than any of the data points. In the homogeneous case, we have defined $\hat{q}_L = \max\{y_i\}$ and $\hat{q}_H = \min\{z_i\}$.

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 that there are $t = 1, \ldots, T$ hospitals and that hospital heterogeneity is captured by covariates x_t, where x_t can include $q^2(t)$ itself and other contract terms as well as observable hospital characteristics. If all hospitals are homogeneous, x_t will include a constant only, and the gap and slope estimates would be identical for all hospitals.

Let I_t be the number of patient observations for hospital t. We specify the 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 equalities, 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 the nonstandard likelihood estimator (cf. Donald and Paarsch (1996), Chernozhukov and Hong (2004)). Each of these approaches has its advantages and disadvantages.

The extreme quantile regression approach of Chernozhukov (2005) 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 $g_H(x_t, \beta_H)$, and are also easy to implement. Given that we are interested in the shape of the distribution of q_i—specifically the slope parameter $\phi_{\text{SLOPE}}(x_t)$—in addition to $q_L(x_t)$ and $q_H(x_t)$, in the empirical estimation we adopt a maximum likelihood approach.

By adopting a parametric functional form on $q_L(x_t)$ and $q_H(x_t)$ we are maintaining a strong specification assumption that 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.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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

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

Conducting statistical inference on $\hat{\phi}_{\mathrm{GAP}}(x_t)$ 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 $n = \sum_t I_t$. The limiting distribution can be obtained by simulation, which we will describe below in the context of the parametric likelihood approach.

To describe the maximum likelihood approach, assume that

$$\varepsilon_{it}^L = g_L(x_t, \beta_L) - y_{it} \sim m_L\big(\varepsilon_{it}^L, x_t, \alpha_L\big) \quad \text{for } y_{it} \le g_L(x_t, \beta_L)$$

and

$$\varepsilon_{it}^H = z_{it} - g_H(x_t, \beta_H) \sim m_H\big(\varepsilon_{it}^H, x_t, \alpha_H\big) \quad \text{for } z_{it} \ge g_H(x_t, \beta_H) \text{ and } z_{it} < M.$$

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

$$(\hat{\alpha}_L, \hat{\beta}_L) = \arg\max_{\alpha_L, \beta_L} \sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(y_{it} > 0) \log m_L\big(g_L(x_t, \beta_L) - y_{it}, x_t, \alpha_L\big)$$
$$\text{such that} \quad y_{it} \le g_L(x_t, \beta_L)\ \forall i = 1, \ldots, I_t, t = 1, \ldots, T$$

and

$$(\hat{\alpha}_H, \hat{\beta}_H) = \arg\max_{\alpha_H, \beta_H} \sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(z_{it} < M) \log m_H\big(z_{it} - g_H(x_t, \beta_H), x_t, \alpha_H\big)$$
$$\text{such that } z_{it} \ge g_H(x_t, \beta_H)\ \forall i = 1, \ldots, I_t, t = 1, \ldots, T.$$

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

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

and

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

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Even though $\hat{\beta}_L$ and $\hat{\beta}_H$ converge at a $1/n$ 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.

We apply appropriate scaling to convert the conditional densities $m_L(\varepsilon_{it}^L, x_t, \hat{\alpha}_L)$ and $m_H(\varepsilon_{it}^H, x_t, \hat{\alpha}_H)$ into unconditional densities $f_L(\varepsilon_{it}^L, x_t, \hat{\alpha}_L)$ and $f_H(\varepsilon_{it}^H, x_t, \hat{\alpha}_H)$:

$$f_L\big(\varepsilon_{it}^L, x_t, \hat{\alpha}_L\big) = m_L\big(\varepsilon_{it}^L, x_t, \hat{\alpha}_L\big)\frac{\sum_{t=1}^{T}\sum_{i=1}^{I_t} 1(y_{it} > 0)}{n},$$

$$f_H\big(\varepsilon_{it}^H, x_t, \hat{\alpha}_H\big) = m_H\big(\varepsilon_{it}^H, x_t, \hat{\alpha}_H\big)\frac{\sum_{t=1}^{T}\sum_{i=1}^{I_t} 1(z_{it} < M)}{n}.$$

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

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

Since $\hat{\phi}_{\text{SLOPE}}(x_t)$ 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,

$$\tilde{\beta}_L = \arg\min_{\beta_L} \sum_{t=1}^{T}\sum_{i=1}^{I_t} m_L\big(0, x_t^*, \hat{\alpha}_L\big)\frac{\partial g_L\big(x_t^*, \hat{\beta}_L\big)'}{\partial \beta_L}\beta_L 1\big(y_{it}^* > 0\big)$$
$$\text{such that } y_{it}^* \le g_L\big(x_t^*, \beta_L\big) \ \forall i, t$$

and

$$\tilde{\beta}_H = \arg\max_{\beta_H} \sum_{t=1}^{T}\sum_{i=1}^{I_t} m_H\big(0, x_t^*, \hat{\alpha}_H\big)\frac{\partial g_H\big(x_t^*, \hat{\beta}_H\big)'}{\partial \beta_H}\beta_H 1\big(z_{it}^* < M^*\big)$$
$$\text{such that } z_{it}^* \ge g_H\big(x_t^*, \beta_H\big) \ \forall i, t,$$

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 $m_L(\varepsilon_{it}^L, x_t, \alpha_L)$ and $m_H(\varepsilon_{it}^H, x_t, \alpha_H)$ are misspecified. In this case, each of

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the alternative methods (linear and quadratic programming, 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 and 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 $\hat{\phi}_{\text{SLOPE}}$. Suppose x_t is continuously distributed with dimension $d = \dim(x)$. Let

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

and

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

where

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

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

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

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

and

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

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 $n(\hat{\beta}_H - \beta_H^0)$.[16]

[16]The validity of the simulated distributions of the programming estimators $n(\tilde{\beta}_L - \hat{\beta}_L)$, $n(\tilde{\beta}_H - \hat{\beta}_H)$, $n(\bar{\beta}_L - \hat{\beta}_L)$, and $n(\bar{\beta}_H - \hat{\beta}_H)$ is implied by the results in Bickel and Freedman (1981), Donald and Paarsch (2002), and Chernozhukov and Hong (2004).

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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. We have acquired hospital contracting data from the largest private insurer of organ and tissue transplants in the United States. The insurer contracts with 127 hospitals in the United States and we have data on the shape of the reimbursement schedule for all of these contracts. Consistent with our modeling framework, the majority of these contracts are piecewise linear with multiple kinks. 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 3 or 4 years.

In addition to the contract information, we also have 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) claims-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 major organ and tissue transplants, the most common being bone marrow transplant (BMT), kidney transplant, and liver transplant. Organ and tissue transplants are a rare and exceedingly expensive procedure. In 2007, 27,578 organs were transplanted in the United States. The average total billed charges for kidney transplantation in our data—the least expensive 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 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 (Department of Health and Human Services (2007)). Typically, the reimbursements made by the payer we study will comprise a significant portion of the transplant revenue of a hospital.

The insurer negotiates a separate contract with each individual hospital for each 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 with two kinks as shown in Figure 1 (the marginal reimbursement rate starts as a 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 as a positive and then remains at zero above a certain expenditure

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

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 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 on hospitals' bargaining power and markups).

In our analysis, we focus on hospitals whose reimbursement schedules display two kinks because our method is applicable to the second kink, which exhibits a gap, but not to the first kink, which exhibits bunching. As a result, our estimates are applicable only to hospitals that negotiate two-kink contracts with the insurer and are not necessarily generalized to hospitals whose contracts feature only one kink.

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.[18] The list prices are set well above marginal and average costs, and are generally determined based on expected costs plus a markup. Since the *charge master* (the file in which list prices are kept) does not vary by patient within a given hospital, changes in the charges within a given hospital reflect changes in quantity. On the contrary, it is well known that the charge master varies significantly across hospitals. Thus, we would observe different charges for two patients in two different hospitals even if they received the same level of treatment. In our analysis, the empirical distribution of q will be computed separately for each hospital so as to address this issue.

One might be concerned that sicker patients might incur lower charges because they 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 do 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 and kidney transplant patients from that data for our analysis ($N = 2980$). We estimate a simple logit model of the likelihood of death as a function of a polynomial of charges with hospital fixed effects for

[18]To be precise, q measures all expenses 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 followups. 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 during the covered period are included in q.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1. Summary statistics

	BMT	Kidney	Liver
Total number of patients	742	353	506
Total number of hospitals	14	11	13
Avg. charge per patient (in $1000)	168.2 (114)	140.6 (79.76)	311 (224.9)
Avg. reimbursement per patient (in $1000)	98.54 (67)	74.85 (40.46)	188.9 (131.7)
Avg. number of patients per hospital	53 (34.7)	32.09 (16.56)	38.92 (25.54)
Avg. q^1 across hospitals (in $1000)	115.9 (17.5)	78.21 (8.6)	186.2 (27.23)
Avg. q^2 across hospitals (in $1000)	159.2 (23.02)	125.3 (13.92)	254.3 (39.68)
Avg. δ^1 across hospitals	0.74 (0.06)	0.8 (0.07)	0.77 (0.06)
Avg. δ^2 across hospitals	0.54 (0.06)	0.5 (0.05)	0.56 (0.04)
Avg. % patients with $q < q^1$	33.73 (19.26)	7.31 (6.18)	14.72 (11.47)
Avg. % patients with $q^1 \leq q \leq q^2$	26.15 (14.21)	40.15 (15.82)	33.53 (13.64)
Avg. % patients with $q > q^2$	40.13 (19.16)	52.54 (13.19)	51.76 (17.92)

Note: Standard deviations are given in parentheses.

privately insured patients. 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.

Another potential concern is that the cutoff points q^1 and q^2 are chosen based on clinical criteria by which there is some standard amount of care that is provided and any care in excess of that base amount is likely to reflect provider agency, for example, through usage of very costly equipment, making the cutoff points endogenous. Our conversations with the insurer and transplant physicians indicate that transplant procedures do not follow such a cost structure. There is a significant degree of patient heterogeneity as the underlying health status of the patient varies widely. Thus, to the best of our understanding, the thresholds q^1 and q^2 are not chosen based on underlying technological features that are related to the agency problem.

One practical issue we encounter is that the number of patients who receive a certain type of organ transplant within a hospital is typically very 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 per hospital, we exclude from the sample (hospital, organ) pairs that have too few observations.[19] Even with this treatment, however, our sample size is quite small. While our data are unique and have the virtue of containing detailed information on contracts, we acknowledge that the small sample size due to the infrequency of the transplant procedures is the main weakness of our data. Table 1 presents summary statistics for our estimation sample.[20]

[19]In our result tables below we report the number of patients for each hospital.

[20]Since we focus on relatively large hospitals in Table 1, cross-hospital variations in marginal reimbursement rates and the locations of the kinks are smaller in Table 1 than in the original sample.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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, and so forth. Our data are essentially the information that the insurer receives from the hospital, and such detail is not transmitted to the insurer and is generally not available.

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, it is well known that hospitals can and do manipulate charges in response to 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, and so forth. Transplant surgeons we interviewed highlighted that there is significant variation in resource use that is attributable to testing and that many of these tests are discretionary and have ambiguous expected benefit. Another interesting example is that hospitals often contract with nearby hotels and step-down facilities and place transplant patients in the advanced stages of recovery in them instead of keeping the inpatient setting: the utilization of such facilities is often discretionary.

In Figure 4 we plot the empirical cdf of q (q on the x axis and cdf on the y axis) for all the hospitals in our estimation sample to illustrate the source of variation we use in estimation. The vertical line reflects the second discontinuity point q^2. To help with visualization, we also plot the lines that are fitted separately for each side of a window around q^2. For all the hospitals in the figure, the marginal reimbursement rate jumps from 0% to a positive number at q^2. The graphs are for 14 BMT hospitals, 11 kidney transplant hospitals, and 13 liver transplant hospitals. The figure suggests that the density function tends to become flatter after q^2 than before q^2 for the majority of hospitals, which is equivalent to the slope of the quantile function becoming steeper after q^2 than before q^2, leading to positive ϕ_{SLOPE}. The plot also suggests that there is no clear sign of gap for many hospital/organ combinations, an issue we will return to below.

5. Application: Results

To estimate the responsiveness of hospitals' health care provision to 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 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 the gap, $\hat{\phi}_{\text{GAP}}(x_t) = g_H(x_t, \hat{\beta}_H) - g_L(x_t, \hat{\beta}_L)$, and the change in the slope of the quantile function, $\hat{\phi}_{\text{SLOPE}}(x_t) = \frac{1}{f_H(0,x_t,\hat{\alpha}_H)} - \frac{1}{f_L(0,x_t,\hat{\alpha}_L)}$, at

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

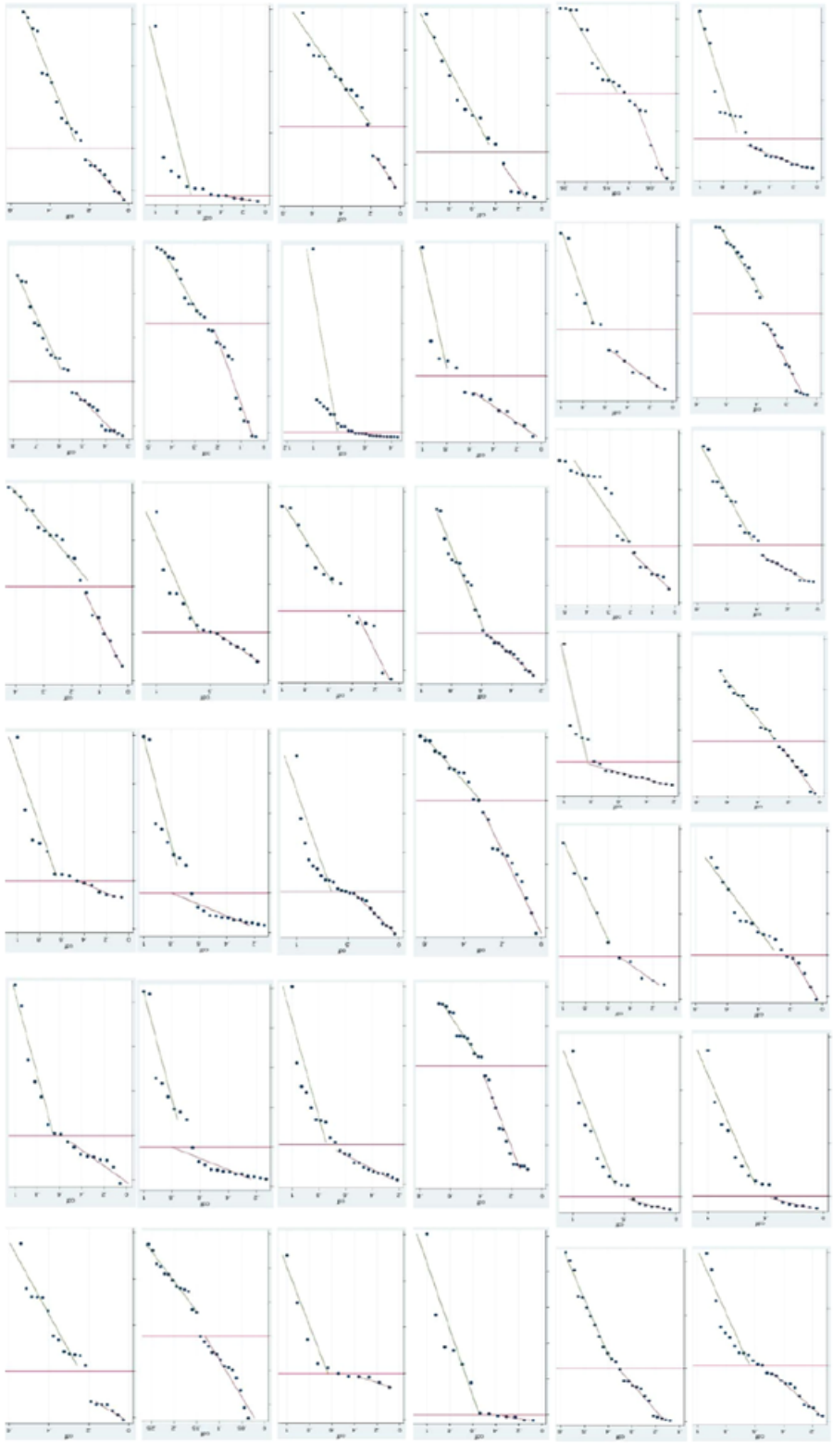

FIGURE 4. Changes in the empirical quantile function at q^2.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2. Maximum likelihood estimates for BMTs (q in \$1000)

	Obs.	δ^2	$\hat{\phi}_{\text{GAP}}$	$\frac{dq}{d\delta}\frac{\delta}{q}$	$\hat{\phi}_{\text{SLOPE}}$
H1	37	0.6	1.68 [−1.64, 4.77]	0.0055	199.09 (29.41)***
H2	17	0.5	0.52 [−3.29, 3.9]	0.0014	46.17 (28.56)
H3	15	0.55	1.37 [−1.64, 4.03]	0.0039	111.88 (19.97)***
H4	47	0.5	0.07 [−3.75, 2.59]	0.0002	26.76 (26.09)
H5	55	0.55	1.75 [−1.92, 5.2]	0.0048	45.29 (25.42)*
H6	41	0.65	−0.01 [−8, 31, 7.13]	−0.00004	302.91 (57.04)***
H7	113	0.6	1.31 [−0.62, 2.75]	0.0043	162.14 (18.79)***
H8	14	0.45	−0.56 [−5.25, 2.85]	−0.0018	−38.09 (33.73)
H9	23	0.55	1.45 [−1.2, 3.88]	0.0038	60.95 (18.23)***
H10	16	0.45	−0.88 [−9.93, 6.73]	−0.0023	−41.94 (58.87)
H11	58	0.5	0.2 [−3.5, 3.3]	0.0006	102.92 (23.25)***
H12	14	0.6	−0.02 [−5.01, 4.73]	−0.00008	197.55 (31.85)***
H13	11	0.5	−0.22 [−6.24, 5.05]	−0.0009	75.64 (36.69)**
H14	28	0.6	1.68 [−4.35, 7.22]	0.0056	19.76 (45.18)

Note: We report bootstrapped standard errors inside the parentheses for $\hat{\phi}_{\text{SLOPE}}$. Due to asymptotic nonnormality of gap estimates, we report the 95% confidence interval instead of standard errors. In particular, the reported lower and upper bounds are computed as 0.025 and 0.975 quantiles of $2\hat{\phi}_{\text{GAP}} - \hat{\phi}_{\text{GAP\_}b}$, where $\hat{\phi}_{\text{GAP}}$ is the sample estimate and $\hat{\phi}_{\text{GAP\_}b}$ is the bth bootstrap estimate. We apply a similar bias correction to the estimates themselves as well. The reported numbers under $\hat{\phi}_{\text{SLOPE}}$ correspond to a 1 unit increase in θ (= 100 percentile increase in sickness), and q is in \$1000. Thus, we need to multiply the reported numbers by 10 so as to compute the change in dollar spending corresponding to 1 percentile increase in sickness. *Significant at 10%; **significant at 5%; ***significant at 1%.

the discontinuity point for each hospital characterized by x_t. We use exponential distribution for conditional densities m_L and m_H with hazard rate parameters $\lambda_L(x_t, \alpha_L)$ and $\lambda_H(x_t, \alpha_H)$, respectively. To compute standard errors, we use parametric bootstrap using 500 simulations. We apply the MLE to each organ type separately.

In the data section, we noted that the substantial variation in the contract type and contract terms across hospitals likely reflects variation in bargaining power as well as heterogeneity in the patient pool across hospitals. In the current parametric approach, patient heterogeneity across hospitals is captured by the linear regression function where the contract terms, 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), enter as regressors. Additionally including other observable hospital characteristics in x_t, such as teaching hospital status, may offer a more refined control of patient heterogeneity across hospitals, but we are not able to implement it empirically due to data limitations. Further, a structural model of how patients self-select into different hospitals is beyond our current scope.

In Tables 2–4, we report the maximum likelihood estimates of ϕ_{GAP} and ϕ_{SLOPE} for each hospital. We report δ^2 and the number of patients above the first discontinuity point 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 the estimation is the sum of observations across hospitals in a given organ type, which is much higher than that indicated by each individual hospital. Table 2 reports estimates for BMT, Table 3 for kidney transplants, and Table 4 for liver transplants. In addition to reporting the estimates of ϕ_{GAP} and ϕ_{SLOPE}, we also report the arc elasticities recovered from the estimates of ϕ_{GAP}, as defined in (12).

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3. Maximum likelihood estimates for kidney transplants (q in \$1000)

	Obs.	δ^2	$\hat{\phi}_{\text{GAP}}$	$\frac{dq}{d\delta}\frac{\delta}{q}$	$\hat{\phi}_{\text{SLOPE}}$
H1	25	0.55	−1.65 [−7.23, 3.47]	−0.0067	26.92 (14.65)*
H2	14	0.45	1.13 [−3.29, 5.34]	0.0048	65.16 (20.54)***
H3	37	0.45	4.07 [−5.29, 12.29]	0.0162	86.52 (34.12)**
H4	27	0.5	0.41 [−2.58, 2.2]	0.0019	61.89 (12.03)***
H5	13	0.55	−1.98 [−5.89, 1.96]	−0.0085	54.73 (13.54)***
H6	43	0.49	−0.56 [−3.35, 2.26]	−0.0021	56.65 (14.25)***
H7	40	0.49	−0.25 [−2.56, 1.32]	−0.0009	50.56 (12.62)***
H8	39	0.49	0.05 [−3.12, 2.61]	0.0002	44.33 (16.29)***
H9	14	0.49	0.31 [−4.35, 4.27]	0.0011	38.79 (21.75)*
H10	15	0.45	0.81 [−4.38, 5.64]	0.0036	63.03 (22.34)***
H11	57	0.6	−0.81 [−5.59, 1.27]	−0.0037	119.02 (24.38)***

Note: We report bootstrapped standard errors inside the parentheses for $\hat{\phi}_{\text{SLOPE}}$. Due to asymptotic nonnormality of gap estimates, we report the 95% confidence interval instead of standard errors. In particular, the reported lower and upper bounds are computed as 0.025 and 0.975 quantiles of $2\hat{\phi}_{\text{GAP}} - \hat{\phi}_{\text{GAP}\_b}$, where $\hat{\phi}_{\text{GAP}}$ is the sample estimate and $\hat{\phi}_{\text{GAP}\_b}$ is the bth bootstrap estimate. We apply a similar bias correction to the estimates themselves as well. The reported numbers under $\hat{\phi}_{\text{SLOPE}}$ correspond to a 1 unit increase in θ (= 100 percentile increase in sickness), and q is in \$1000. Thus, we need to multiply the reported numbers by 10 so as to compute the change in dollar spending corresponding to 1 percentile increase in sickness. *Significant at 10%; **significant at 5%; ***significant at 1%.

TABLE 4. Maximum likelihood estimates for liver transplants (q in \$1000)

	Obs.	δ^2	$\hat{\phi}_{\text{GAP}}$	$\frac{dq}{d\delta}\frac{\delta}{q}$	$\hat{\phi}_{\text{SLOPE}}$
H1	17	0.55	−2.32 [−17.83, 6.21]	−0.0042	282.83 (66.69)***
H2	10	0.55	−1.32 [−11.82, 4.6]	−0.0023	180.89 (47.41)***
H3	23	0.55	−2.72 [−57.85, 52.55]	−0.0044	44.76 (119.16)
H4	38	0.55	2.83 [−11.53, 9.39]	0.0087	175.49 (49.35)***
H5	14	0.55	28.41 [−13.03, 36.37]	0.0492	0.82 (133.87)
H6	81	0.65	−1.57 [−8.95, 1.96]	−0.0038	271.92 (36.91)***
H7	31	0.49	6.29 [−16.85, 12.88]	0.0111	0.66 (84.07)
H8	21	0.55	0.76 [−30.37, 24.25]	0.0018	260.29 (88.9)***
H9	27	0.55	−0.97 [−11.6, 5.82]	−0.0018	206.54 (47.95)***
H10	34	0.55	0.31 [−14.33, 7.13]	0.0006	271.71 (65.31)***
H11	33	0.55	−1.26 [−10.93, 3.76]	−0.0024	185.81 (46.72)***
H12	82	0.62	2 [−8.27, 6.11]	0.004	246.55 (47.05)***
H13	23	0.62	17.63 [−23.95, 31.67]	0.0354	85.58 (72.03)

Note: We report bootstrapped standard errors inside the parentheses for $\hat{\phi}_{\text{SLOPE}}$. Due to asymptotic nonnormality of gap estimates, we report the 95% confidence interval instead of standard errors. In particular, the reported lower and upper bounds are computed as 0.025 and 0.975 quantiles of $2\hat{\phi}_{\text{GAP}} - \hat{\phi}_{\text{GAP}\_b}$, where $\hat{\phi}_{\text{GAP}}$ is the sample estimate and $\hat{\phi}_{\text{GAP}\_b}$ is the bth bootstrap estimate. We apply a similar bias correction to the estimates themselves as well. The reported numbers under $\hat{\phi}_{\text{SLOPE}}$ correspond to a 1 unit increase in θ (= 100 percentile increase in sickness), and q is in \$1000. Thus, we need to multiply the reported numbers by 10 so as to compute the change in dollar spending corresponding to 1 percentile increase in sickness. *Significant at 10%; **significant at 5%; ***significant at 1%.

From the results in Tables 2–4, we see that $\hat{\phi}_{\text{SLOPE}}$ is positive and statistically significant for 29 out of 38 cases (76.3%). When the hospitals have to bear a larger fraction of the additional expenditures for patient treatment, health care spending tends to go down more for (marginally) sicker patients, indicating that the effects of cost shar-

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ing would fall more heavily on sicker patients. To interpret the magnitude of the coefficients, take the result for hospital 1 in Table 2. In response to a drop in the marginal reimbursement rate from 60% to 0%, the drop in health care spending is greater by \$1990 for marginally sicker patients (specifically, for a 1 percentile increase in illness severity). Similarly, take the result for hospital 1 in Table 3.[21] In response to a drop in the marginal reimbursement rate from 55% to 0%, the drop in health care spending is greater by \$269 for patients with 1 percentile higher illness severity. Similarly, for hospital 1 in Table 4, the change in health care spending in response to a drop in the marginal reimbursement rate from 55% to 0% is greater by \$2828 for patients with 1 percentile higher illness severity. Overall, the change in health care spending in response to a 45–65 percentage points drop in the marginal reimbursement rate is greater by \$269–\$2828 for patients with 1 percentile higher illness severity across all hospitals and organ types with significant slope estimates.[22]

Another pattern we observe in Tables 2–4 is that $\hat{\phi}_{\text{GAP}}$ is not statistically different from zero in all cases. As a result, the corresponding elasticity of health expenditures with respect to the marginal reimbursement rate is not statistically distinguishable from zero. As discussed earlier, it seems unlikely that there exists heterogeneity in price sensitivity across patients (significant slope estimate), while there is zero response for the focal patient (insignificant gap estimate). Further, the prior literature has found a nonzero price elasticity among health care providers (Gaynor and Gertler (1995), Dafny (2005), Ho and Pakes (2014)). We will return to this issue of insignificant $\hat{\phi}_{\text{GAP}}$ in the next section.

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 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 slope estimates tell us that the average density on the right-hand side (RHS) is smaller than the average density on the left-hand side (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 (after the first

[21] The kth hospital in Table 2 is different from the kth hospital in Tables 3 or 4.

[22] To see whether there is any systematic difference in the slope estimates across transplant types, we regressed the slope estimates (from Tables 2–4) on δ^2 and transplant type dummies (we do not examine the gap estimates since they are mostly insignificant). The analysis, available from the authors on request, reveals that, as expected, higher δ^2 is associated with a larger slope estimate. The analysis also shows that the slope estimates for liver transplants are greater than those for BMT or kidney transplants. However, this effect goes away once we control for the fact that liver transplants are the most expensive among the three by including the average q for each transplant and hospital combination in the regression. Therefore, the evidence from the data seems to suggest that the degree of heterogeneity in price sensitivity across patients does not differ much across different types of transplants.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5. Maximum likelihood estimates for the control group (q in \$1000)

	Obs.	$\hat{\phi}_{\text{GAP}}$	$\frac{dq}{d\delta}\frac{\delta}{q}$	$\hat{\phi}_{\text{SLOPE}}$
H1 (BMT)	28	5.68 [−3.83, 10.36]	0.0177	3.35 (34.16)
H1 (kidney)	10	8 [4.85, 9.81]**	0.0303	3.84 (13.86)
H2 (kidney)	14	7.86 [4.73, 9.9]**	0.0298	11.49 (14.55)
H3 (kidney)	33	0.45 [−1.5, 1.78]	0.0017	68.75 (11.18)***
H4 (kidney)	16	−2.5 [−4.92, −0.75]**	−0.0098	95.58 (16.13)***
H5 (kidney)	16	1.29 [−3.68, 3.63]	0.005	14.86 (13.33)
H6 (kidney)	30	0.01 [−6.36, 3.3]	0.00004	14.03 (16.2)
H7 (kidney)	14	−0.79 [−2.84, 0.39]	−0.0031	62.65 (10.52)***
H8 (kidney)	18	8.19 [4.82, 9.91]**	0.0321	−5.72 (14.16)
H9 (kidney)	13	7.72 [4.39, 10.49]**	0.0292	19.63 (16.28)
H10 (kidney)	39	−2.51 [−4.99, −0.71]**	−0.0097	97.09 (16.55)***
H1 (liver)	18	16.48 [−25.4, 32.6]	0.0353	261.77 (97.18)***
H2 (liver)	47	−2.85 [−17.13, 3.46]	−0.0057	306.04 (66.93)***

kink, the marginal reimbursement rate is always zero). We then impose an artificial cutoff point, equal to the average q^2 among our estimation sample hospitals, and perform similar analysis as in Tables 2–4. If our earlier results are an artifact of, for example, 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 5.

A comparison of Table 5 against Tables 2–4 suggests that our earlier results were at least partially reflective of hospitals' true behavioral responses to financial incentives. In Tables 2–4 we saw that the estimates of ϕ_{SLOPE} were positive and statistically significant for 29 out of 38 hospitals (76%) in our estimation sample, with a mean value of 109.3. In contrast, we see that only 6 out of 13 hospitals in the control group have significant $\hat{\phi}_{\text{SLOPE}}$ (46%), with a mean value of 73.3. These results from the *control* group help partially alleviate concerns that our findings may be mainly an artifact of concavity of the density function of spending around q^2 for reasons other than financial incentives. At the same time, however, the nonzero results from the control group suggest that our estimates in Tables 2–4 partly capture effects unrelated to financial incentives, so this is a caveat in interpreting the magnitudes of our slope estimates.

To test the robustness of our results against parametric assumptions under MLE, we perform our analysis at a finer level of aggregation: at the individual hospital level. In this second set of results, we apply the estimators discussed in Section 3.2.1 to estimate ϕ_{SLOPE} and ϕ_{GAP} 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. In our slope estimates, half-normal kernels were used to construct the weights. We use Silverman's plug-in estimates for bandwidths. Table 6 reports estimates of ϕ_{GAP} and ϕ_{SLOPE} for each hospital and each organ type.[23] As a placebo test, we again redo the esti-

[23]Since estimation is done separately for each hospital, only hospitals with sufficient numbers of observations are used in Tables 6 and 7. As a result, the number of hospitals reported in Table 6 is smaller than

1759733l, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6. Kernel estimates (q in \$1000)

	Obs.	δ^2	$\hat{\phi}_{\text{GAP}}$	$\frac{dq}{d\delta}\frac{\delta}{q}$	$\hat{\phi}_{\text{SLOPE}}$
H5 (BMT)	55	0.55	5.13 (0.142)	0.0141	78.14 (38.18)**
H6 (BMT)	41	0.65	4.7 (0.334)	0.019	187.2 (59.1)***
H7 (BMT)	113	0.6	7.1 (0.048)**	0.0231	51.25 (74.17)
H11 (BMT)	58	0.5	3.44 (0.461)	0.0101	18.54 (49.88)
H14 (BMT)	28	0.6	10.27 (0.455)	0.0339	238.6 (197.8)
H1 (kidney)	25	0.55	0.78 (0.892)	0.0032	3.8 (19.99)
H3 (kidney)	37	0.45	6.69 (0.405)	0.0268	183.4 (85.78)**
H6 (kidney)	43	0.49	4.67 (0.363)	0.0179	−20.47 (55.5)
H7 (kidney)	40	0.49	3.8 (0.563)	0.014	−39.68 (71.5)
H8 (kidney)	39	0.49	1.95 (0.808)	0.0067	114.1 (60.87)*
H11 (kidney)	57	0.6	1.44 (0.688)	0.0066	66.27 (31.47)**
H7 (liver)	31	0.49	13.99 (0.383)	0.0247	136.1 (92.41)
H8 (liver)	21	0.55	10.01 (0.708)	0.0233	325.7 (140.3)**
H10 (liver)	34	0.55	7.77 (0.566)	0.0153	86.09 (85.85)
H11 (liver)	33	0.55	26.02 (0.164)	0.0491	279.2 (161.4)*
H12 (liver)	82	0.62	6.54 (0.15)	0.013	76.78 (49.74)
H13 (liver)	23	0.62	33.3 (0.235)	0.0668	318.5 (173.6)*

Note: The p-values for $\hat{\phi}_{\text{GAP}}$ and standard errors for $\hat{\phi}_{\text{SLOPE}}$ are given in parentheses.

TABLE 7. Kernel estimates for the *control* group (q in \$1000)

	Obs.	$\hat{\phi}_{\text{GAP}}$	$\frac{dq}{d\delta}\frac{\delta}{q}$	$\hat{\phi}_{\text{SLOPE}}$
H3 (kidney)	33	1.51 (0.772)	0.0059	39.7 (31.85)
H10 (kidney)	39	4.95 (0.119)	0.0191	14.99 (27.66)
H2 (liver)	47	4.24 (0.619)	0.0085	124.6 (79.83)

Note: The p-values for $\hat{\phi}_{\text{GAP}}$ and standard errors for $\hat{\phi}_{\text{SLOPE}}$ are given in parentheses.

mation for each individual hospital in the control group, as discussed earlier. Estimation results for this control group of hospitals are reported in Table 7.

A comparison of Table 6 against Table 7 again suggests that our results are likely reflective of hospitals' true behavioral responses to financial incentives. None of the slope estimates is statistically significant for the control group in Table 7 (with a mean value of 59.7), while the estimates of ϕ_{SLOPE} are positive and statistically significant for about half of the hospitals in our estimation sample in Table 6 (with a mean value of 123.7). The gap estimates are mostly insignificant in both samples, again similar to the MLE results. While the magnitudes and significance differ, the fact that our global estimator (Tables 2–5) and local estimator (Tables 6 and 7) lead to similar conclusions is reassuring.

The overall results suggest that the impact of the reimbursement rate on hospitals' provision of health care services differs across patients. Although the lack of informa-

those in Tables 2–4, and similarly for Table 7. In Tables 6 and 7, we report the number of patients used in estimation for each hospital.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

tion on the components of the final charges prevents us from examining whether this is mainly due to underprovision for sicker patients (relative to healthier patients) below the threshold (necessary care is withheld) or overprovision for sicker patients (relative to healthier patients) above the threshold (unnecessary care is provided), the finding that the hospitals' sensitivity to financial incentives differs across patients provides insights on the distribution of effects under cost-sharing policies.

6. Discussion

In this section, we discuss other settings to which our estimator can be applied and how our proposed methodology is related to existing methods. We also discuss the lack of significance for the estimates of ϕ_{GAP}. 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 price is equivalent to a sudden jump in marginal reimbursement rate in our model in 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 (a common assumption 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 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. 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 the marginal reimbursement rate jumps. In a related work, an empirical application of our method is found in Marsh (2014), who employs a variant of our estimator in the case of health savings accounts to infer the sensitivity of patients' expenditures to price.

We view our work as complementary to the work by Saez (2010) and Chetty et al. (2011), who use the size of bunching to infer the sensitivity of labor supply to marginal tax rates. In the 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 Figure 4. We, on the other hand, propose to use the size of the 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

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 slope—gave somewhat different answers: we found no evidence of significant gap, but the slope estimate was statistically significant. As we discussed earlier, finding significant slope estimates along with insignificant gap estimates seems unlikely. Furthermore, the prior literature has found a nonzero response by health care providers to financial incentives (Gruber and Owings (1996), Dafny (2005), Ho and Pakes (2014)). Thus, it is puzzling that our gap estimates are insignificant.

Although we cannot provide a definitive answer on why this is the case, we think one possibility might be the presence of optimization error or measurement error. It is well known that estimators based on order statistics are sensitive to outliers. In the extreme case, the gap can entirely disappear if the hospital behaves suboptimally for just two patients, while its impact on the slope estimator would be much smaller. Thus, the insignificant gap estimates could be due to such optimization error. Further, simulation exercises we conducted to investigate the performance of the gap estimator under optimization error suggest that even a small degree of optimization error could make the gap estimates insignificant when combined with a small sample size.[24]

We considered the possibility of extending the gap estimator to make it robust to a certain degree of optimization error. The general idea is that instead of a complete absence of data (i.e., zero density) in an interval, one might use a lower density within an interval compared to nearby areas to identify a *gap*. A similar idea was used in the stochastic frontier production function literature, but most of the literature relies on distributional assumptions on the error. Accounting for optimization error nonparametrically is very difficult (e.g., Kneip, Simar, and Van Keilegom (2015)), and the challenge is even greater in our setting since the presence of two boundaries—instead of one boundary in the case of stochastic frontier production function—creates an issue of how to classify each observation to either the lower boundary or the upper boundary. Given the difficulty encountered in the frontier production function literature, an extension of the gap estimator to make it less sensitive to optimization error without making strong parametric assumptions is beyond the scope of this paper and we plan to investigate the issue further in future work.

7. Conclusion

In this paper, we propose an estimator that exploits discontinuity without bunching in a nonlinear pricing schedule to recover the 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

[24]We discuss the simulation results in the Appendix.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

our estimator to contracts in the health care market reveals that the impact of financial incentives on hospitals' health care spending significantly differs across patients.

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.

Appendix

We conduct simulations to investigate the performance of the gap estimator under optimization error or measurement error. Using the specifications of the payoff function and reimbursement schedule provided in Section 2, we calculate the optimal health care level for each patient, whose type θ is drawn uniformly from $[0, 100]$. We then assume that the observed choice of health care for a patient differs from the patient's optimal choice either due to optimization error or measurement error. In the investigation, we vary the sample size as well as the magnitude and the nature of the optimization/measurement error to see the impact of those factors on the performance of the gap estimator. Since we simulate data for one hospital with a sufficiently large sample size, there is no need to pool data across hospitals. Accordingly, we use the estimation method proposed in Section 3.2.1, which computes the gap estimate using sample maximum and minimum.

For the simulations, we choose the following parameter values, under which the resulting utility function satisfies assumptions (2)–(7):

$$\begin{aligned} u(q, \theta) &= 5\big(\theta q^{0.1} - 20\theta\big) - q + r(q), \\ r(0) &= 0, \\ r'(q) &= 0.2 \quad \text{for } 0 < q < 30, \\ r'(q) &= 0 \quad \text{for } 30 \le q \le 50, \\ r'(q) &= 0.1 \quad \text{for } q > 50. \end{aligned}$$

Under the chosen parameter values, the true gap size is 5.86. In Table A1, we summarize the estimation results obtained using the simulated data. The reported estimates are the average of the gap estimates over 500 simulated data sets. We report the average of the associated p-values over the 500 simulated data sets in parentheses. Since the gap estimates based on sample minimum and maximum are always positive, the 95% confidence interval constructed using the empirical distribution of the gap estimates always lies above 0, thus making it impossible to reject the null of zero gap. Thus we report the average p-values instead of the 95% confidence interval.

In Scenario 0, the observed q for each patient is equal to the optimal q ($\hat{q} =$ optimal q).

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A1. Gap estimates using simulated data

	$n = 5000$	$n = 1000$	$n = 500$	$n = 100$
Scenario 0	5.897 (0)***	6.035 (0)***	6.197 (0)***	7.555 (0.011)**
Scenario 1	3.87 (0)***	4.513 (0)***	4.989 (0)***	7.166 (0.035)**
Scenario 2	1.606 (0)***	2.454 (0.001)***	3.18 (0.005)***	6.003 (0.151)
Scenario 3	0.155 (0.334)	0.734 (0.277)	1.454 (0.237)	4.708 (0.339)
Scenario 4	0.084 (0.463)	0.408 (0.446)	0.864 (0.406)	3.712 (0.449)
Scenario 5	1.443 (0.045)**	3.94 (0.009)***	4.905 (0.011)**	7.22 (0.122)
Scenario 6	0.804 (0.11)	3.185 (0.031)**	4.301 (0.03)**	6.94 (0.151)

Note: ***Significant at 1% level; **significant at 5% level; *significant at 10% level.

In Scenario 1, the observed q for each patient is subject to error whose magnitude is up to 2.5% of the optimal q ($\hat{q} =$ optimal $q \times (1 + 0.025 \times \text{uniform}[-1, 1])$).

In Scenario 2, the observed q for each patient is subject to error whose magnitude is up to 5% of the optimal q ($\hat{q} =$ optimal $q \times (1 + 0.05 \times \text{uniform}[-1, 1])$).

In Scenario 3, the observed q for each patient is subject to error whose magnitude is up to 7.5% of the optimal q ($\hat{q} =$ optimal $q \times (1 + 0.075 \times \text{uniform}[-1, 1])$).

In Scenario 4, the observed q for each patient is subject to error whose magnitude is up to 10% of the optimal q ($\hat{q} =$ optimal $q \times (1 + 0.1 \times \text{uniform}[-1, 1])$).

In Scenario 5, for 10% of patients the observed q is subject to error whose magnitude is up to 7.5% of the optimal q, and for the remaining 90% of patients, the observed q is equal to the optimal q.

In Scenario 6, for 10% of patients the observed q is subject to error whose magnitude is up to 10% of the optimal q, and for the remaining 90% of patients, the observed q is equal to the optimal q.

A summary of the simulation results follows. First we find that the performance of the gap estimator, not surprisingly, depends on the nature and extent of error as well as the sample size. If everybody is subject to optimization (or measurement) error, gap estimates start to lose significance when the degree of optimization error reaches a nontrivial level (e.g., Scenarios 3 and 4 with $n = 5000$). When only a fraction of agents are subject to optimization error, gap estimates often retain statistical significance even when the extent of error for those agents is not negligible (e.g., Scenario 5 with $n = 5000$). Since statistical significance takes into account the impact of sample size, even seemingly small gap estimates are statistically significant with a large sample (e.g., Scenario 2 with $n = 5000$).

The magnitude of gap estimates is a different story. Gap estimates quickly become small in magnitude when error is not negligible. With a sufficiently large sample size, the gap estimates, while still statistically different from zero, almost always suffer from downward bias, with the size of downward bias increasing with the magnitude of optimization error.

We also note that with a small sample ($n = 100$), gap estimates are statistically indistinguishable from zero even with quite small optimization error (Scenario 2). The gap estimate is statistically different from zero under sample $n = 100$ only when there is no or very little optimization error (Scenarios 0 and 1).

1759733l, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In our empirical analysis, the sample size is quite small and the gap estimates are mostly insignificant. Thus, the lack of statistical significance in the gap estimates in our empirical analysis could be due to a combination of a small sample size and some degree of optimization or measurement error.

The overall message from the simulations suggests that the gap estimator is sensitive to optimization error or measurement error, which is inherent in an estimator that is based on extremal points. The results also show that adequate performance of the gap estimator requires a reasonable sample size. For reasonably large samples, the estimated gap in the presence of optimization or measurement error would be a lower bound for the true gap, while the estimated gap might overestimate the true gap if the sample size is very small (although most likely those will be statistically indistinguishable from zero due to large standard errors).

References

Arrow, K. (1963), "Uncertainty and the welfare economics of medical care." *American Economic Review*, 53, 941–973. [398]

Bickel, P. and D. Freedman (1981), "Some asymptotic theory for the bootstrap." *The Annals of Statistics*, 9 (6), 1196–1217. [415]

Bresnahan, T. (1987), "Competition and collusion in the American automobile industry: The 1955 price war." *Journal of Industrial Economics*, 35 (4), 457–482. [405]

Burtless, G. and J. Hausman (1978), "The effect of taxation on labor supply: Evaluating the Gary income maintenance experiment." *Journal of Political Economy*, 86, 1103–1130. [400]

Cardon, J. and I. Hendel (2001), "Asymmetric information in health insurance: Evidence from the national medical expenditure survey." *RAND Journal of Economics*, 32 (3), 408–427. [410]

Chandra, A., D. Cutler, and Z. Song (2012), "Who ordered that? The economics of treatment choices in medical care." In *Handbook of Health Economics*, Vol. 2 (M. Pauly, T. McGuire, and P. P. Barros, eds.), 397–432, North-Holland. [398]

Chernozhukov, V. (2005), "Extremal quantile regression." *The Annals of Statistics*, 33 (2), 806–839. [412]

Chernozhukov, V. and C. Hansen (2005), "An IV model of quantile treatment effects." *Econometrica*, 73 (1), 245–261. [410]

Chernozhukov, V. and H. Hong (2004), "Likelihood estimation and inference in a class of nonregular econometric models." *Econometrica*, 72 (5), 1445–1480. [412, 415]

Chetty, R., J. Friedman, T. Olsen, and L. 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. [400, 426]

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Copeland, A. and C. Monnet (2009), "The welfare effects of incentive schemes." *Review of Economic Studies*, 76, 93–113. [400]

Cutler, D. (1995), "The incidence of adverse medical outcomes under prospective payment." *Econometrica*, 63 (1), 29–50. [398, 409]

Dafny, L. (2005), "How do hospitals respond to price changes?" *American Economic Review*, 95 (5), 1525–1547. [398, 409, 423, 427]

Dalton, C., G. Gowrisankaran, and R. Town (2015), "Myopia and complex dynamic incentives: Evidence from medicare part D." Working Paper. [400]

Department of Health and Human Services (2007), "Organ procurement and transplantation network and scientific registry of transplant recipients, Annual data report", Rockville, MD: Health Resources and Services Administration. [416]

Donald, S. and H. Paarsch (1996), "Identification, estimation, and testing in parametric empirical models of auctions within independent private values paradigm." *Econometric Theory*, 12, 517–567. [412]

Donald, S. and H. Paarsch (2002), "Superconsistent estimation and inference in structural econometric models using extreme order statistics." *Journal of Econometrics*, 109 (2), 305–340. [415]

Dranove, D. and P. Wehner (1994), "Physician-induced demand for childbirths." *Journal of Health Economics*, 13 (1), 61–73. [398]

Einav, L., A. Finkelstein, and P. Schrimpf (2015), "The response of drug expenditure to non-linear contract design: Evidence from medicare part D." *Quarterly Journal of Economics*, 130 (2), 841–899. [400, 401, 409]

Einav, L., A. Finkelstein, and P. Schrimpf (2017), "Bunching at the kink: Implications for spending responses to health insurance contracts." *Journal of Public Economics*, 146, 27–40. [400, 401]

Gaynor, M. and P. Gertler (1995), "Moral hazard and risk spreading in partnerships." *RAND Journal of Economics*, 26 (4), 591–613. [398, 409, 423]

Gaynor, M., J. Rebitzer, and L. Taylor (2004), "Physician incentives in health maintenance organizations." *Journal of Political Economy*, 112 (4), 915–931. [398, 409]

Gowrisankaran, G., A. Nevo, and R. Town (2015), "Mergers when prices are negotiated: Evidence from the hospital industry." *American Economic Review*, 175, 172–203. [401]

Gruber, J. and M. Owings (1996), "Physician financial incentives and cesarean section delivery." *RAND Journal of Economics*, 27 (1), 99–123. [398, 409, 427]

Hahn, J., P. Todd, and W. Van der Klaauw (2001), "Identification and estimation of treatment effects with a regression-discontinuity design." *Econometrica*, 69 (1), 201–209. [409]

Hausman, J. (1979), "The econometrics of labor supply on convex budget sets." *Economics Letters*, 3, 171–174. [400]

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hausman, J. (1985), "The econometrics of nonlinear budget sets." *Econometrica*, 53 (6), 1255–1282. [400]

Ho, K. (2009), "Insurer-provider networks in the medical care market." *American Economic Review*, 99 (1), 393–430. [417]

Ho, K. and A. Pakes (2014), "Hospital choices, hospital prices and financial incentives to physicians." *American Economic Review*, 104 (12), 3841–3884. [398, 409, 423, 427]

Hodgkin, D. and T. McGuire (1994), "Payment levels and hospital response to prospective payment." *Journal of Health Economics*, 13, 1–29. [398, 409]

Ketcham, J., P. Léger, and C. Lucarelli (2011), "Standardization under group incentives." Working Paper. [398]

Kneip, A., L. Simar, and I. Van Keilegom (2015), "Frontier estimation in the presence of measurement error with unknown variance." *Journal of Econometrics*, 184, 379–393. [427]

Kowalski, A. (2015), "Estimating the tradeoff between risk protection and moral hazard with a nonlinear budget set model of health insurance." *International Journal of Industrial Organization*, 43, 122–135. [400, 410]

Kowalski, A. (2016), "Censored quantile instrumental variable estimates of the price elasticity of expenditure on medical care." *Journal of Business and Economic Statistics*, 34 (1), 107–117. [409]

Lindrooth, R., G. Bazzoli, and J. Clement (2007), "The effect of reimbursement on the intensity of hospital services." *Southern Economic Journal*, 73 (3), 575–587. [409]

Marsh, C. (2014), "Estimating demand elasticities using nonlinear pricing." *International Journal of Industrial Organization*, 37, 178–191. [426]

Matzkin, R. L. (2003), "Nonparametric estimation of nonadditive random functions." *Econometrica*, 71 (5), 1339–1375. [410]

McClellan, M. (2011), "Reforming payments to healthcare providers: The key to slowing healthcare cost growth while improving quality?" *Journal of Economic Perspectives*, 25 (2), 69–92. [398]

McCrary, J. (2008), "Manipulation of the running variable in the regression discontinuity design: A density test." *Journal of Econometrics*, 142 (2), 698–714. [409]

McGuire, T. (2000), "Physician agency." In *Handbook of Health Economics* (A. Cuyler and J. Newhouse, eds.), 467–536, North-Holland. [398]

Milgrom, P. and C. Shannon (1994), "Monotone comparative statics." *Econometrica*, 62 (1), 157–180. [406]

Misra, S. and H. Nair (2011), "A structural model of sales-force compensation dynamics: Estimation and field implementation." *Quantitative Marketing and Economics*, 9 (3), 211–225. [400]

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Moffitt, R. (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. [400]

Mussa, M. and S. Rosen (1978), "Monopoly and product quality." *Journal of Economic Theory*, 18, 301–317. [405]

Nekipelov, D. (2010), "Empirical content of a continuous-time principal-agent model: The case of the retail apparel industry." Working Paper. [400]

Pudney, S. (1989), *Modelling Individual Choice: The Econometrics of Corners, Kinks, and Holes*. Blackwell Publishers, Cambridge, UK. [400]

Reiss, P. and M. White (2005), "Household electricity demand, revisited." *Review of Economic Studies*, 72, 853–883. [400]

Rosenthal, M. (2004), "Donut-hole economics." *Health Affairs*, 23 (6), 129–135. [403]

Saez, E. (2010), "Do taxpayers bunch at kink points?" *American Economic Journal: Economic Policy*, 2, 180–212. [400, 426]

Spence, M. (1973), "Job market signaling." *Quarterly Journal of Economics*, 87 (3), 355–374. [405]

Topkis, D. (1978), "Minimizing a submodular function on a lattice." *Operations Research*, 26, 305–321. [406]

Wilson, R. (1997), *Nonlinear Pricing*. Oxford University Press, London. [426]

Yip, W. (1998), "Physician response to medicare fee reductions: Changes in the volume of coronary artery bypass graft (CABG) surgeries in the medicare and private sectors." *Journal of Health Economics*, 17 (6), 675–699. [398]

Co-editor Rosa L. Matzkin handled this manuscript.

Manuscript received 12 August, 2015; final version accepted 14 November, 2016; available online 30 January, 2017.

17597331, 2017, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.3982/QE602 by Test, Wiley Online Library on [13/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License