Health Insurer Gatekeeping

Natalia Serna*

October 25, 2024

Abstract

In managed care health systems, insurers can impose barriers in access to healthcare services by engaging in gatekeeping practices. Using the discontinuity in cost-sharing introduced by the out-of-pocket maximum, I show that there is a substantial price elasticity when prices are zero for consumers but positive for insurers, suggestive of insurer gatekeeping. I find that gatekeeping significantly reduces healthcare utilization and spending, even in services that patients with chronic diseases need for treatment. Estimates from a model of hospital demand indicate that gatekeeping induces patients to choose less preferred hospitals.

Keywords: Health insurance; Gatekeeping; Cost-sharing; Information Frictions.

JEL codes: I10, I11, I13, I18.

*e-mail: nserna@stanford.edu. I am deeply grateful to the Colombian Ministry of Health for providing the data for this research and to Lin-Tung Tsai for excellent research assistance. I thank Marguerite Burns, Ken Hendricks, Riley League, Grant Miller, Corina Mommaerts, Maria Polyakova, Parker Rogers, and Alan Sorensen for their comments and advice. I also thank participants to the 2024 International Industrial Organization Conference. The findings of this paper do not represent the views of any institution involved. All errors are my own.

1 Introduction

Barriers in access to healthcare services in the form of claim denials (League, 2023), prior authorization requirements (Brot-Goldberg et al., 2023), or narrow provider networks (Ho and Lee, 2017; Buitrago et al., 2024) are widely used across health systems with managed care competition. In the U.S., recent news articles highlight extreme cases of patients who have died perhaps as a result of insurers' use of mechanisms to gatekeep their enrollees from the care they need.[1] While some view these gatekeeping practices as hurting patient health and imposing administrative burdens on doctors and patients,[2] gatekeeping may also be a powerful cost-minimizing strategy, suggesting its impact on social welfare can be ambiguous.

Some literature to date has analyzed the trade-off associated with gatekeeping practices, but only in the context of prescription drugs (e.g., Brot-Goldberg et al., 2017) or with a single payer (e.g., League, 2023). In this paper, I document novel evidence of insurer gatekeeping along multiple types of healthcare and in a setting with managed care competition. Specifically, I quantify the impacts of insurer gatekeeping on healthcare utilization and spending, health outcomes, and where consumers go to seek care. I show that insurer gatekeeping is a much more effective cost-containment mechanism than patient cost-sharing, but that it may undermine health outcomes by reducing the utilization of services that patients with chronic diseases need for treatment.

I study the effects of gatekeeping in the context of Colombia's contributory

[1] See https://www.nytimes.com/2024/03/14/opinion/health-insurance-prior-authorization.html?smid=nytcore-ios-share&referringSource=articleShare

[2] See https://www.nytimes.com/2024/05/25/science/medicare-seniors-authorization.html?smid=nytcore-ios-share&referringSource=articleShare&sgrp=c-cb and https://www.nytimes.com/2014/08/04/opinion/adventures-in-prior-authorization.html?smid=nytcore-ios-share&referringSource=articleShare&sgrp=c-cb

healthcare system. This system covers individuals who pay payroll taxes and provides access to the national health insurance plan. Several aspects of this plan are strictly regulated such as premiums, cost-sharing, and health service coverage. Cost-sharing rules (copays, coinsurance rates, and maximum out-of-pocket (OOP) amounts in the year) are indexed to the enrollee's monthly income level but are standardized across insurers and providers.

To identify gatekeeping I leverage the discontinuity in coinsurance rates introduced by the OOP maximum. Coinsurance rates drop to zero after patients reach this maximum and the insurer has to cover the full cost of care. Gatekeeping incentives are therefore more salient after patients reach the maximum, but are present along the entire distribution of healthcare expenditures. I show that reaching the OOP maximum in my setting is a random and sudden event, typically a hospitalization, and thus individuals cannot preempt it. This provides a unique setting to estimate the causal effect of insurer gatekeeping on different outcomes. To do so, I use enrollment and health claims data from a random panel of 8 million enrollees from 2009 to 2011, who did not switch their insurer over the sample period, who were diagnosed with a chronic disease, and who were at the bottom of the income distribution, making less than 2 times the monthly minimum wage ($270). These sample restrictions help to control for the evolution of health status, a known source of confounding bias.

I compare average claim prices and number of claims between individuals who reach their OOP maximum and those who do not reach it, in a dynamic difference-in-differences design. Focusing on the sub-sample of individuals within a bandwidth of 90 thousand pesos ($49) around the OOP maximum, I find that treated individuals consume significantly cheaper services and make substantially fewer claims than controls after reaching the OOP maximum. These findings are at odds with behavioral assumptions

about consumers when they face zero prices and they are also inconsistent with individuals' health status worsening due to sudden health shocks such as hospitalizations. Instead, results are in line with insurers gatekeeping claims and with information frictions that make consumers unaware of zero prices.

To separate the effect of information frictions from gatekeeping, I estimate my event study specification separately for cohorts of patients that reach their OOP maximum in different months of the year. If information frictions disappear over time, then cohorts who reach their OOP maximum early in the year should consume more expensive services the closer they are to the end of the year before cost-sharing resets. My findings show no evidence that negative treatment effects vanish for any cohort, suggesting that information frictions are not the main driver of reductions in spending.

Event study results conform to the idea that insurer gatekeeping is an important source of responsiveness to prices. Even if consumers can perfectly assess their zero out-of-pocket prices after reaching the OOP maximum, demand slopes down because of insurer gatekeeping. This evidence extends prior findings in Prager (2020) who showed that consumers engage in price-shopping across hospitals when prices are positive and there are no information frictions.

I further show that insurers are less likely to gatekeep claims made in an inpatient setting, but are more likely to gatekeep discretionary care made in an outpatient setting, such as imaging and laboratory tests. I also show that the utilization of services associated with appropriate disease management also falls after reaching the OOP maximum, suggesting potentially negative impacts on patient health.

To provide more direct evidence of the gatekeeping mechanism, in the last part of the paper I turn to examining the types of providers that consumers choose to receive care before and after reaching the OOP maximum. While

insurers in Colombia can deny claims or require prior authorization, I do not observe these practices in the data. I hypothesize that the main gatekeeping mechanism is to steer patients towards cheaper providers given that insurers compete only on their provider networks.[3] This mechanism can explain why patients consume cheaper services after reaching the OOP maximum conditional on making claims, something that claims denials alone cannot explain for example.

I develop and estimate a structural model of hospital demand that incorporates consumers' and insurers' responsiveness to prices in two states of the world: before and after patients reach their OOP maximum. My model estimates show significant responsiveness to prices in the two states, in line with the reduced-form evidence. Estimates also show that consumers dislike commuting to visit healthcare providers. In a partial equilibrium exercise where I prohibit insurer gatekeeping, I find that patients on average would choose hospitals that are twice as expensive and have 38% higher quality relative to the observed scenario. Unlike gatekeeping, information frictions have negligible effects on consumers' choices.

1.1 Related literature

Non-price mechanisms to contain healthcare costs and unnecessary spending have become increasingly popular since the advent of managed care (Glied, 2000; Glazer and McGuire, 2000). Although traditionally physicians served the role of gatekeepers in different health systems (Godager et al., 2015; Dumontet et al., 2017; Brekke et al., 2007), managed care companies now influence how

[3]Other elements of the public health insurance contract such as premiums and cost-sharing rules (coinsurance rates, copayments, and out-of-pocket maxima) are regulated and standardized across insurers.

consumers access healthcare (Baker and Corts, 1996).

This paper contributes to a growing literature on insurers' use of gatekeeping mechanisms to reduce healthcare spending, such as prior authorization requirements (Roberts et al., 2021; Brot-Goldberg et al., 2023) and claim denials (Gottlieb et al., 2018; League, 2023; Dunn et al., 2024). Complementing these papers, I show evidence of an alternative gatekeeping mechanism, namely, whether insurers steer patients towards cheaper providers when they have to cover the full cost of healthcare. I provide causal estimates of the impacts of gatekeeping on healthcare utilization and spending on several types of services.

I show that insurer gatekeeping is much more effective than patient cost-sharing at containing spending. In doing so, I build on prior empirical work that finds mixed evidence of patient cost-sharing being effective on this front (Chandra et al., 2010; Shigeoka, 2014; Chandra et al., 2014; Baicker et al., 2015; Brot-Goldberg et al., 2017; Chandra et al., 2021; Buitrago et al., 2021). Specifically, the finding that demand for healthcare always responds to the shadow price of care even when out-of-pocket prices are zero are in contrast to Brot-Goldberg et al. (2017) and Lin and Sacks (2019) who find no evidence of responsiveness to prices after patients hit their deductible.

Finally, my paper also relates to the literature that quantifies changes in demand when healthcare prices are zero (Chandra et al., 2010; Dague, 2014; Drake et al., 2023; Iizuka and Shigeoka, 2022) and that quantifies the relative costs of information frictions in healthcare (Handel and Kolstad, 2015; Handel et al., 2019; Brown, 2019).

The remainder of this paper is structured as follows: section 2 describes the empirical setting, section 3 describes my data, section 4 presents the empirical analysis to identify gatekeeping, section 5 presents the structural model of hos-

pital demand, section 6 presents results from the partial equilibrium analyses, and section 7 concludes.

2 Cost-Sharing in Colombia

The Colombian healthcare system was established in 1993 and is divided into a contributory and a subsidized regime. The first covers individuals who are employed or self-employed and can pay their taxes. The second covers individuals who are poor enough to qualify and is fully funded by the government through tax revenue. In both regimes, enrollees have access to the national health insurance plan, which is provided by private insurers.

The government regulates several aspects of the national plan: insurance premiums are zero in both regimes, individuals in the contributory system have to pay a fraction of their healthcare expenditures through cost-sharing, and healthcare is free in the subsidized system (apart from small copays). Insurers have no discretion on how to design these elements of the insurance plan, but they can decide on their network of covered providers.

TABLE 1: Cost-Sharing Rules in the Contributory Health Care System

Income level	Copay	Coinsurance rate Per claim	OOP Maximum Per year
Low: Income$<$ 2 MMW	1,900	11.5%	57.5%
Middle: Income$\in [2,5]$ MMW	7,600	17.3%	230%
High: Income$>$ 5 MMW	20,100	23.0%	460%

Note: Table shows the copays, coinsurance rates, and OOP maximum per income level that apply to individuals enrolled in Colombia's contributory health care system. The monthly minimum wage (MMW) in 2009 equals 535,600 COP or roughly $270. The coinsurance rates are percentages of claim prices, whereas the OOP maximum is a percentage of the MMW.

Cost-sharing rules in the contributory system are a function of the enrollee's monthly income level but are standardized across insurers and hospitals. These

rules involve a three-tier system of copayments, coinsurance rates, and maximum out-of-pocket (OOP) amounts in the year as seen in Table 1. Individuals are assigned specific cost-sharing rules depending on whether they make less than 2, between 2 and 5, or more than 5 times the monthly minimum wage (MMW), which equaled roughly $270 in 2011. For example, for individuals who make less than 2 times the MMW, the copay equals 1,900 pesos (nearly $1), the coinsurance rate is 11.5% of the price per health claim, and the OOP maximum is 57.5% of the MMW, which resets every year. Enrollees pay copays every time they go to a primary care doctor or a specialist and they pay coinsurance rates for every health service that they claim.[4] After individuals reach their OOP maximum in the year, copays and coinsurance rates drop to zero and the insurer covers the full cost of healthcare. These cost-sharing rules have not changed since the establishment of the healthcare system.

Previous studies have analyzed the impacts of coinsurance rate and copay discontinuities in the Colombian healthcare system on utilization, spending, and health outcomes (Serna, 2021; Buitrago et al., 2021). These studies have leveraged comparisons across consumers within an income tier using a regression discontinuity (RD) design. In this paper, I analyze insurer gatekeeping comparing within-patient changes in outcomes before and after they reach their OOP maximum, using a dynamic difference-in-differences (*did*) design.[5]

Insurer gatekeeping refers to mechanisms by which the insurer deters, denies, or steers health claims made by its enrollees in order to control costs or risk select. These mechanisms may include requiring prior authorization, denying claims, steering patients towards cheaper providers, or having long

[4] Although coinsurance rates apply only to family members who are enrolled through the individual who is a tax contributor, these coinsurance rates are usually paid for by the contributor.

[5] I also provide RD estimates following the prior literature as a robustness check.

lines to file claims or receive medications. This definition differs from settings where primary care providers or general practitioners serve as gatekeepers by deciding whether to refer a patient for more specialized care as in the U.K. (Forrest, 2003), France (Dumontet et al., 2017), or Norway (Godager et al., 2015).

3 Data and Descriptives

My raw data consist of all the health claims from a random sample of nearly 8 million enrollees in Colombia's contributory system from 2009 to 2011, who made at least one claim and who did not switch their insurer during the sample period. For every individual, I observe socio-demographic characteristics such as sex, age, income, and municipality of residence (similar to a county in the U.S.). For every claim the data report insurer, provider, service, International Classification of Diseases Code 10 (ICD-10), and negotiated price. I do not observe claim denials. Using the enrollee's income, I recover their level of cost-sharing and the OOP maximum that applies to each of them. With the health claims data, I construct different measures of monthly utilization and spending and determine whether and when they reach their OOP maximum.

I consider observations from one individual in different years as different individuals because cost-sharing resets at the beginning of each calendar year.[6] To control for the confounding bias arising from changes in health status, I perform my main analysis in the sample of enrollees who were diagnosed with a chronic disease at any moment during the sample period and who were at the bottom of the income distribution, making less than 2 times the MMW. I also

[6]This assumption implies that I consider my data as repeated cross-sections, and therefore I will exploit the variation within years.

exclude individuals who reach their OOP maximum during the first quarter of every year because I do not observe healthcare utilization prior to reaching the OOP maximum. For tractability, I choose a random sample of 200,000 individuals in each year (600,000 in total).

Table 2 presents some summary statistics of my sample. An observation is a person-month. Column (1) shows descriptives for the full sample, column (2) for the analysis sample focusing on low income individuals diagnosed with any chronic condition, column (3) for the analysis sample that reaches their OOP maximum, and column (4) for the analysis sample that does not reach their OOP maximum. In my analysis sample, 5% of individuals reach their OOP maximum. These individuals are on average older than those who do not reach it. 76% of those in column (3) have a cardiovascular disease (such as hypertension), while 67% of those in column (4) have this diagnosis. Consumers who reach their OOP maximum have higher healthcare utilization than their counterparts, a difference that is driven only by the health claim made during the month in which they reach the OOP maximum. For example, the average claim price is over 5 times higher and the likelihood of being hospitalized in a month is 9 percentage points (p.p) higher for those in column (3) relative to column (4).

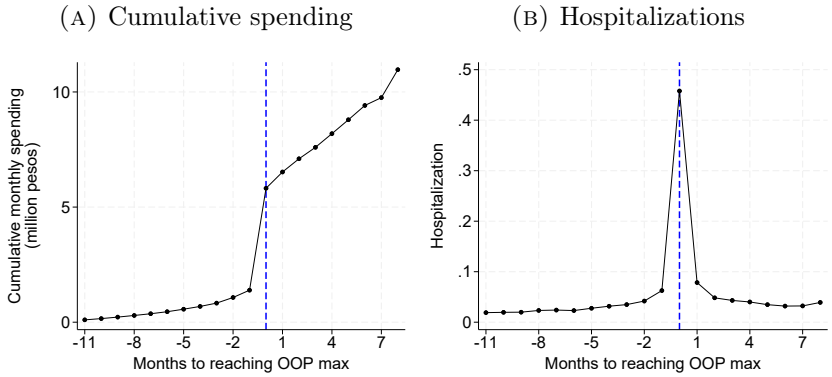
Reaching the OOP maximum in my setting is a sudden event. Panel A of Figure 1 shows that cumulative monthly spending increases smoothly until the month before reaching the OOP maximum and has a sharp discontinuity when this maximum is reached. This sudden event is typically a hospitalization as seen in Panel B. A little under 50% of individuals who reach the OOP maximum have a hospitalization and the other half either claim an expensive imaging service or an expensive visit to the specialist as seen in Appendix Figure 1.

TABLE 2: Summary Statistics

	Full sample	Analysis sample	Above OOP max	Below OOP max
	(1)	(2)	(3)	(4)
Socio-demographic				
Male	0.42 (0.49)	0.35 (0.48)	0.42 (0.49)	0.34 (0.47)
Age	50.0 (17.5)	55.4 (17.1)	60.7 (17.0)	54.9 (17.0)
Low income	0.77 (0.42)	1	1	1
Medium/High income	0.23 (0.42)	—	—	—
Health				
Any chronic disease	0.64 (0.48)	1	1	1
Cancer	0.22 (0.41)	0.35 (0.48)	0.37 (0.48)	0.35 (0.48)
Cardiovascular	0.43 (0.50)	0.68 (0.47)	0.76 (0.43)	0.67 (0.47)
Pulmonary	0.07 (0.25)	0.11 (0.31)	0.23 (0.42)	0.10 (0.30)
Renal	0.05 (0.22)	0.08 (0.27)	0.16 (0.37)	0.07 (0.26)
Health care use				
Hospitalization	0.01 (0.12)	0.02 (0.13)	0.10 (0.31)	0.01 (0.10)
Relative OOP spending	-0.50 (0.59)	-0.23 (0.24)	0.10 (0.77)	-0.26 (0.05)
Average claim price	0.04 (0.26)	0.04 (0.27)	0.17 (0.93)	0.03 (0.06)
Total claims	3.73 (5.97)	4.08 (6.68)	8.48 (18.7)	3.71 (4.07)
Inpatient spending	0.04 (0.72)	0.05 (0.84)	0.49 (2.95)	0.01 (0.10)
Outpatient spending	0.07 (0.28)	0.08 (0.33)	0.32 (1.06)	0.06 (0.11)
Imaging spending	0.01 (0.09)	0.02 (0.11)	0.06 (0.36)	0.01 (0.05)
Laboratory spending	0.01 (0.05)	0.01 (0.06)	0.04 (0.18)	0.01 (0.03)
Any chemo/Radio therapy	0.002 (0.04)	0.002 (0.05)	0.02 (0.15)	0.001 (0.03)
Any A1C blood test	0.01 (0.08)	0.01 (0.10)	0.02 (0.13)	0.01 (0.10)
Any AMI	0.003 (0.06)	0.01 (0.07)	0.02 (0.16)	0.004 (0.06)
Individuals	600,000	234,824	12,073	222,751
Individuals × Months	2,615,407	1,314,209	103,766	1,210,443

Note: Table presents mean and standard deviation in parenthesis of consumer characteristics in the full sample in column (1), in the analysis sample that conditions on low income and individuals with any chronic condition in column (2), in the analysis sub-sample that reach their OOP maximum in column (3), and in the analysis sub-sample that do not reach their OOP maximum in column (4). An observation is an individual-month. The full sample consists of a random sample of 200 thousand individuals per year from 2009 to 2011. Price and spending variables are measured in millions of pesos. The average exchange rate in 2011 was 1,848 COP/USD.

FIGURE 1: Cumulative Spending and Hospitalizations by Month



Note: Figure shows average cumulative monthly spending in Panel A and average number of hospitalizations in Panel B by month relative to when the individual reaches the OOP maximum. Figure uses the full sample of 600,000 individuals.

4 Identifying Gatekeeping

To identify insurer gatekeeping, I compare people who reach the OOP maximum (treatment group) and those who do not reach it (control group), within a bandwidth of spending around their OOP maximum. The rationale for this approach is that gatekeeping incentives should be stronger among the group of patients who have their insurer completely cover the cost of care, compared to those who pay a fraction of their healthcare cost through cost-sharing. My empirical strategy is a dynamic *did* design (following Colonnelli et al., 2020).

The regression specification is as follows:

$$y_{it} = \sum_{\substack{k=-11 \\ k\neq -1}}^{8} \beta_k \mathbf{1}\{t - t^* = k\} \times \text{Treated}_i + \theta S_{it} + \alpha_i + \gamma_t + \varepsilon_{it} \qquad (1)$$

where y_{it} is an outcome of individual i in month t, t^* is the calendar month in which the treated individual reaches the OOP maximum; S_{it} is consumer

i's relative OOP spending up to month t which, in the style of an RD design, imposes a linear trend on the treatment assignment variable; α_i is an individual fixed effect and γ_t is a calendar month fixed effect.[7] I focus on individuals that are within 90 thousand pesos ($43) around their OOP maximum and consider alternative bandwidths as robustness checks in the appendix.

The coefficients β_k measure the average treatment effect in month k relative to the month when individuals reach their OOP maximum. For those in the control group, I normalize $k = -1$. I use De Chaisemartin and d'Haultfoeuille (2020)'s estimator to deal with staggered treatment and possibly heterogeneous treatment effects. Standard errors are clustered at the individual level, which defines the level of treatment. Appendix 3 presents all the associated coefficients and standard errors.

4.1 Threats to Identification

The empirical strategy of using individuals who face zero prices to identify the magnitude of gatekeeping has several identification threats. The first is a selection bias problem: people who reach their OOP maximum may be unobservably sicker and less responsive to prices compared to those who do not reach it. This type of unobserved heterogeneity might lead a researcher to underestimate the effects of insurer gatekeeping. I address this concern by focusing on the sub-sample of enrollees who were diagnosed with a chronic

[7] Recent literature on *did* has devoted attention to the issues that arise when including time-variant and time-invariant covariates (e.g., Caetano et al., 2022). If cumulative OOP spending in my specification is affected by treatment or the true underlying relation between my outcome and this covariate is non-linear, then estimates for β_k will be biased. However, note first that cumulative OOP spending determines treatment, not the other way around; and second outcomes such as the average claim price contribute linearly to the cumulative OOP spending that is used to determine whether the patient reaches the OOP maximum. Colonnelli et al. (2020) use a similar specification for studying patronage in context of close electoral races in Brazil.

disease at any moment during the sample period and who have low incomes. I further address this concern by estimating treatment effects in a relatively small bandwidth of cumulative OOP spending around the OOP maximum. Within this sub-sample, I show that reaching the OOP maximum is a sudden event as seen in Figure 1, such that there are no reasons to believe that individuals can anticipate it.

The second is a confounding bias problem: changes in demand after reaching the OOP maximum may come from patients facing zero prices, facing information frictions, experiencing changes in health status, or experiencing reversions to the mean. These confounding factors might lead a researcher to overestimate the effects of gatekeeping. I will address these concerns in sections 4.3 and 4.4.

4.2 Main Results

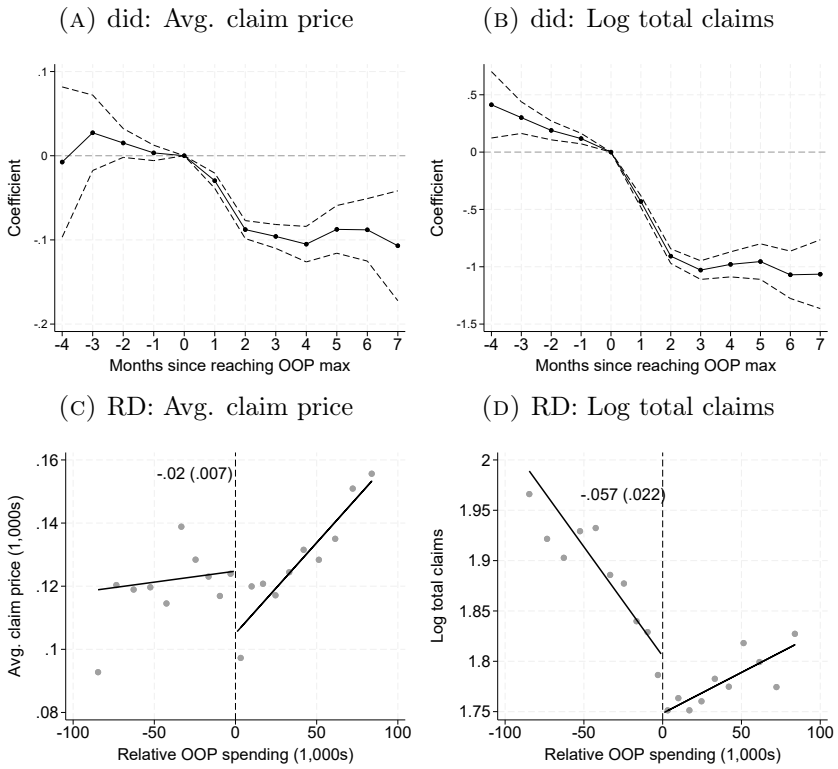
Figure 2 presents event study coefficients and 95% confidence intervals. The outcomes variables are the average claim price in Panel A and the log of total claims in Panel B. Panels C and D present robustness checks using an RD design conditional on treated individuals. Appendix 2 presents tests of the assumptions for a valid RD design.[8] I use a bandwidth of 90 thousand pesos ($49) around the cutoff of zero relative OOP spending (relative to the OOP maximum), which is the optimal bandwidth according Calonico et al. (2014)'s algorithm. Appendix Figure 4 presents robustness checks using different bandwidths.

I find that trend differences in the average claim price before the event

[8] Appendix Figure 2 shows that the distribution of the running variable (OOP spending relative to the OOP maximum) has an apparent spike after the cutoff of zero, but this is not a significant discontinuity based on the McCrary (2008) test. However, due to this potential limitation of the RD design, my preferred specification is the event study.

between treated and control groups are negligible, suggestive of parallel pre-trends. Although I cannot rule out that treated individuals are on different trends relative to controls for the log of total claims, the fact that trend differences reverse after reaching the OOP maximum is suggestive of substantial treatment effects.

FIGURE 2: Utilization and Average Prices after Reaching the OOP Maximum



Note: Panels A and B show coefficients and 95% confidence intervals of the event study specification following equation (1) for the average claim price and the log of total claims, respectively. Estimation uses De Chaisemartin and D'Haultfouille's estimator. Standard errors are clustered at the individual level. Panels C and D show a regression discontinuity plot for the average claim price and the log of total claims, respectively, conditional on treated individuals. Labels in these panels show the RD point estimate and its standard error in parenthesis. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum. All specifications control for relative OOP spending.

People who reach the OOP maximum claim services that are on average 85 thousand pesos ($46) cheaper than individuals in the control group (a 17%

decrease based on the RD estimate). Individuals who reach the OOP maximum not only consume relatively cheaper services after the event, but they make fewer claims overall as seen in Panel B. For both the average claim price and the log of total claims, reductions in the post-period are persistent.

TABLE 3: Claims-level Difference-in-Differences Regression for the Log of Claim Price

	Full sample		Analysis sample	
	(1)	(2)	(3)	(4)
Treated×Post	-0.197	-0.065	-0.191	-0.060
	(0.008)	(0.005)	(0.009)	(0.005)
Fixed effects				
Individual	✓	✓	✓	✓
Month	✓	✓	✓	✓
Service	No	✓	No	✓
Observations	8,596,692	8,596,058	4,637,268	4,636,546

Note: Table shows a regression of the log of claim price on the interaction between the treatment indicator (for individuals who reach the OOP maximum) and post-period indicator (for months after reaching the OOP maximum). Columns (1) anbd (2) use the full sample and columns (3) and (4) use the analysis sample that conditions on individuals with any chronic condition who have low incomes. Columns (1) and (3) include individual and month fixed effects. Columns (2) and (4) additionally include service fixed effects. An observation is a claim. Standard errors in parenthesis are clustered at the service level.

Table 3 finds confirming results of the negative treatment effect on the average claim price by exploiting the claims-level data. In this data I estimate a *did* regression of the log of claim prices on individual, month, and service fixed effects for the full sample and for the analysis sample. In the specifications that include service fixed effects in columns (2) and (4), findings indicate that individuals who reach the OOP maximum consume services that are between 6% and 7% cheaper than controls.

4.3 Zero-Price and Health Shock Effects

Despite the substantial reductions in average claim prices and utilization, estimates in Figure 2 confound the effect of two changes at play: the change in health status due to sudden hospitalizations ("health shock effect") and the change in OOP prices ("zero-price effect"). These effects are highly correlated in the sense that individuals face sudden hospitalizations and zero prices at the same time.

Disentangling the zero-price effect from the health shock effect is important because gatekeeping incentives may differ across people who face sudden hospitalizations versus those who face zero prices without having a hospitalization. For instance, we might expect an insurer to be more willing to send patients to expensive providers when it has to cover the full cost of care if the patient is relatively sick compared to when the patient is relatively healthy. Hence, health shocks and zero prices may have opposite impacts on utilization and spending.

To decompose the zero-price and the health-shock effects, let $t = 0$ denote the period before reaching the OOP maximum and $t = 1$ the period after reaching the maximum. Let $H(i,t)$ be an indicator for individual i having a hospitalization in period t, $D(i,t)$ be the price of healthcare that individual i faces in period t, and $Y(i,t)$ be the average claim price of individual i in period t. The *did* specification in equation (1) identifies the average treatment effect on the treated (ATT) generally as

$$\begin{aligned}\beta = &(E[Y(i,1)|D(i,1)=0] - E[Y(i,1)|D(i,1)>0]) \\ &- (E[Y(i,0)|D(i,1)=0] - E[Y(i,0)|D(i,1)>0])\end{aligned}$$

Underlying this treatment effect is the effect of sudden changes in health status driven primarily by hospitalizations. Both treated and control units may face hospitalizations in the pre- or post-periods. This implies that each element of the previous equation is a weighted average across hospitalization status. Using this fact, we can rewrite the ATT as:

$$
\begin{aligned}
\beta = & \Big(E[Y(i,1)|D(i,1)=0,H(i,1)=0] - E[Y(i,1)|D(i,1)>0,H(i,1)=0]\Big) \\
& \underbrace{-\Big(E[Y(i,0)|D(i,1)=0,H(i,1)=0] - E[Y(i,0)|D(i,1)>0,H(i,1)=0]\Big)}_{\text{Zero-price effect}} \\
& + x\Big(E[Y(i,1)|D(i,1)=0,H(i,1)=1] - E[Y(i,1)|D(i,1)=0,H(i,1)=0]\Big) \\
& - x\underbrace{\Big(E[Y(i,0)|D(i,1)=0,H(i,1)=1] - E[Y(i,0)|D(i,1)=0,H(i,1)=0]\Big)}_{\text{Treated health shock effect}} \\
& - \Big(y\Big(E[Y(i,1)|D(i,1)>0,H(i,1)=1] - E[Y(i,1)|D(i,1)>0,H(i,1)=0]\Big) \\
& - y\underbrace{\Big(E[Y(i,0)|D(i,1)>0,H(i,1)=1] - E[Y(i,0)|D(i,1)>0,H(i,1)=0]\Big)\Big)}_{\text{Control health shock effect}}
\end{aligned}
$$

where $x = P(H(i,1)=1|D(i,1)=0)$ is the probability of having a hospitalization conditional on being above the OOP maximum and $y = P(H(i,1)=1|D(i,1)>0)$ is the probability of having a hospitalization conditional on being below the OOP maximum. The expression above shows that the ATT is the sum of the zero-price effect and the marginal health shock effect on the treated relative to controls.

Figure 3 presents results of this decomposition exercise using as outcome the average claim price. Panel A depicts event study coefficients for the zero-price effect, that is, conditional on people who never have a hospitalization in the study period. The regression specification in this case is the same as

equation (1). Panel B depicts event study coefficients for the treated health shock effect, that is, conditional on individuals who reach the OOP maximum. For this effect, time indicators are constructed relative to the month when the individual is last hospitalized, h^*. Formally, the regression equation for the treated health shock effect is:

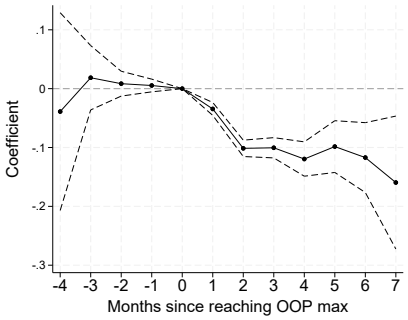
$$y_{it} = \sum_{\substack{k=-11 \\ k\neq -1}}^{8} \beta_k \mathbf{1}\{t - h^* = k\} \times \text{Hospitalized}_i + \theta S_{it} + \alpha_i + \gamma_t + \varepsilon_{it}$$

where, as before, I consider individuals within a bandwidth of 90 thousand pesos around their OOP maximum.

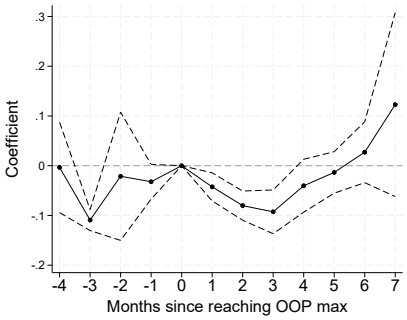
Findings show that the zero-price effect is significantly negative in the post-period. On average, individuals who never have a hospitalization make claims that are 104 thousand pesos ($56) cheaper than controls after reaching the OOP maximum (a 39% decrease based on the RD estimate in Panel C). In contrast, the treated health shock effect is positive by the end of the sample period, although coefficients are noisy. The estimates of the health shock effect go in line with the intuition that individuals with poor health tend to be less price sensitive. But, the negative treatment effects associated with facing zero prices are irreconcilable with healthcare demand being perfectly inelastic after consumers reach their OOP maximum. This suggests that factors other than consumers' OOP prices may explain why healthcare demand responds to cost-sharing.

FIGURE 3: Effect Decomposition

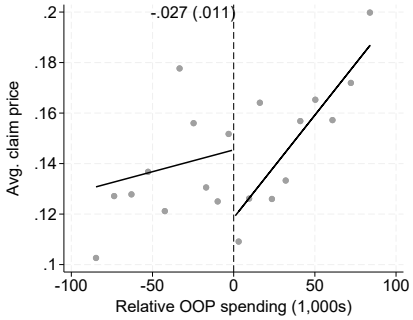
(A) did: Zero-price Effect



(B) did: Health Shock Effect



(C) RD: Zero-price Effect



Note: Figure presents coefficients and 95% confidence intervals of the event study specifications for the zero-price effect in Panel A and for the health shock effect in Panel B. In both specifications the outcome variable is the average claim price. The zero-price effect uses the sub-sample of individuals who are never hospitalized during the sample period. The health shock effect uses the sample of treated individuals comparing those who are hospitalized are those who are not, before and after the hospitalization. Specifications use De Chaisemartin and D'Haultfouille's estimator. Time indicators are constructed relative to reaching the OOP maximum for the zero-price effect and relative to the month when the individual is hospitalized for the health shock effect. Both specifications control for relative OOP spending. Standard errors are clustered at the individual level. Panel C presents the zero-price effect using a regression discontinuity design conditional on treated individuals who never have a hospitalization during the study period. The label in this panel shows the RD point estimate and its standard error in parenthesis. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum.

4.4 Information Frictions and Mean Reversion

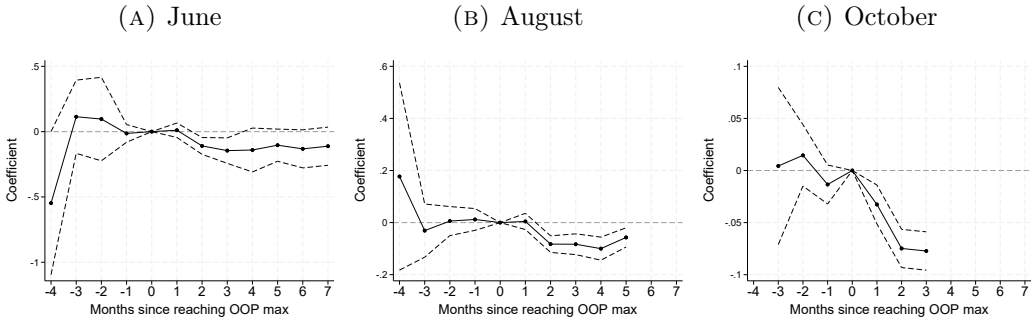
After reaching the OOP maximum consumers face zero prices because their coinsurance rate drops to zero. If consumers were forward-looking and did not face information frictions (such that they know exactly what their OOP prices

are at every point in time), they would either consume more healthcare services or more expensive services after the event than before the event. Healthcare spending can also increase over time after reaching the OOP maximum if consumers foresee that prices will be non-zero at the start of the next calendar year when cost-sharing resets. Even if consumers dislike going to the doctor, we would at least expect to see no changes in healthcare utilization in the post-period.

When compared to individuals who do not reach the OOP maximum, the zero-price effect in Panel A of Figure 3 is at odds with these behavioral assumptions about consumers when they face zero prices. The negative treatment effects associated with zero prices are consistent with insurers steering patients towards cheaper care, experiencing reversions to the mean, or facing information frictions after they reach their OOP maximum. I move now to exploring whether information frictions and mean reversion play a role in my findings.

The month in which a patient reaches their OOP maximum is unusual because it represents a discontinuity in utilization and spending as seen in Figure 1. After this month, it is reasonable to think that patient outcomes would reverse to their pre-period averages if there are no meaningful changes in health status. Moreover, if information frictions are meaningful and consumers are unaware about having reached their OOP maximum, they might behave as if they face non-zero prices after the event. If information frictions disappear over time, then we should see consumers either making more expensive claims or claiming more services the further they are from having reached their OOP maximum and the closer they are to the end of the calendar year before cost-sharing resets. We would expect to see a similar pattern in utilization and spending if individuals were experiencing reversions to the mean.

FIGURE 4: Average Prices by Cohort



Note: Figure presents coefficients and 95% confidence intervals of the event study specifications following equation (1) using as outcome the average claim price. Specifications use De Chaisemartin and D'Haultfouille's estimator on the sub-sample of treated individuals who reach their OOP maximum in June in Panel A, August in Panel B, and October in Panel C. Standard errors are clustered at the individual level. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum. All specifications control for relative OOP spending.

To check whether information frictions or mean reversion can explain the spending patterns in Panel A of Figure 3, I estimate separate event study specifications conditional on treated individuals who reach the OOP maximum in different months of the year. Findings in Figure 4 show that information frictions and mean reversion are not the main sources of reductions in spending in this setting. Individuals who reach their OOP maximum in June, August, or October see reductions in average claim prices every month after the event. In fact, treatment effects are greater in magnitude for cohorts that are treated earlier in the year. While these significant effects across cohorts do not rule out the presence of information frictions, negative estimates even for consumers who are treated in June suggest a role for insurer gatekeeping in explaining my results.

Overall, my findings of substantial responses of healthcare demand to the shadow price of care are in contrast to Brot-Goldberg et al. (2017), who show no evidence of consumers responding to prices after they hit their deductible.

Figures 2 and 4 in this paper indicate that demand responds significantly to prices even when these prices are zero for consumers because of insurer gatekeeping. Thus, failure to account for coinsurance rate discontinuities might lead a researcher to underestimate the price elasticity of demand for healthcare.

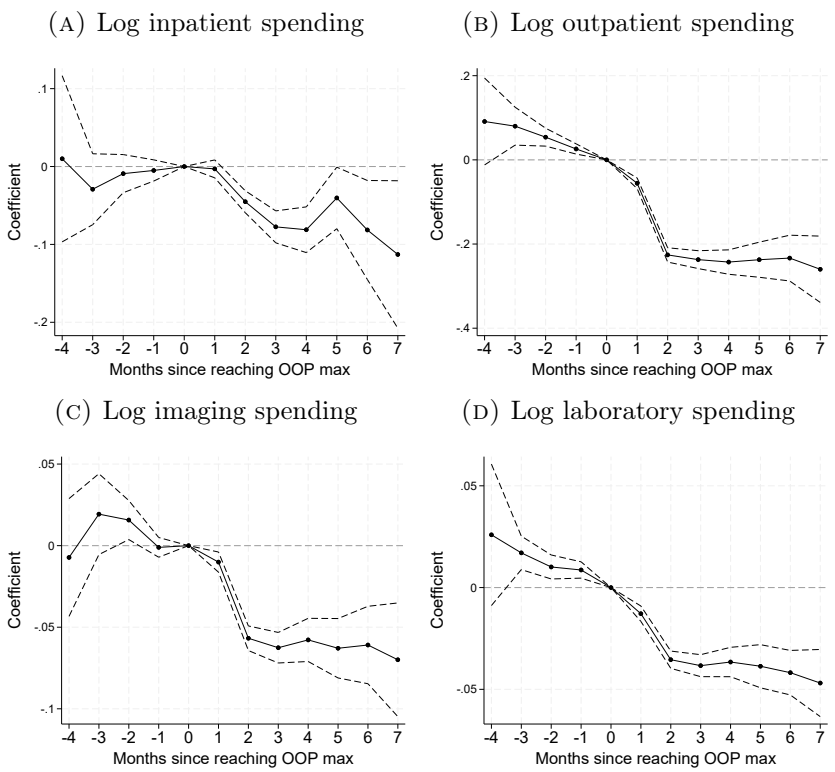
4.5 Impact of Gatekeeping on Specific Health Services

In this section I investigate the types of services that insurers are more likely to gatekeep. This analysis will provide evidence on gatekeeping incentives as well as on potential impacts on patient health outcomes. I estimate my event study specification on healthcare utilization and spending among subsets of health services. Figure 5 presents results of my event study specification using as outcomes the log of spending in inpatient, outpatient, imaging, and laboratory services.

The figure shows that gatekeeping efforts are present across all types of care but are smaller in magnitude for acute or necessary care. Reductions in spending on inpatient services are smaller than for outpatient services, imaging, and laboratory. For example, Panel A shows that treated individuals see a reduction in spending on inpatient services equal to 7% relative to controls 3 months after reaching the OOP maximum. Instead, the reduction in spending on outpatient services equals 20% in that same month. Panels C and D also show substantial reductions in spending on relatively discretionary services such as imaging and laboratory.

Figure 6 explores changes in the utilization of services that patients with certain chronic diseases need for adequate disease management. These services include chemotherapy or radiotherapy for patients with Cancer and A1C blood tests for people with diabetes in Panels A and B, respectively. Panel C uses

FIGURE 5: Spending by Type of Service

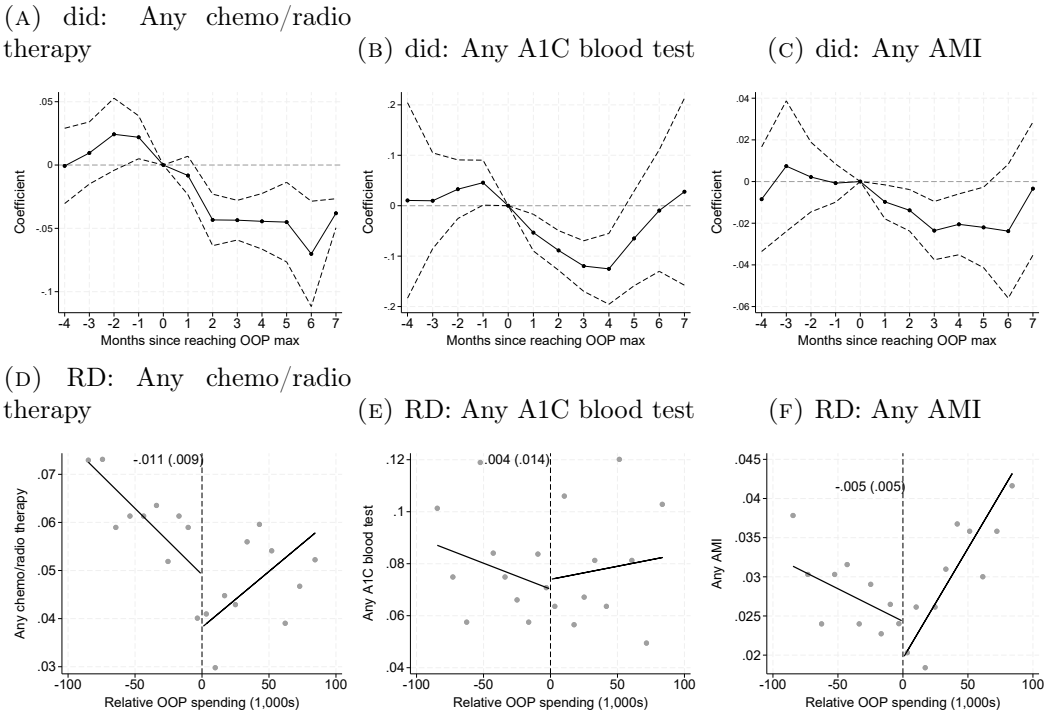


Note: Panels A to D show coefficients and 95% confidence intervals of the event study specification following equation (1) for the log of inpatient, outpatient, imaging, and laboratory spending, respectively. Estimation uses De Chaisemartin and D'Haultfouille's estimator. Standard errors are clustered at the individual level. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum. All specifications control for relative OOP spending.

as outcome an indicator for having an Acute Myocardial Infarction (AMI) conditional on patients with cardiovascular disease. Panels D to F report robustness checks using an RD design.

In Panels A and B I estimate significant reductions in the use of chemo/radio therapy and A1C blood tests, which are persistent over time. These reductions suggest worsened disease management and likely worse health outcomes for individuals with cancer and diabetes. In Panel C I also estimate a lower likelihood of AMIs among patients with cardiovascular disease. This finding

FIGURE 6: Utilization of Disease Management Services



Note: Panels A to C show coefficients and 95% confidence intervals of the event study specification following equation (1) using as outcomes an indicator for making chemotherapy or radiotherapy claims conditional on patients with Cancer, an indicator for having A1C blood tests conditional on patients with diabetes, and an indicator for having an Acute Myocardial Infarction (AMI) conditional on patients with cardiovascular disease. Estimation uses De Chaisemartin and D'Haultfouille's estimator. Standard errors are clustered at the individual level. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum. All specifications control for relative OOP spending.

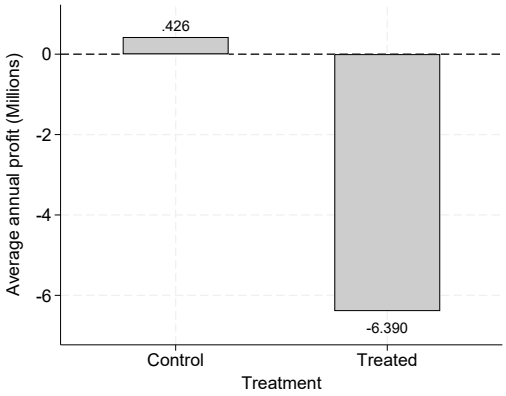
does not necessarily mean that health outcomes improve for these patients after reaching the OOP maximum, since the reduction in AMIs can also be due to patients who did not make it to the hospital after experiencing a stroke.

4.6 Gatekeeping Incentives

Why would a seemingly small change in insurer costs generate a large change in consumers' healthcare spending after reaching the OOP maximum? Note that

for a low-income consumer, insurer costs increase 11.5 percentage points after patients reach their OOP maximum because the insurer goes from covering 88.5% to 100% of the enrollee's healthcare costs (see Table 1). It is unlikely that gatekeeping incentives change discontinuously on this (intensive) cost-sharing margin, but they may change discontinuously based on the consumer's risk type. For example, if an individual reaches the OOP maximum, they are at risk of being very expensive to the insurer. Insurers may respond to this change in risk (extensive margin) by restricting the number and the type of services that their enrollees claim.

FIGURE 7: Average Insurer Profit per Enrollee



Note: Figure presents the average annual profit per enrollee conditional on individuals who do not reach their OOP maximum in the year (control) and conditional on those who reach it (treated). The profit per enrollee is the risk-adjusted transfer from the government minus total health care cost net of the patient's coinsurance payments. Averages use the full sample of individuals across all income groups. In the full sample, 15,393 individuals reach their OOP maximum and 584,607 do not.

Figure 7 provides suggestive evidence of these changes in consumer risk. The figure presents the average annual insurer profit per enrollee separately for individuals who stay below their OOP maximum (control) and for those who reach their OOP maximum (treated) in the year. The profit per enrollee is calculated as the annual risk-adjusted transfer from the government minus annual health care spending net of the patient's coinsurance payments. The

figure shows that individuals who stay below their OOP maximum are profitable, representing an average profit of 426 thousand pesos ($231). Instead, those who reach their OOP maximum and have the insurer completely cover the cost of care are substantially unprofitable, representing losses of 6.4 million pesos ($3,463). In the full sample, total losses incurred in treated individuals account for 21% of total profits accrued from controls.

5 Gatekeeping Mechanism

The analysis so far suggests that insurer gatekeeping is a likely source of price sensitivity, but it does not speak to the mechanisms that insurers use to gatekeep their enrollees. While insurers in Colombia can engage in claim denials and prior authorization practices, I do not observe these practices in the data. Yet, another mechanism by which insurers can gatekeep their patients is by directing them towards cheaper providers or giving them access to a narrow network of providers. Patient steering is the most salient gatekeeping mechanism in Colombia, because insurers compete mainly on their network of covered providers. Offering a narrow network and directing patients to cheaper providers would explain why after reaching the OOP maximum we see patients claiming cheaper services despite their worsened health and conditional on making claims, something that claim denials alone would not be able to explain.

The challenge with the exercise of showing that insurers steer patients at risk of reaching their OOP maximum to cheaper providers is separating consumer preferences from insurer gatekeeping. If we see the patient visit a cheaper hospital is it because they have an unobserved preference for this hospital? or is it because the insurer directed them to this hospital?

To separate the effect of insurer gatekeeping, I explicitly model patient preferences for hospitals before and after they reach their OOP maximum. Suppose consumer i is enrolled with insurer j. The consumer chooses a hospital h in the network of their insurer based on the indirect utility in two states of the world, before and after reaching the OOP maximum:

$$u_{ijh} = \begin{cases} (\alpha_i + \sigma_p \omega_i) r_i p_{jh} + \beta_i p_{jh} + \tau d_{ih} + \kappa x_{ih} + \xi_h + \varepsilon_{ijh} & \text{if } c_i + \nu_i \leq oop_i \\ \beta_i p_{jh} + \tau d_{ih} + \kappa x_{ih} + \xi_h + e_{ijh} & \text{o.w} \end{cases} \tag{2}$$

In this utility function, p_{jh} is the price that insurer j pays at hospital h for an admission, r_i is the coinsurance rate, d_{ih} is the distance from patient i to hospital h, and x_{ih} is an indicator for whether patient i had previously visited hospital h, which controls for the potential bias arising from provider inertia.[9] Price coefficients are given by $\alpha_i = x_i'\alpha$, $\beta_i = x_i'\beta$, where x_i is a vector of consumer demographics (dummies for sex and every age group) and diagnoses. I consider the following diagnoses: cancer, cardiovascular disease, diabetes, pulmonary disease, renal disease, high-cost disease (HIV-AIDS, autoimmune disorders, genetic anomalies, and transplants), and other diseases (arthritis, tuberculosis, epilepsy, asthma). Moreover, $\omega_i \sim N(0,1)$ captures unobserved heterogeneity in price sensitivity across consumers with dispersion parameter given by σ_p. Finally, ξ_h is a hospital fixed effect representing shared preferences for hospital h across consumers.

Consumer utility is a function of admission prices in both states of the world because insurer gatekeeping is present along the entire distribution of health care expenditures. However, its relative importance on the decision of

[9] For example, if a patient visited a relatively expensive hospital in the previous year and continues to visit this hospital, then the model would interpret this patient as having low price-sensitivity.

which hospital to visit is higher after patients reach their OOP maximum. A finding that both α and β are negative would imply that hospital choices are characterized by consumer and insurer sensitivity to prices.

States of the world differ on the source of responsiveness to prices. Before reaching the OOP maximum, the consumer and the insurer respond to prices up to the coinsurance rate. After reaching the OOP maximum, the insurer solely responds to prices since it covers the full cost of care. I specify the probability of staying below the OOP maximum as

$$\gamma_i = E[\mathbf{1}\{c_i + \nu_i \leq oop_i\}]$$

where, c_i is the consumer i's OOP spending up to but not including the hospital admission and oop_i is the OOP maximum.

Price sensitivity in both states of the world depends not only on gatekeeping but also on the magnitude of information frictions. These frictions may arise either because the consumer does not know the true shadow price of healthcare or because insurer j is uncertain about the patient's total OOP spending up to the hospital admission. I capture the impact of these information frictions on the probability of each state of world through $\nu_i \sim N(0, \sigma_\nu^2)$. This parameterization implies that

$$\gamma_i = \Phi\Big(\frac{oop_i - c_i}{\sigma_\nu}\Big)$$

I further assume that ν_i, ω_i, ε_{ijh}, and e_{ijh} are independent of each other, and that ε_{ijh} and e_{ijh} follow a type-I extreme value distribution.

Let H_j denote the set of hospitals in the network of insurer j. Given the

distribution of the preference shocks, the log-likelihood function is:

$$L = \sum_i \Big(\sum_{h \in H_j} y_{ijh} \log(P_{ijh}) + (1 - y_{ijh}) \log(1 - P_{ijh}) \Big) \tag{3}$$

where $P_{ijh} = \gamma_i P^1_{ijh} + (1 - \gamma_i) P^2_{ijh}$,

$$P^1_{ijh} = \int \frac{\exp(\delta^1_{ijh})}{\sum_{k \in H_j} \exp(\delta^1_{ijk})} d\phi(\omega), \qquad P^2_{ijh} = \frac{\exp(\delta^2_{ijh})}{\sum_{k \in H_j} \exp(\delta^2_{ijk})} \tag{4}$$

and

$$\delta^1_{ijk} = (\alpha_i + \sigma_p \omega_i) r_i p_{jh} + \beta_i p_{jh} + \tau d_{ih} + \kappa x_{ih} + \xi_h, \qquad \delta^2_{ijk} = \beta_i p_{jh} + \tau d_{ih} + \kappa x_{ih} + \xi_h$$

Identification. To separately identify the coefficients associated with admission prices in the two states of world, α_i and β_i, I use the discontinuity in coinsurance rates introduced by the OOP maximum. Price variation within hospital (across insurers) and coinsurance rate variation across patients are therefore needed to identify these coefficients. However, the price variation within hospital might be endogenous if consumers choose insurers that have negotiated low prices with their preferred hospitals or if there is some unobserved insurer quality that is correlated with prices. To deal with this potential price endogeneity, I use a Hausman-style instrument as follows.

When insurers and hospitals engage in bilateral price negotiations, they use as starting point the reference prices created by the government with a group of medical experts in 2005 (Ruiz et al., 2008).[10] There is a separate reference price for hospital admissions by basic, intermediate, and intensive care as well

[10] The reference prices were created to reimburse hospitals in the event of car accidents, natural disasters, and terrorist attacks (See Decree 2423 of 1996).

as by number of beds in the hospital room. To generate variation of reference prices across insurer-hospital pairs, I first calculate the average reference price for each pair conditional on admissions that happened during 2009, which are excluded from the analysis. Then, for insurer j and hospital h I calculate the average reference price across other hospitals $-h$ in the network of insurer j weighting by number of beds. I use the resulting average reference price and the number of beds as instruments for the negotiated price, which I implement using a control function approach (Petrin and Train, 2010). Appendix 5 describes the estimation details.

I impose that β_i is the same across the two states of the world because conditional on being below the OOP maximum, this coefficient is not identified separately from α_i when there is not enough variation in coinsurance rates. Imposing that β_i is the same across states therefore forces identification to come only from the state after reaching the OOP maximum. The unobserved preference heterogeneity parameter σ_p is identified from observationally identical patients that have not reached their OOP maximum but choose different hospitals and from variation in the choice set across consumers. Finally, the impact of information frictions captured by σ_ν is identified from comparing the choices made by patients who reach their OOP maximum and are observationally identical except for their OOP costs prior to the admission.

Data and sample restrictions. I estimate the hospital demand model using data from the full sample of enrollees described in column (1) of Table 2. I drop hospital admissions that happen during 2009 because lagged hospital choices (x_{ih}) will be missing for this year. A consumer's choice set is given by the hospitals that their insurer covers in their municipality of residence. I obtain this choice set from the claims data, considering a hospital as in-network for an insurer if it provides 10 or more admissions during the sample

period for that insurer.

Because I do not observe the patient's residence address but only their municipality of residence, I complement my enrollment and claims data with information on the distribution of population density by age across census blocks ("manzanas" for their Spanish name) within a municipality.[11] This information comes from the 2018 population census of Colombia. With this data I approximate distance to hospitals as: $d_{ih} \approx \sum_{l \in m} q_{\theta l} d_{lh}$, where $q_{\theta l}$ is the fraction of consumers type θ that live in census block l within municipality m and d_{lh} is the distance in kilometers from census block l's centroid to hospital h. Consumer types are defined by a combination of sex and ten-year age group.

I limit my analysis to the 13 largest municipalities in the country, for which census block-level information exists. These municipalities represent 75% of admissions. Appendix 4 describes the census data by reporting maps of the 4 largest municipalities in my sample with their census blocks and hospital geolocations.[12]

The admission prices reported in the claims data correspond to the negotiated prices between insurers and hospitals. However, pricing units may vary across insurer-hospital pairs in ways that make it difficult to predict prices for every hospital in a consumer's choice set. For example, some insurer-hospital pairs may negotiate an admission price that is specific to an age group and a category of length of stay. To overcome this challenge and express prices in the same unit across all insurer-hospital pairs, I follow Gowrisankaran et al. (2015). I regress the claims-level price on patient characteristics and hospital fixed effects separately for every insurer, and then average the predictions from these regressions to the level of an insurer-hospital pair. Appendix 4 describes

[11] Census blocks have an average area of 128 squared kilometers.

[12] I obtain hospital geolocations using Google's API.

this methodology in more detail.

Estimates. To estimate the hospital demand model, I use simulated maximum likelihood to approximate the integrals in equation (4). Results are presented in Table 4 and first-stage results of the regression of prices on the instruments are in Appendix Table 6. Since the instruments are a proxy for the hospitals' marginal cost, I find that there is a positive relation between the instruments and the negotiated prices. In the second stage, consistent with the reduced form evidence, I find that hospital demand responds to prices before and after consumers reach their OOP maximum. Before reaching this maximum, a 10,000 pesos increase in OOP prices (1/5 of the mean) reduces the probability of choosing a hospital by 18 percent. After reaching the maximum, a 10,000 pesos increase in admission prices (2% of the mean) reduces the choice probability by 25 percent. Because OOP prices are zero after patients reach their OOP maximum, price sensitivity in this state of the world can only be explained by insurer gatekeeping.

Interactions of prices with consumer demographics and diagnoses are in line with intuition and previous literature (e.g., Ho, 2006). Patients with chronic diseases are significantly less sensitive to OOP prices than patients without diagnoses. Gatekeeping incentives are stronger among older individuals who are potentially more expensive to the insurer. But, insurers are less likely to gatekeep admissions from individuals with chronic diseases compared to healthy individuals. Price sensitivity in both states of the world is substantially heterogeneous across consumers as seen by the estimate of σ_p. I find that patients dislike commuting to the hospitals in their choice set: if they have to travel one additional kilometer to visit a hospital, the probability of choosing this hospital decreases 10%. There is also evidence of information frictions given by the statistically significant estimate for σ_ν.

TABLE 4: Hospital Demand Estimates

		coef	se
OOP price		-17.83	(0.469)
Price		-25.01	(0.052)
Distance		-0.100	(0.002)
Lag visit		1.737	(0.001)
σ_p		10.40	(0.241)
σ_ν		0.072	(0.001)
Interactions			
OOP price	Male	0.733	(0.470)
	Age 20-30	(ref)	
	Age 31-40	10.15	(0.313)
	Age 41-50	0.744	(0.329)
	Age 51-60	3.039	(0.235)
	Age 61-70	3.115	(0.381)
	Age 71 or older	0.695	(0.328)
	Sick	5.834	(0.109)
Price	Male	0.532	(0.029)
	Age 20-30	(ref)	
	Age 31-40	0.229	(0.080)
	Age 41-50	0.026	(0.027)
	Age 51-60	0.080	(0.013)
	Age 61-70	-0.070	(0.020)
	Age 71 or older	-0.056	(0.008)
	Sick	0.746	(0.017)
Observations		271,910	

Note: Table shows simulated maximum likelihood estimates of hospital demand model. Prices are measured in millions of COP. Distance is measured in kilometers. Includes hospital fixed effects. Bootstrap standard errors in parenthesis based on 100 resamples of consumers.

6 Partial Equilibrium Analysis

Using my model estimates, I conduct two partial equilibrium analyses that reveal the relative importance of gatekeeping and information frictions on access to care. To quantify the effect of gatekeeping, I set $\beta_i = 0$, and to quantify the impact of information frictions, I set $\sigma_\nu = 0$. In each scenario, I recompute individuals' choice probabilities and present summary statistics of different measures: quality rank of the chosen alternative, price of the chosen alterna-

tive, distance of the chosen alternative, consumer surplus, and the demand elasticity with respect to distance and OOP price. Appendix 6 reports the expressions to compute these measures.[13]

TABLE 5: Partial Equilibrium Results

Measure	Observed	No gatekeeping	No information frictions
Quality rank	48.0 [28.7, 63.2]	30.3 [16.8,43.8]	48.0 [28.4, 63.6]
Price†	0.23 [0.15, 0.27]	0.42 [0.30, 0.44]	0.23 [0.15, 0.27]
Distance	7.62 [4.64, 9.79]	7.52 [4.62, 9.95]	7.62 [4.64, 9.79]
Consumer surplus per capita†	-0.30 [-0.53, -0.04]	0.16 [-0.01, 0.34]	-0.30 [-0.53, -0.04]
Distance elasticity	-0.49 [-0.76, -0.21]	-0.55 [-0.87, -0.28]	-0.49 [-0.76, -0.21]
OOP price elasticity	-0.17 [-0.21, -0.05]	-0.24 [-0.31,-0.16]	-0.17 [-0.22, -0.05]

Note: Table shows the mean and 25th and 75th percentiles of the distribution of several measures in the observed scenario, the scenario without gatekeeping (setting $\beta_i = 0$), and the scenario without information frictions (setting $\sigma_\nu = 0$). Quality rank is the average rank of the $\hat{\xi}_h$s weighted by the choice probabilities, price is the average price weighted by the choice probabilities, consumer surplus is the maximum expected utility divided by $-(\alpha_i + \sigma_p \omega_i)$, distance elasticity is the demand elasticity with respect to d_{ih}, and OOP price elasticity is the demand elasticity with respect to $r_i p_{jh}$. (†) Measured in million COP.

Table 5 presents the mean and 25th and 75th percentiles of the distribution (in brackets) of each measure under the observed scenario and the scenarios without gatekeeping and without information frictions. First of all, I find virtually no impacts of information frictions on the choices that consumers make relative to the observed scenario, which goes in line with σ_ν being relatively small in my estimates.

Instead, eliminating gatekeeping has strong effects on consumers' choices. I find that without gatekeeping consumers would choose significantly higher-quality hospitals relative to the observed scenario (as seen by the average quality rank), which are also nearly twice as expensive on average. These hospitals are relatively closer to where consumers live, although differences

[13] My results correspond to a partial equilibrium because I assume that admission prices do not change as a result of banning gatekeeping practices or eliminating information frictions. A full counterfactual analysis would require a pricing model such as Nash-in-Nash bargaining to predict prices under each policy and is left for future research.

in expected distance compared to the observed scenario are small. These results emphasize a classic price-quality trade-off: gatekeeping is an important mechanism to contain health care spending in managed care insurance markets, but its use comes at the cost of having consumers visit less preferred hospitals. The trade-off is further substantiated by the finding that consumer surplus is negative on average in the observed scenario, but eliminating gatekeeping increases this surplus by a factor of 1.5.

In the last two rows of Table 5 I also explore changes in the elasticity with respect to distance and OOP price. Note that without gatekeeping the relative weight of distance and OOP price on consumers' choices should be higher, and thus we should expect to find greater elasticities (in absolute value). Indeed, the elasticity with respect to distance increases 12% and the elasticity with respect to OOP prices increases 41% (in absolute value) relative to the observed scenario.

7 Conclusions

In this paper I show that health insurers shape the way in which healthcare is provided to patients by engaging in gatekeeping practices. Gatekeeping has stronger effects on healthcare utilization and spending compared to demand-side cost-sharing. To identify the impact of gatekeeping, I leverage the discontinuity in coinsurance rates introduced by the out-of-pocket (OOP) maximum. I use data from the Colombian healthcare system where cost-sharing rules are determined by the government and standardized across insurers and hospitals.

I show that patients in my setting reach their OOP maximum as-if-randomly. Those who reach their maximum consume significantly cheaper services and are substantially less likely to make claims afterwards. These results are at

odds with behavioral assumptions about consumers when they face zero prices and are in favor of insurer gatekeeping driving consumer choices. Estimates from a structural model of hospital demand show that gatekeeping induces patients to choose lower-quality, cheaper hospitals, while information frictions have negligible effects on consumer choices.

In the discussion of how best to deliver health insurance coverage, these findings suggest a role for private insurers as buffers of patient moral hazard. However, recent media attention to cases where gatekeeping has prevented patients from receiving adequate care for their chronic health conditions, indicates that government oversight of gatekeeping practices is needed.

References

Baicker, K., Mullainathan, S., and Schwartzstein, J. (2015). Behavioral hazard in health insurance. *The Quarterly Journal of Economics*, 130(4):1623–1667.

Baker, L. C. and Corts, K. S. (1996). Hmo penetration and the cost of health care: Market discipline or market segmentation? *The American economic review*, 86(2):389–394.

Brekke, K. R., Nuscheler, R., and Straume, O. R. (2007). Gatekeeping in health care. *Journal of health economics*, 26(1):149–170.

Brot-Goldberg, Z. C., Burn, S., Layton, T., and Vabson, B. (2023). Rationing medicine through bureaucracy: Authorization restrictions in medicare. Technical report, National Bureau of Economic Research.

Brot-Goldberg, Z. C., Chandra, A., Handel, B. R., and Kolstad, J. T. (2017). What does a deductible do? the impact of cost-sharing on health care prices,

quantities, and spending dynamics. *The Quarterly Journal of Economics*, 132(3):1261–1318.

Brown, Z. Y. (2019). Equilibrium effects of health care price information. *Review of Economics and Statistics*, 101(4):699–712.

Buitrago, G., Miller, G., and Vera-Hernández, M. (2021). Cost-sharing in medical care can increase adult mortality risk in lower-income countries. *medRxiv*, pages 2021–03.

Buitrago, G., Rodriguez-Lesmes, P., Serna, N., and Vera-Hernandez, M. (2024). The role of hospital networks in individual mortality. *Working paper*.

Caetano, C., Callaway, B., Payne, S., and Rodrigues, H. S. (2022). Difference in differences with time-varying covariates. *arXiv preprint arXiv:2202.02903*.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Chandra, A., Flack, E., and Obermeyer, Z. (2021). The health costs of cost-sharing.

Chandra, A., Gruber, J., and McKnight, R. (2010). Patient cost-sharing and hospitalization offsets in the elderly. *American Economic Review*, 100(1):193–213.

Chandra, A., Gruber, J., and McKnight, R. (2014). The impact of patient cost-sharing on low-income populations: Evidence from massachusetts. *Journal of health economics*, 33:57–66.

Colonnelli, E., Prem, M., and Teso, E. (2020). Patronage and selection in public sector organizations. *American Economic Review*, 110(10):3071–99.

Dague, L. (2014). The effect of medicaid premiums on enrollment: A regression discontinuity approach. *Journal of Health Economics*, 37:1–12.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996.

Drake, C., Anderson, D., Cai, S.-T., and Sacks, D. W. (2023). Financial transaction costs reduce benefit take-up: Evidence from zero-premium health insurance plans in colorado. *Journal of Health Economics*, 89:102752.

Dumontet, M., Buchmueller, T., Dourgnon, P., Jusot, F., and Wittwer, J. (2017). Gatekeeping and the utilization of physician services in france: Evidence on the médecin traitant reform. *Health policy*, 121(6):675–682.

Dunn, A., Gottlieb, J. D., Shapiro, A. H., Sonnenstuhl, D. J., and Tebaldi, P. (2024). A denial a day keeps the doctor away. *The Quarterly Journal of Economics*, 139(1):187–233.

Forrest, C. B. (2003). Primary care gatekeeping and referrals: Effective filter or failed experiment? *BMJ*, 326(7391):692–695.

Glazer, J. and McGuire, T. G. (2000). Optimal risk adjustment in markets with adverse selection: an application to managed care. *American Economic Review*, 90(4):1055–1071.

Glied, S. (2000). *Managed care*, volume 1. Elsevier.

Godager, G., Iversen, T., and Ma, C.-t. A. (2015). Competition, gatekeeping, and health care access. *Journal of Health Economics*, 39:159–170.

Gottlieb, J. D., Shapiro, A. H., and Dunn, A. (2018). The complexity of billing and paying for physician care. *Health Affairs*, 37(4):619–626.

Gowrisankaran, G., Nevo, A., and Town, R. (2015). Mergers when prices are negotiated: Evidence from the hospital industry. *American Economic Review*, 105(1):172â203.

Handel, B. R. and Kolstad, J. T. (2015). Health insurance for "humans": Information frictions, plan choice, and consumer welfare. *American Economic Review*, 105(8):2449–2500.

Handel, B. R., Kolstad, J. T., and Spinnewijn, J. (2019). Information frictions and adverse selection: Policy interventions in health insurance markets. *Review of Economics and Statistics*, 101(2):326–340.

Ho, K. (2006). The welfare effects of restricted hospital choice in the us medical care market. *Journal of Applied Econometrics*, 21(7):1039–1079.

Ho, K. and Lee, R. S. (2017). Insurer competition in health care markets. *Econometrica*, 85(2):379–417.

Iizuka, T. and Shigeoka, H. (2022). Is zero a special price? evidence from child health care. *American Economic Journal: Applied Economics*, 14(4):381–410.

League, R. (2023). Administrative burden and consolidation in health care: Evidence from medicare contractor transitions.

Lin, H. and Sacks, D. W. (2019). Intertemporal substitution in health care demand: Evidence from the rand health insurance experiment. *Journal of Public Economics*, 175:29–43.

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

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47(1):3–13.

Prager, E. (2020). Healthcare demand under simple prices: Evidence from tiered hospital networks. *American Economic Journal: Applied Economics*, 12(4):196–223.

Roberts, J., McDevitt, R., Eliason, P., Leder-Luis, J., and League, R. (2021). Enforcement and deterrence of medicare fraud: The case of non-emergent ambulance rides. *NBER Working Paper*, (29491).

Ruiz, F., Amaya, L., Garavito, L., and Ramírez, J. (2008). *Precios y Contratos en Salud: Estudio Indicativo de Precios y Análisis Cualitativo de Contratos*. Ministerio de la Protección Social.

Serna, N. (2021). Cost sharing and the demand for health services in a regulated market. *Health Economics*, 30(6):1259–1275.

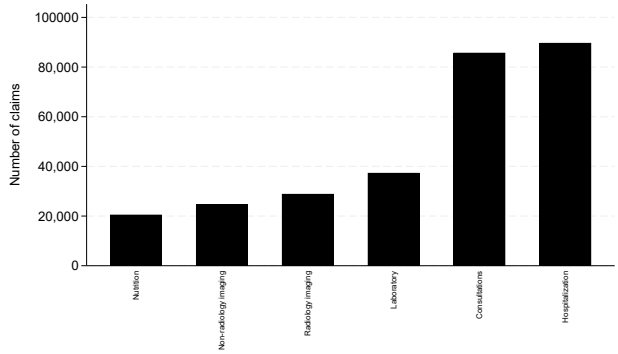
Shigeoka, H. (2014). The effect of patient cost sharing on utilization, health, and risk protection. *American Economic Review*, 104(7):2152–84.

Svanberg, K. (2002). A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM journal on optimization*, 12(2):555–573.

Appendix 1 Additional descriptives

This appendix presents the frequency of services delivered during the week when individuals reach their OOP maximum.

Appendix Figure 1: Most Expensive Types of Claims

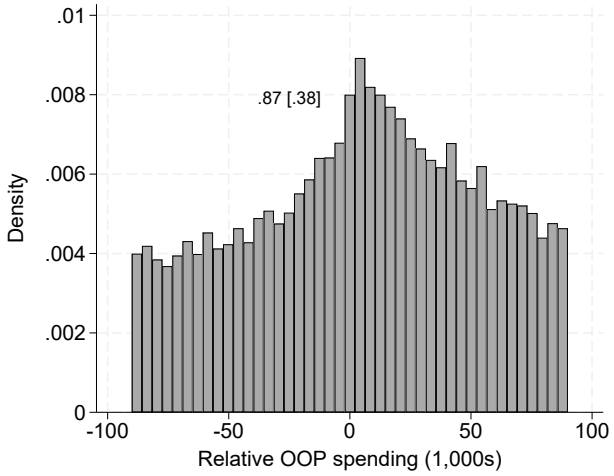


Note: Figure shows the frequency of the top 6 most expensive types of services claimed by individuals who reach their OOP maximum in the week prior to reaching this limit. Figure uses claims-level data conditional on individuals who ever reached their OOP maximum.

Appendix 2 Test of RD Assumptions

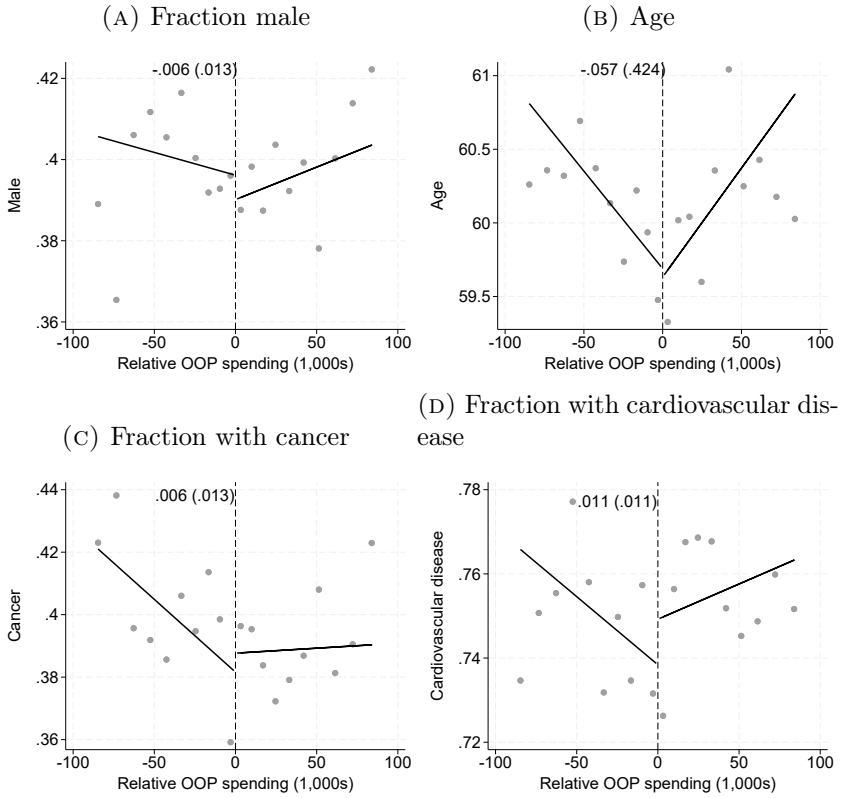
This appendix presents evidence towards the assumptions for a valid RD design. The RD design in the main text focuses on individuals with low incomes who were ever diagnosed with a chronic disease. Appendix Figure 2 presents the distribution of the running variable and Appendix Figure 3 presents RD plots on covariates to test for covariate smoothness around the cutoff.

Appendix Figure 2: Distribution of Running Variable



Note: Figure presents the distribution of relative OOP spending relative to the OOP maximum within a bandwidth of 90 thousand pesos around the cutoff of zero. Figure uses the sample of individuals who ever reached the OOP maximum, had income below 2 times the monthly minimum wage, and were ever diagnosed with a chronic disease. The label in the figure presents the McCrary (2008) test statistic and its associated p-value in brackets.

Appendix Figure 3: Covariate Smoothness Around Cutoff for Relative OOP Spending

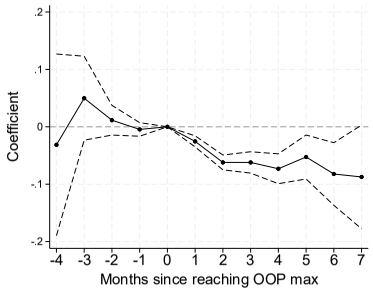


Note: Figure presents RD plots for the fraction of males, age, fraction of individuals with any cancer, and fraction of individuals with cardiovascular disease in Panels A to D, respectively. Estimation uses the sample of individuals with relative OOP spending (relative to the OOP maximum) within a bandwidth of 90 thousand pesos around the cutoff of zero, who ever reached the OOP maximum, had income below 2 times the monthly minimum wage, and were ever diagnosed with a chronic disease. The label in each Panel presents the RD estimate and its standard error in parenthesis.

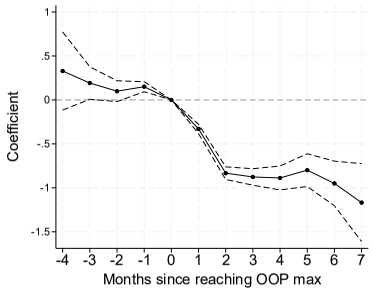
Appendix Figure 4: Utilization and Average Prices with Alternative Bandwidths

Panel (i): Bandwidth = 60 thousand pesos

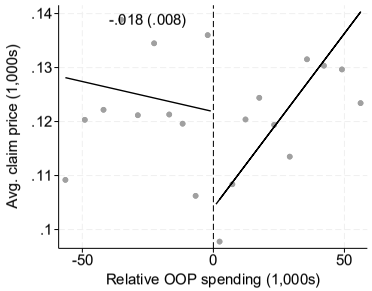
(A) did: Avg. claim price



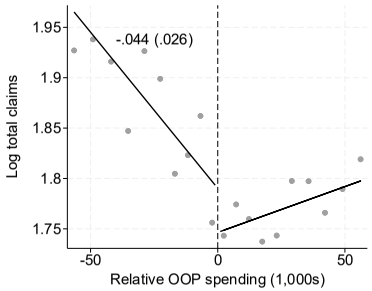
(B) did: Log total claims



(C) RD: Avg. claim price

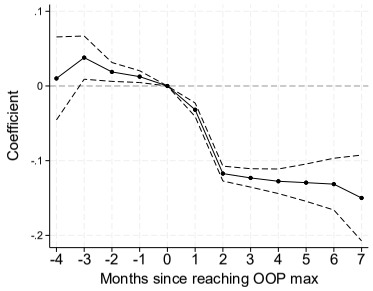


(D) RD: Log total claims

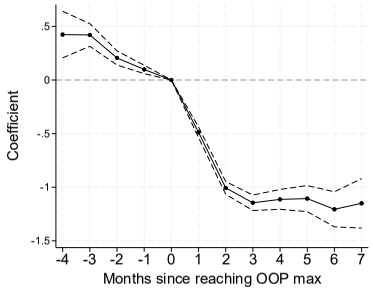


Panel (ii): Bandwidth = 120 thousand pesos

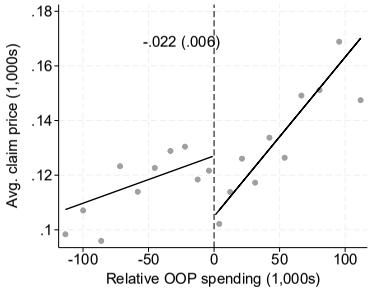
(E) did: Avg. claim price



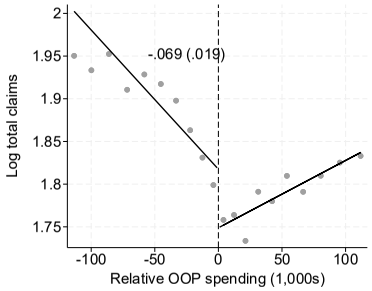
(F) did: Log total claims



(G) RD: Avg. claim price



(H) RD: Log total claims



Note: Panels A to D show event study and RD results for the average claim price and the log of total claims using a bandwidth of 60 thousand pesos around the zero cutoff for relative OOP spending. Panels E to H use a bandwidth of 120 thousand pesos. In Panels A, B, E, F estimation uses De Chaisemartin and D'Haultfouille's estimator. Standard errors are clustered at the individual level. Panels C, D, G, H are conditional on treated individuals. Labels in these panels show the RD point estimate and its standard error in parenthesis. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages and were diagnosed with a chronic disease at any point during the sample period. All specifications control for relative OOP spending.

Appendix 3 Event Study Coefficients

In this appendix I present event study coefficients and standard errors used to construct each figure in the main text.

Appendix Table 1: Main Event Study Coefficients

	Avg. claim price		Log total claims	
Months to treat	coef	se	coef	se
-4	-0.008	(0.046)	0.413	(0.148)
-3	0.027	(0.023)	0.301	(0.070)
-2	0.015	(0.009)	0.189	(0.042)
-1	0.003	(0.005)	0.118	(0.023)
0	—	—	—	—
1	-0.030	(0.005)	-0.431	(0.026)
2	-0.088	(0.005)	-0.907	(0.032)
3	-0.096	(0.007)	-1.029	(0.041)
4	-0.105	(0.011)	-0.979	(0.056)
5	-0.087	(0.014)	-0.955	(0.079)
6	-0.088	(0.019)	-1.070	(0.105)
7	-0.107	(0.033)	-1.065	(0.153)
Observations	45,511		45,511	

Note: Coefficients and standard errors in parenthesis of the event study specifications using the average claim price and the log of total claims as outcomes. Estimation uses De Chaisemartin and D'Haultfouille's estimator. Standard errors are clustered at the individual level. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum. All specifications control for relative OOP spending.

Appendix Table 2: Event Study Coefficients for Zero-price and Health Shock Effects

	Zero-price		Health shock	
Months to treat	coef	se	coef	se
-4	-0.039	(0.086)	-0.003	(0.046)
-3	0.019	(0.028)	-0.109	(0.011)
-2	0.008	(0.011)	-0.021	(0.066)
-1	0.005	(0.006)	-0.032	(0.018)
0	—	—	—	—
1	-0.034	(0.006)	-0.042	(0.014)
2	-0.101	(0.007)	-0.080	(0.015)
3	-0.100	(0.009)	-0.093	(0.022)
4	-0.120	(0.015)	-0.040	(0.027)
5	-0.098	(0.022)	-0.014	(0.021)
6	-0.117	(0.030)	0.027	(0.031)
7	-0.160	(0.058)	0.123	(0.094)
Observations	30,845		24,284	

Note: Coefficients and standard errors in parenthesis of the event study specifications using the average claim price as outcome. The zero-price effect uses the sub-sample of individuals who are never hospitalized during the sample period. The health shock effect uses the sample of treated individuals comparing those who are hospitalized are those who are not, before and after the hospitalization. Specifications use De Chaisemartin and D'Haultfouille's estimator. Time indicators are constructed relative to reaching the OOP maximum for the zero-price effect and relative to the month when the individual is hospitalized for the health shock effect. Both specifications control for relative OOP spending. Standard errors are clustered at the individual level. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum.

Appendix Table 3: Event Study Coefficients by Cohort

	June		August		October	
Months to treat	coef	se	coef	se	coef	se
-4	-0.547	(0.279)	0.177	(0.183)		
-3	0.114	(0.143)	-0.031	(0.052)	0.004	(0.038)
-2	0.096	(0.163)	0.006	(0.028)	0.015	(0.015)
-1	-0.014	(0.035)	0.012	(0.021)	-0.013	(0.010)
0	—	—	—	—	—	—
1	0.011	(0.028)	0.004	(0.016)	-0.033	(0.010)
2	-0.109	(0.033)	-0.083	(0.016)	-0.075	(0.009)
3	-0.145	(0.050)	-0.083	(0.020)	-0.077	(0.009)
4	-0.141	(0.086)	-0.100	(0.023)		
5	-0.104	(0.063)	-0.057	(0.019)		
6	-0.132	(0.074)				
7	-0.112	(0.075)				
Observations	16,764		17,573		17,948	

Note: Coefficients and standard errors in parenthesis of the event study specifications using as outcome the average claim price. Specifications use De Chaisemartin and D'Haultfouille's estimator on the sub-sample of treated individuals who reach their OOP maximum in June, August, and October. Standard errors are clustered at the individual level. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum. All specifications control for relative OOP spending.

Appendix Table 4: Event Study Coefficients by Service

	Log inpatient cost		Log outpatient cost		Log imaging cost		Log labs cost	
Months to treat	coef	se	coef	se	coef	se	coef	se
-4	0.010	(0.054)	0.091	(0.053)	-0.007	(0.018)	0.026	(0.018)
-3	-0.029	(0.023)	0.080	(0.023)	0.019	(0.013)	0.017	(0.004)
-2	-0.009	(0.013)	0.054	(0.011)	0.016	(0.006)	0.010	(0.003)
-1	-0.005	(0.007)	0.026	(0.006)	-0.001	(0.003)	0.009	(0.002)
0	—	—	—	—	—	—	—	—
1	-0.003	(0.006)	-0.056	(0.006)	-0.010	(0.003)	-0.013	(0.002)
2	-0.045	(0.007)	-0.226	(0.009)	-0.057	(0.004)	-0.035	(0.002)
3	-0.077	(0.010)	-0.237	(0.011)	-0.063	(0.005)	-0.038	(0.003)
4	-0.081	(0.015)	-0.243	(0.015)	-0.058	(0.007)	-0.037	(0.004)
5	-0.040	(0.020)	-0.237	(0.021)	-0.063	(0.009)	-0.039	(0.005)
6	-0.081	(0.032)	-0.233	(0.028)	-0.061	(0.012)	-0.042	(0.006)
7	-0.113	(0.048)	-0.260	(0.040)	-0.070	(0.018)	-0.047	(0.008)
Observations	45,511		45,511		45,511		45,511	

Note: Coefficients and standard errors in parenthesis of the event study specifications using as outcomes the log of inpatient, outpatient, imaging, and laboratory spending. Estimation uses De Chaisemartin and D'Haultfouille's estimator. Standard errors are clustered at the individual level. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum. All specifications control for relative OOP spending.

Appendix Table 5: Event Study Coefficients for Disease Management Service

	Any chemo/radio		Any A1C		Any AMI	
Months to treat	coef	se	coef	se	coef	se
-4	-0.001	(0.015)	0.010	(0.099)	-0.008	(0.013)
-3	0.010	(0.013)	0.010	(0.048)	0.007	(0.016)
-2	0.024	(0.015)	0.033	(0.030)	0.002	(0.009)
-1	0.022	(0.009)	0.046	(0.023)	-0.001	(0.005)
0	—	—	—	—	—	—
1	-0.008	(0.008)	-0.053	(0.019)	-0.010	(0.004)
2	-0.043	(0.010)	-0.089	(0.020)	-0.014	(0.005)
3	-0.044	(0.008)	-0.120	(0.026)	-0.024	(0.007)
4	-0.044	(0.011)	-0.125	(0.036)	-0.021	(0.007)
5	-0.045	(0.016)	-0.065	(0.048)	-0.022	(0.010)
6	-0.070	(0.021)	-0.010	(0.061)	-0.024	(0.016)
7	-0.038	(0.006)	0.028	(0.095)	-0.003	(0.016)
Observations	17,146		9,195		33,412	

Note: Coefficients and standard errors in parenthesis of the event study specifications using as outcomes using as outcomes an indicator for making chemotherapy or radiotherapy claims conditional on patients with Cancer, an indicator for having A1C blood tests conditional on patients with diabetes, and an indicator for having an Acute Myocardial Infarction (AMI) conditional on patients with cardiovascular disease. Estimation uses De Chaisemartin and D'Haultfouille's estimator. Standard errors are clustered at the individual level. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. All specifications use the sub-sample of individuals who make less than 2 monthly minimum wages, were diagnosed with a chronic disease at any point during the sample period, and are within a bandwidth of 90 thousand pesos at each side of the relative OOP spending, relative to the OOP maximum. All specifications control for relative OOP spending.

Appendix 4 Census Tract Data and Admission Prices

While the claims data reports admission prices that each insurer negotiated with each hospital in its network, these prices sometimes vary with admission characteristics that are unobserved to insurers when they bargain. To average out these characteristics, I estimate the following regression separately for every insurer:

$$p_{cjh} = \lambda_1 + x_c'\lambda_2 + \lambda_h + \upsilon_{cjh}$$

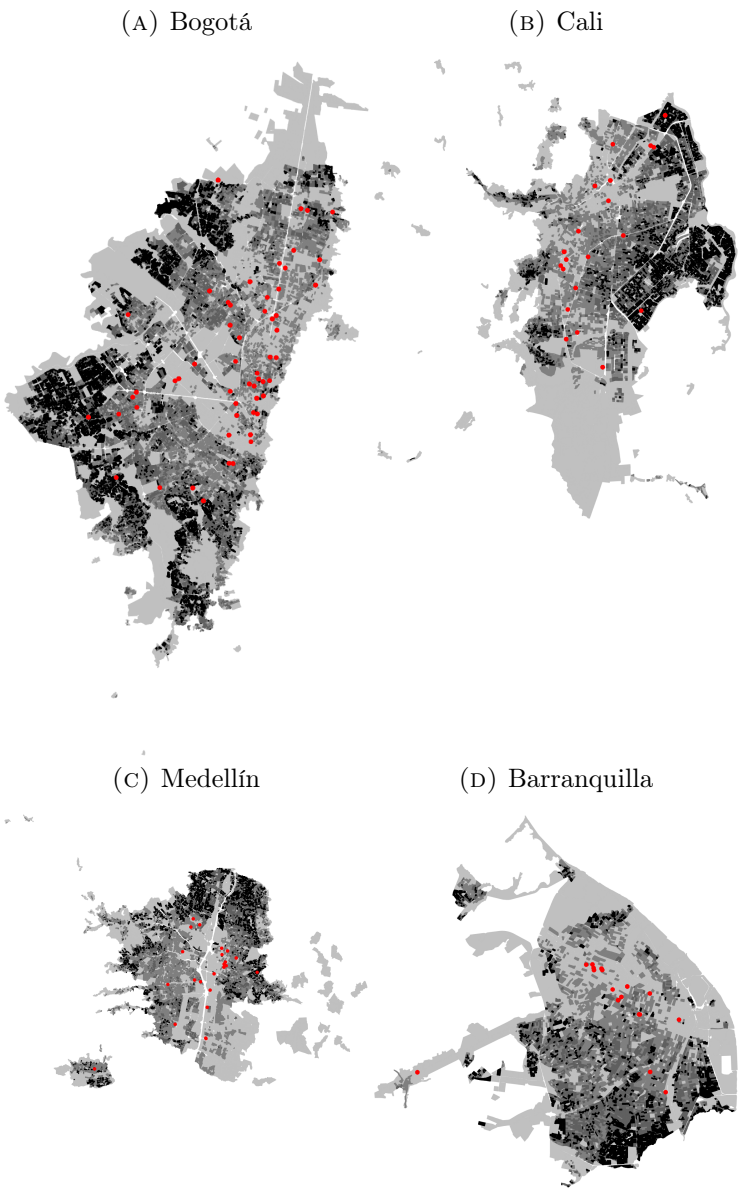
where c is a claim, j is an insurer, and h is a hospital. Moreover, x_c are claim characteristics including patient's sex, age, and length-of-stay; and λ_h are hospital fixed effects. From these regressions I obtain price predictions $\hat{p}_{cjh}$, which I then average across claims for every insurer-hospital pair to calculate the final prices used in my model.

To construct my population-weighted distance measure I use data from the 2018 Colombian census. This data reports population density in each locality within a municipality by age quintile. I limit my analysis sample to the 14 main capital cities in the country. Appendix figure 5 presents the maps for the 4 largest municipalities and their localities: Bogotá, Cali, Medellín, and Barranquilla. Darker colors represent denser localities and red dots correspond to hospitals.

Appendix 5 Estimation Details

Control function. I estimate the model in Section 5 using Simulated Maximum Likelihood. I implement the instrumental variable approach using control

Appendix Figure 5: Hospital locations and census tracts



Note: Census tract level maps for the main capital cities in Colombia using data from the 2018 census: Bogotá in panel A, Cali in panel B, Medellín in panel C, and Barranquilla in panel D. Darker colors represent denser census tracts in terms of population. Red dots correspond to hospitals.

Appendix Table 6: First-Stage Control Function Results

	Negotiated price
Instrument	1.723
	(0.057)
Number of beds	0.004
	(0.0004)
Constant	0.295
	(0.004)
Observations	271,910
R^2	0.004
Adjusted R^2	0.004
F Statistic	485.63

Note: Table shows results from a regression of negotiated prices on the price instrument and the hospital's number of beds. The price instrument if the average reference price from the government weighted by the fraction of beds represented by other hospitals in the choice set. This instrument is measured in millions. Number of beds is meaasured in 100s. Robust standard errors in parenthesis.

function (Petrin and Train, 2010). In the first stage I estimate a regression of the negotiated price between insurer j and provider h on the price instrument (z_{jh}) and the provider's number of beds (b_h). The price instrument is the average government's reference price weighted by the fraction of beds represented by h's competitors in the network for insurer j. Formally, the first stage regression is:

$$p_{jh} = \beta_0 + \beta_1 z_{jh} + \beta_2 b_h + \epsilon_{jh}.$$

The residuals from this regression, $\tilde{p}_{jh} = p_{jh} - \hat{p}_{jh}$, are then added to the utility function in equation (2) as $\gamma_i \tilde{p}_{jh}$, allowing for the same interactions with consumer demographics.

Choice probabilities. To calculate the choice probability implied by the model in equation (2), I approximate the integral over ω numerically as follows:

$$\int \frac{\exp(\delta^1_{ijh}(\omega))}{\sum_{k\in H_j} \exp\left(\delta^1_{ijk}(\omega)\right)} d\phi(\omega) \simeq \frac{1}{R} \sum_{l=1}^{R} \frac{\exp(\delta^1_{ijh}(\omega_{il}))}{\sum_{k\in H_j} \exp\left(\delta^1_{ijk}(\omega_{il})\right)}.$$

For each individual, I generate $R = 300$ independent draws from the standard normal distribution $N(0, 1)$. I use the same set of draws when conducting the partial equilibrium analysis.

Simulated Maximum Likelihood. To find the parameter values that maximize the log-likelihood function, I use the moving asymptotes (MMA) algorithm (Svanberg, 2002), an iterative gradient-based method for nonlinear optimization. To arrive at the gradient of the objective function, I take the derivative of the log-likelihood function in equation (3) with respect to the parameter vector $\tilde{\theta}$:

$$\begin{aligned}\frac{\partial}{\partial\tilde{\theta}}L(y;\tilde{\theta}) =& \sum_i \sum_{h\in H_j} y_{ihj}\frac{\frac{\partial}{\partial\tilde{\theta}}P_{ihj}}{P_{ihj}} - (1-y_{ih})\frac{\frac{\partial}{\partial\tilde{\theta}}P_{ihj}}{1-P_{ihj}} \\ =& \sum_i \sum_{h\in H_j} \left(\frac{y_{ihj}}{P_{ihj}} - \frac{1-y_{ihj}}{1-P_{ihj}}\right)\frac{\partial}{\partial\tilde{\theta}}P_{ihj}\end{aligned}$$

where

$$\frac{\partial}{\partial\tilde{\theta}}P_{ihj} = \gamma_i\frac{\partial}{\partial\tilde{\theta}}P^1_{ihj} + (1-\gamma_i)\frac{\partial}{\partial\tilde{\theta}}P^2_{ihj}.$$

Moreover, the derivative of the choice probability with respect to the parameter vector in each state s is given by:

$$\begin{aligned}\frac{\partial}{\partial\tilde{\theta}}P^s_{ihj} =& \frac{\partial}{\partial\tilde{\theta}}\int \frac{\exp(\delta^s_{ijh})}{\sum_{k\in H_j}\exp(\delta^s_{ijk})}d\phi(w) \\ =& \int \frac{\partial}{\partial\tilde{\theta}}\frac{\exp(\delta^s_{ijh})}{\sum_{k\in H_j}\exp(\delta^s_{ijk})}d\phi(w) && \text{Leibniz integral rule} \\ \simeq& \sum_{l=1}^{R}\frac{\partial}{\partial\tilde{\theta}}\frac{\exp(\delta^s_{ijh})}{\sum_{k\in H_j}\exp(\delta^s_{ijk})}, && \text{Numerical approximation}\end{aligned}$$

with

$$\frac{\partial}{\partial\widetilde{\theta}}\frac{\exp(\delta^s_{ijh})}{\sum_{k\in H_j}\exp(\delta^s_{ijk})} = P^s_{ihj}\Big(\frac{\partial}{\partial\widetilde{\theta}}\delta^s_{ijh} - \sum_k P^s_{ihk}\frac{\partial}{\partial\widetilde{\theta}}\delta^s_{ijk}\Big),$$

$\frac{\partial}{\partial\widetilde{\theta}}\delta^s_{ijh}$ is straightforward as all the parameters enter the utility function linearly, except for σ_ν, which affects the choice probability through γ_i. To ensure that the variance coefficients σ_ν and σ_p are non-negative, I use their logarithm in estimation and modify the gradient accordingly.

I select the starting values by estimating two auxiliary models. First, for the values of σ_ν and σ_p, I estimate a model setting to zero all the provider fixed effects $\xi_h = 0$ and allowing price coefficients to vary only with the indicator for chronic conditions, $\beta_i = \beta + \text{Sick}_i\beta_s$. All starting values for this first auxiliary estimation are set to zero. I take the estimates of the information friction parameter σ_ν as the starting value for the next step.

Then, I estimate a model imposing no unobserved price heterogeneity $\sigma_p = 0$. I also test different starting OOP price coefficients $\beta_0 = 0, -5, ..., -30$ and instrument coefficients $\gamma_0 = -\beta_0$ and select the starting values from the estimation with the maximum log-likelihood. This process yields the following starting values: $\tau = -0.09, \kappa = 1.73, \alpha_0 = -16.03, \beta_0 - 25.30, \gamma_0 = 24.28, \ln\sigma_\nu = -2.80, \ln\sigma_p = 1.36$. The starting values for interaction variables and provider fixed effects are set to 0.

Appendix 6 Expressions for Partial Equilibrium Measures

Individual i's choice probability for hospital h in the network of insurer j is:

$$P_{ijh} = \gamma_i P^1_{ijh} + (1-\gamma_i)P^2_{ijh}$$

where

$$P^1_{ijh} = \int \frac{\exp(\delta^1_{ijh})}{\sum_k \exp(\delta^1_{ijk})} d\phi(\omega), \qquad P^2_{ijh} = \frac{\exp(\delta^2_{ijh})}{\sum_k \exp(\delta^2_{ijk})}$$

and

$$\delta^1_{ijk} = (\alpha_i + \sigma_p \omega_i) r_i p_{jh} + \beta_i p_{jh} + \tau d_{ih} + \kappa x_{ih} + \xi_h, \qquad \delta^2_{ijk} = \beta_i p_{jh} + \tau d_{ih} + \kappa x_{ih} + \xi_h$$

Let ξ^k_h denote the rank order of the fixed effect for hospital h in the demand function, with a lower rank denoting a higher quality. The quality rank of the chosen alternative for consumer i is given by $\sum_{h \in H_j} \xi^k_h P_{ijh}$ and similarly the price and distance of the chosen alternative are given by $\sum_{h \in H_j} p_{jh} P_{ijh}$ and $\sum_{h \in H_j} d_{ih} P_{ijh}$, respectively. Consumer i's surplus is computed as $\log(\sum_{h \in H_j} (\gamma_i \exp(\delta^1_{ijh}) + (1 - \gamma_i) \exp(\delta^2_{ijh})))$. In the main text, I report the mean and 25th and 75th percentiles of these measures across all consumers.

Finally, we report the average demand elasticity with respect to distance and with respect to the OOP price weighted by the choice probability which are computed respectively as:

$$\frac{\sum_i P_{ijh} \tau (1 - P_{ijh}) d_{ih}}{\sum_i P_{ijh}}$$

and

$$\frac{\sum_i P_{ijh} \int (\alpha_i + \sigma_p \omega_i)(1 - P^1_{ijh}) r_i p_{jh} d\phi(\omega)}{\sum_i P_{ijh}}$$