What Do Health Insurers Do? Patient Choices and Utilization Management in Healthcare

Natalia Serna*

March 17, 2025

Abstract

This paper demonstrates that health insurers actively engage in utilization management by influencing patients' medical care choices. First, I demonstrate that healthcare demand responds to the full price of care, even when consumers face zero out-of-pocket expenses, by analyzing the discontinuity in cost-sharing created by the out-of-pocket maximum. I then identify insurer utilization management as the key mechanism driving this effect, examining quasi-random variations in insurer costs. Finally, estimates from a structural model of hospital demand reveal that utilization management manifests in insurers steering patients toward lower-cost, lower-quality hospitals. Findings underscore the crucial role insurers play in controlling healthcare spending.

Keywords: Health insurance, Utilization management, Healthcare spending, Cost-sharing.

JEL codes: I10, I11, I13, I18.

*Stanford University, 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 Maria Polyakova for her 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

Classic economic theory proposes two main ways for regulating spending on goods: price controls and quantity controls. Price controls, such as price ceilings, and quantity controls, such as quantity quotas, can create social costs by preventing access to consumers who are willing to pay more for a good than its price. Both types of controls can decrease the overall number of transactions in a market and lower total spending. The dichotomy between price and quantity controls is present in many markets, including health insurance, where rising healthcare spending presents ongoing challenges for insurers and regulators. Insurers use cost-sharing strategies to directly control the prices patients face for healthcare services, while increasingly relying on utilization management to nudge patients' medical care choices and regulate the quantity of care received. The questions of whether to apply price or quantity controls in healthcare markets, and how utilization management impacts healthcare market outcomes, deserve more research.

Within this context, utilization management practices—such as prior authorization, narrow provider networks, and claim denials—have attracted significant media attention due to concerns over their potential to adversely affect patient health and to create administrative burdens for both doctors and patients (Ofri, 2014; Span, 2024; Stockton, 2024). However, evidence regarding the existence and impacts of utilization management has been limited, as directly observing these practices in data often proves challenging. In this paper, I demonstrate that insurers effectively employ utilization management to guide patients toward less preferred healthcare providers, resulting in significantly lower healthcare spending. I outline my findings in three steps: first, I establish that healthcare demand responds to the full price of care even when consumers face zero out-of-pocket (OOP) prices. Next, I identify utilization management as the underlying mechanism driving this phenomenon. Finally, I provide evidence that this management approach often manifests in insurers directing patients to lower-cost, lower-quality hospitals.

My empirical setting is the Colombian contributory health care system, where private insurers operate under a heavily regulated environment. These insurers offer a single contract and compete primarily based on their network of covered healthcare providers. Eligibility

for this system extends to individuals who either pay payroll taxes or are dependents of taxpayers, accounting for nearly half of the country's population. Enrollees do not pay premiums but are responsible for copayments and coinsurance rates for each healthcare service up to the OOP maximum. Beyond this threshold consumer OOP prices drop to zero and the insurer covers the full cost of care. These cost-sharing rules are indexed to the enrollee's monthly income and are designed to be progressive, meaning that low-income consumers face a lower OOP maximum compared to their higher-income counterparts.

To explore the impacts of utilization management, I analyze individual-level panel data from the contributory system covering more than 8 million enrollees between 2009 and 2011. I start by presenting descriptive evidence that patients appear to respond to the full price of care even when they are not responsible for these prices out of pocket. In other words, healthcare demand slopes down in the full price of care while controlling for consumers' OOP prices; an effect that is identified from the discontinuity in cost-sharing created by the OOP maximum.

The cost-sharing structure in Colombia, which is linked to monthly income levels, allows me to assess whether the relationship between full healthcare prices and demand is causal. I analyze variations in OOP maximums across different income groups, specifically comparing low- and high-income consumers with similar health statuses who experience a sudden hospitalization.[1] Notably, this hospitalization pushes only the low-income consumer over the OOP maximum, creating quasi-random variation in OOP prices across consumers. My findings reveal that low-income consumers, even with zero OOP prices, file 42% fewer claims and incur 38% lower healthcare spending after reaching the OOP maximum compared to their high-income counterparts.

The reduction in utilization and spending spans all healthcare services. Low-income consumers make 26% fewer primary care claims, 16% fewer specialist care claims, and 23% fewer urgent care claims. These trends emphasize the importance of utilization management once patients hit their OOP maximum. In Colombia, patients must obtain a referral from their primary care physician to access specialized services. Therefore, the decrease in primary care

[1] Consumers in different income groups are observationally equivalent in the sense that their utilization patterns before the OOP maximum are parallel. Additionally, the onset of hospitalization is "sudden," as it cannot be anticipated based on consumers' income levels.

usage suggests that insurers may have incentives to control downstream costs by restricting patient access to entry-level services.

I perform a series of robustness checks to strengthen the causal interpretation of these results. For instance, I demonstrate that the findings remain consistent when analyzing consumers within a narrow bandwidth near the income cutoff that establishes the cost-sharing rules, resembling a difference-in-discontinuities design. The results also hold when excluding high-income consumers who reach their OOP maximum and when examining dependents rather than the primary enrollees, to whom slightly different cost-sharing rules apply.

The next part of the analysis investigates the mechanisms that contribute to the reduction in claims when OOP prices are zero. By comparing individuals across income groups who share similar characteristics, I can eliminate several potential explanations for this effect, such as mean reversion, changes in health status, and information frictions. I then explore whether provider responses to payment structures factor into these results, as it is possible that providers receive lower reimbursements or encounter greater financial risk after patients reach their OOP maximum, which could discourage the provision of services. However, I find no evidence to support this as a significant mechanism underlying my findings.

A key mechanism that may explain the price sensitivity of healthcare demand is insurers' engagement in utilization management. To examine this, I leverage a 2011 policy change in which the Colombian government expanded the list of covered services under the national insurance plan while maintaining fixed cost-sharing rules—additionally, premiums are always zero and enrollment is mandatory. This policy disproportionately increases insurers' costs for low-income consumers at zero OOP prices relative to high-income ones, but does not affect consumer total OOP spending. My findings indicate that low-income consumers who reach their OOP maximum after 2011 utilize significantly fewer services compared to their counterparts, which aligns with the concept of insurers disproportionately engaging in utilization management when patients become more expensive.

In Colombia, insurers primarily implement utilization management by directing patients to lower-cost or lower-quality hospitals, as most elements of the insurance contract are closely regulated except for the networks of covered providers. To investigate this practice, I develop

and estimate a structural model of hospital demand that captures consumer price sensitivity in two scenarios: before and after reaching their OOP maximum. In this model patients choose a hospital for an admission to maximize their total expected utility across the two scenarios. My estimates show significant responsiveness to prices before and after the OOP maximum, consistent with the reduced-form evidence.

In a partial equilibrium analysis where I eliminate utilization management by setting demand responsiveness to full prices at zero, I find that patients, on average, would choose hospitals that are 13% more expensive and of higher quality relative to the observed scenario, resulting in a 30% increase in consumer surplus per capita. While this exercise does not account for general equilibrium effects associated with the elimination of utilization management, it effectively illustrates that these practices involve steering patients toward different providers.

Ultimately, my results underscore that quantity controls in the form of utilization management are essential for controlling healthcare spending by nudging patients' medical care choices. In this context, quantity controls prove more effective than price controls. This finding is particularly significant given the limited evidence surrounding the existence and effects of these practices by insurers and the growing interest in regulating them (Kyle and Song, 2023; Anderson et al., 2022; Gaines et al., 2020). While lower healthcare utilization and suboptimal patient choices may jeopardize health outcomes (Buitrago et al., 2024), my results confirm common critiques of utilization management while simultaneously providing evidence that these strategies effectively reduce healthcare spending.

1.1 Related literature

The study of price controls in healthcare has captivated a significant portion of the health economics literature. For instance, previous empirical research has demonstrated that healthcare demand is highly responsive to patient cost-sharing (e.g., Chandra et al., 2010; Shigeoka, 2014; Chandra et al., 2014; Baicker et al., 2015; Serna, 2021; Chandra et al., 2021; Buitrago et al., 2021), and that the magnitude of these responses can vary depending on whether consumers react to the spot or shadow price of care (Aron-Dine et al., 2015; Brot-Goldberg et al., 2017; Lin and Sacks, 2019). This paper advances the literature by examining quantity

controls in the form of utilization management. I demonstrate that patients respond to the full price of care, even when they do not incur these prices, and that insurers' utilization management practices are the primary mechanism driving this effect. This contrasts with findings from Iizuka and Shigeoka (2022); Drake et al. (2023), who find that zero OOP prices lead to increased healthcare utilization and insurance coverage, respectively.

This paper also adds to the growing literature on insurers' utilization management strategies, including narrow provider networks (Ho, 2006), spending monitoring programs (Alpert et al., 2024), prior authorization requirements (Brot-Goldberg et al., 2023), and claim denials (Gottlieb et al., 2018; League, 2023; Dunn et al., 2024). I provide evidence of an alternative utilization management approach, specifically that insurers direct patients toward lower-cost, lower-quality hospitals within their networks. These non-price, steering mechanisms to contain potentially unnecessary healthcare spending have become increasingly popular since the advent of managed care (Glied, 2000; Glazer and McGuire, 2000). However, empirical evidence demonstrating their existence and causal impacts has remained limited.

The remainder of this paper is structured as follows: section 2 describes the empirical setting, section 3 describes the data, section 4 presents the empirical analysis to identify price sensitivity under zero OOP prices, section 5 presents the empirical analysis to identify utilization management, section 6 presents the structural model of hospital demand and results from the partial equilibrium analyses, and section 7 concludes.

2 Utilization Management and Cost-Sharing in Colombia

Colombia's healthcare system, established in 1993, operates under two primary schemes: contributory and subsidized. The contributory scheme serves individuals who pay payroll taxes, along with their dependents, while the subsidized scheme is designed for those living in poverty and is entirely funded by the government through tax revenues. Enrollees in both schemes have access to the national health insurance plan, which is delivered by private

insurers.[2]

The government oversees various aspects of the national health plan: insurance premiums are waived in both schemes, while individuals in the contributory scheme are responsible for a portion of their healthcare costs through cost-sharing. In contrast, healthcare is largely free for those in the subsidized scheme, with only minimal copayments required. While insurers have no leeway in shaping these key components of the insurance plan, they do have the authority to choose which healthcare providers are available to their enrollees.

Even within the typically limited provider networks they set up (Serna, 2024), insurers may mandate prior authorization for specialized care, inpatient or urgent care admissions, and pharmaceuticals not covered by the health plan. Additionally, insurers have the right to deny claims when patients fail to make their monthly payments to the system and when they request healthcare services that fall outside the scope of the national health plan. In these cases, a scientific committee within the insurer determines appropriate covered services that can serve as alternatives and offers those instead.[3]

TABLE 1: Cost-Sharing Rules in the Contributory Health Care System in 2011

Monthly income level	Copay	Coinsurance rate per claim	OOP maximum per year
Income $<$ 2×MMW	2,100	11.5%	57.5% × MMW
Income $\in [2,5]$×MMW	8,300	17.3%	230% × MMW
Income $>$ 5×MMW	21,700	23.0%	460% × MMW

Note: Table shows the copay, coinsurance rate, and OOP maximum by income level that apply to individuals enrolled in Colombia's contributory health care system. The monthly minimum wage (MMW) in 2011 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.

Besides the quantity controls available to insurers, the Colombian health care system also imposes price controls such as cost-sharing to control patient moral hazard. In the contributory scheme, cost-sharing rules are determined by the enrollee's monthly income but are uniform across insurers and hospitals. These rules comprise a three-tiered system that includes copayments, coinsurance rates, and annual maximum out-of-pocket (OOP)

[2]Some insurers offer supplementary plans, which offer services carved-out of the national health plan, and have relatively low market shares.

[3]For more detailed description see Decree 780 of 2016, Article 2.5.3.2.7. and Law 1438 of 2011, Article 27. For instance, inpatient admissions must receive prior approval from the insurer within 2 hours of the patient receiving urgent care. Similarly, any additional services during the hospital stay require approval within 6 hours.

expenses, as illustrated in Table 1 for 2011. Enrollees are categorized based on their income: those earning less than 2 times the monthly minimum wage (MMW), between 2 and 5 times, or more than 5 times the MMW, which was approximately $270 in 2011. For example, individuals earning less than 2 times the MMW face a copayment of 2,100 pesos (around $1), a coinsurance rate of 11.5% on each health claim, and an OOP maximum that is reset annually at 57.5% of the MMW.

All enrollees are required to make copayments whenever they visit a primary care doctor, a specialist, or undergo laboratory or diagnostic tests in an outpatient setting. Dependents—family members eligible for the contributory scheme through the primary contributor or head of the household—are the only enrollees responsible for paying coinsurance rates for any health services they use, except in cases where copayments are applicable.[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. For the rest of the analysis, I denote as "low-income" all individuals who make less than 2 times the MMW and as "high-income" those who make at least 2 times the MMW.

Enrollees in the contributory system must report their income monthly using the Integrated Payments Settlement Form (PILA, from its Spanish acronym). Independent workers are responsible for reporting any income changes, while employers handle these updates for formal workers. Changes in income reported during month t will take effect on cost-sharing rules at least 30 days later. Additionally, the government establishes deadlines for income reports and contributions based on the last two digits of the enrollee's ID number.[5]

3 Data and Descriptives

My raw data comprise health claims from a random sample of nearly 8.7 million enrollees in Colombia's contributory system between 2009 and 2011, all of whom filed at least one

[4]Information on cost-sharing rules can be found in https://www.minsalud.gov.co/Normatividad_Nuevo/ACUERDO%20260%20DE%202004.pdf

[5]Decree 1406 of 1999 establishes the rules governing income reports to the contributory health care system.

claim and did not change their insurer during this period. These data were organized by the Colombian Ministry of Health and Social Protection. For each individual, I have access to socio-demographic information, including sex, age, municipality of residence (comparable to a county in the U.S.), and their status as either a contributor (main enrollee) or dependent. Additionally, for each health claim, the dataset provides details on the insurer, provider, type of service, diagnosis codes, and negotiated prices.

I observe the average monthly income per year for the contributor or main enrollee allowing me to determine their level of cost-sharing. Even though I do not observe income every month, evidence suggests there is little movement of individuals across income groups within a year. For instance, Serna (2021) found that individuals with average monthly incomes exceeding twice the MMW report earnings within this category for 80% of the months in a year. Those with average monthly incomes below twice the MMW consistently report income in this bracket for 99% of the months in a year.

In the data, contributors are not matched to their dependents. Thus, to obtain the income level that applies to dependents for their cost-sharing rules I proceed as follows. First, I obtain each individual's list of visited providers as well as their number of claims and healthcare spending at each provider. Then, I use the subsample of contributors to estimate a linear regression of income on the number of claims per provider, healthcare spending per provider, and provider, insurer, and year fixed effects. An observation in this regression is an individual-provider-year. Finally, I use the estimates from this regression to predict income for dependents. This procedure matches dependents to contributors based on the providers they visited during the year and the insurer they enroll with. Appendix Figure 1 presents the resulting distribution of contributors and dependents by income group. Predictions show that roughly 82% of dependents are in the low-income bracket compared to 76% of contributors.

With the health claims data, I construct different measures of monthly utilization and spending and determine whether and when consumers 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] For tractability, I choose a random

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

sample of 1 million individuals in each year (3 million in total) and construct a balanced panel of person-months resulting in 36 million observations. I refer to these data as the "full sample." In months during which the individual is not observed in the data, I assign healthcare utilization and spending equal to zero. To control for the confounding bias arising from changes in health status, I perform my main analysis in the sample of enrollees who never received a chronic disease diagnosis and who had a "sudden" hospitalization. I refer to these data as the "analysis sample."[7]

While the claims data does not include denied claims or those requiring prior authorization, I can descriptively assess the presence of utilization management strategies that ultimately influence patients' medical care decisions by examining how healthcare demand responds to pricing when patients are not directly responsible for these prices out of pocket. This approach separates the potential influence of insurer utilization management from price controls in the form of cost-sharing, which do not apply when consumers reach their OOP maximum and face zero OOP prices.

In Table 2 I regress the total number of claims on the OOP price and on the full price of healthcare.[8] Columns (1)-(3) show results from the full sample and columns (4)-(6) from the analysis sample. When accounting for individual health shocks, such as hospitalization status, and recognizing that OOP prices are perfectly collinear with the individual fixed effects, demand for healthcare declines with the full price of care, as seen in columns (1) and (4). This coefficient is identified from the cost-sharing discontinuities introduced by the OOP maximum. I obtain similar results when focusing on individuals who reach the OOP maximum in columns (2) and (5). Furthermore, using only observations following the OOP maximum in columns (3) and (6) yields a negative and significant coefficient for the full price of healthcare. This raises the question: why does healthcare demand respond to full prices when OOP expenses are zero?

The remainder of this paper is dedicated to addressing this question. I will proceed in two steps: first, quantifying the causal impact of full prices on healthcare demand to demonstrate

the variation within years.

[7] I determine the list of diseases for each individual by mapping diagnosis codes to diseases following https://www.alvaroriascos.com/researchDocuments/healthEconomics/CLD_xCIE10.tab.

[8] The OOP price equals the copay for contributors and equals the copay plus the coinsurance rate times the health service price for dependents.

that results in Table 2 are meaningful rather than noise, and second investigating utilization management as a key mechanism for this effect.

TABLE 2: Healthcare Demand: Total Number of Claims

Variable	Full sample			Analysis sample		
	(1)	(2)	(3)	(4)	(5)	(6)
OOP price	—	—	—	—	—	—
Full price	-0.468***	-0.936***	-1.074***	-1.071***	-1.085**	-1.130**
	(0.038)	(0.159)	(0.194)	(0.300)	(0.469)	(0.550)
Any hospitalization	19.667***	32.137***	30.795***	16.158***	29.800***	28.953***
	(0.073)	(0.341)	(0.420)	(0.112)	(0.782)	(0.952)
Constant	1.725***	5.887***	7.484***	1.985***	2.956***	4.310***
	(0.001)	(0.038)	(0.071)	(0.009)	(0.081)	(0.146)
Fixed effects						
Individual	✓	✓	✓	✓	✓	✓
Month	✓	✓	✓	✓	✓	✓
Observations	36000000	295920	144651	835932	53892	25740
R-squared	0.384	0.313	0.428	0.293	0.199	0.360

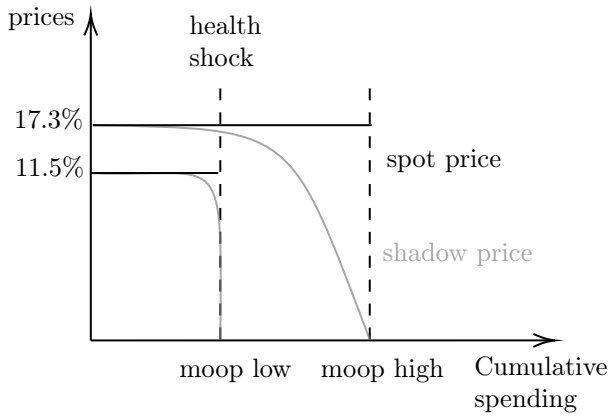
Note: Table presents regression results using as outcome the total number of claims. An observation is a person-month. Columns (1)-(3) use the full sample. Columns (4)-(6) use the analysis sample. Columns (1) and (3) use information from all individuals, columns (2) and (4) from individuals who ever reach the OOP maximum, and columns (3) and (6) from individuals who ever reach the OOP maximum after they reach it. All specifications include individual and month fixed effects. Standard errors in parenthesis are clustered at the individual level.

4 Healthcare Demand at Zero Prices

My empirical approach in the first step of the analysis consists of comparing high- versus low-income individuals, before and after they reach the low OOP maximum. The comparison between income groups helps isolate endogenous changes in health status from potentially sudden changes in OOP prices. The rationale, presented in Figure 1, is as follows: suppose there are two patients who are identical before receiving a sudden health shock except for their income level. This health shock pushes the low-income individual over the low OOP maximum (denoted "moop low" in the figure). The spot price of care for the low-income individual falls to zero at the spending threshold. At the same time, the health shock pushes the high-income individual over the low OOP maximum but not over the high maximum so their spot price of care remains positive.

These quasi-random changes in spot prices for the low-income consumer relative to the

FIGURE 1: Illustration of Identification Strategy for Impact of OOP Prices



Note: Figure illustrates the identifying variation for the price sensitivity of demand for healthcare. The solid black lines represent the spot price of care for the high- and low-income consumer. The solid gray lines represent the shadow price of care for both types of individuals. The dashed vertical lines represent the maximum OOP amount ("moop") for each individual. The health shock impacts both individuals at the same time and pushes the low-income consumer over the OOP maximum.

high-income counterpart help identify the price sensitivity of demand for healthcare. Specifically, I aim to evaluate whether low-income consumers opt for cheaper and fewer services or visit lower-cost providers after reaching the low OOP maximum, which would be consistent with insurers engaging in utilization management as in Table 2.

As an aside, my identification strategy will further highlight the significance of shadow prices: while the comparison of low- versus high-income individuals should identify the correct sign of changes in healthcare demand when OOP prices are zero, the magnitude of the estimate will depend on whether high-income consumers respond to the spot or the shadow price of care. If only spot prices matter, the magnitude of the estimate should not change across different cuts of the data because these prices are constant throughout the distribution of healthcare expenditures for the high-income consumer before they reach the high OOP maximum.

To determine whether income groups are comparable, Table 3 presents some summary statistics of my analysis sample conditional on being below the low OOP threshold. An observation is a person-month. The table shows that high-income consumers are more likely to be male, tend to be older, but have similar total spending, relative spending, and total number of claims before the low OOP maximum. Appendix Figure 3 shows that conditional on being within a narrow bandwidth around the income cutoff (2 times the MMW) and

below the low OOP maximum, these level differences between income groups do not translate into significant trend differences with respect to the relative OOP spending. I will provide robustness exercises focusing on individuals within this narrow bandwidth around the income cutoff, who are more similar in terms of sociodemographic characteristics and healthcare utilization trends and levels as seen in Appendix Table 2.

TABLE 3: Summary Statistics By Income Group Before low OOP Maximum

Variable	High income	Low income
Male	0.45	0.36
	(0.50)	(0.48)
Age	35.47	29.49
	(22.28)	(20.95)
Any hospitalization	0.08	0.09
	(0.28)	(0.29)
Spending relative to low OOP max	-0.23	-0.26
	(0.08)	(0.06)
Average claim price	0.03	0.03
	(0.26)	(0.28)
Total spending	0.18	0.17
	(2.11)	(1.21)
Total number of claims	3.06	3.18
	(8.56)	(8.27)
Outpatient claims	1.57	1.81
	(3.58)	(4.00)
Inpatient claims	0.98	0.96
	(6.20)	(5.62)
Prescription claims	0.46	0.79
	(2.07)	(2.91)
Individuals × Months	120712	667394
Individuals	11873	57441

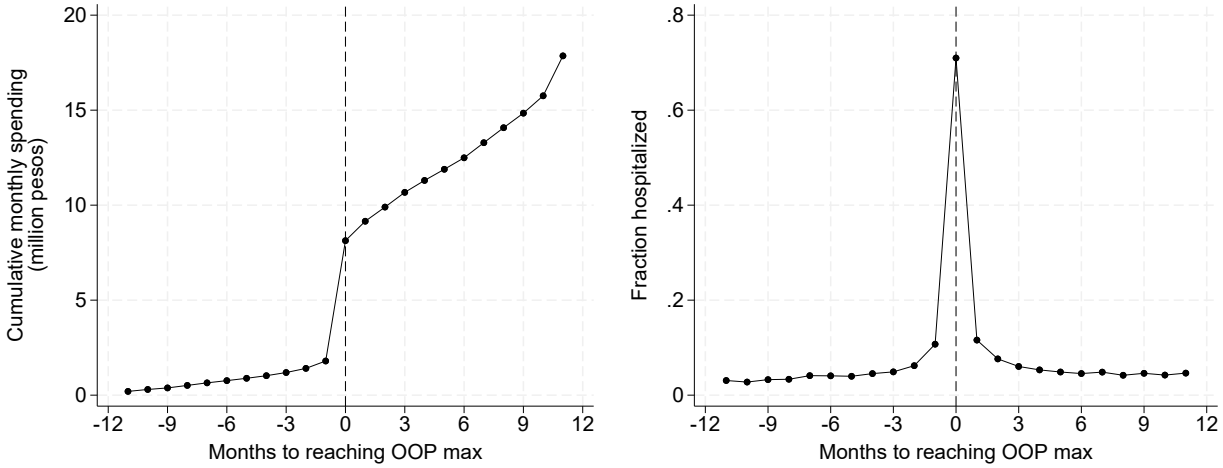
Note: Table presents the mean and standard deviation in parenthesis of consumer characteristics conditional on the period before reaching the low OOP maximum and on individuals in the analysis sample who did not receive a chronic disease diagnosis before reaching the low OOP maximum.

Reaching the OOP maximum in my setting is typically a "sudden" event. Figure 2, Panel A illustrates that for low-income individuals who eventually reach the low OOP maximum, cumulative monthly spending rises steadily until the month before reaching the threshold, followed by a sharp discontinuity at the point when the maximum is reached. This sudden event is typically a hospitalization as seen in Panel B. Around 70% of low-income consumers reach the OOP maximum due to a hospitalization, while the rest either claim an expensive inpatient drug or an expensive doctor consultation during the hospitalization as seen in Appendix Figure 2. Importantly for my empirical analysis, Appendix Table 1 shows that

differences between income groups cannot predict whether individuals have a hospitalization, and in that sense the health shock is sudden. This table reports regression results using as outcome an indicator for having a hospitalization and as regressors lagged healthcare spending, lagged number of claims, and their respective interactions with income level; these interactions are all statistically zero. In other words, potentially unobserved differences between income groups are uncorrelated with the health shocks that create quasi-random variation in OOP prices.

FIGURE 2: Cumulative Spending and Hospitalizations by Month for Low-Income Consumers

(A) Cumulative spending (B) Hospitalizations



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

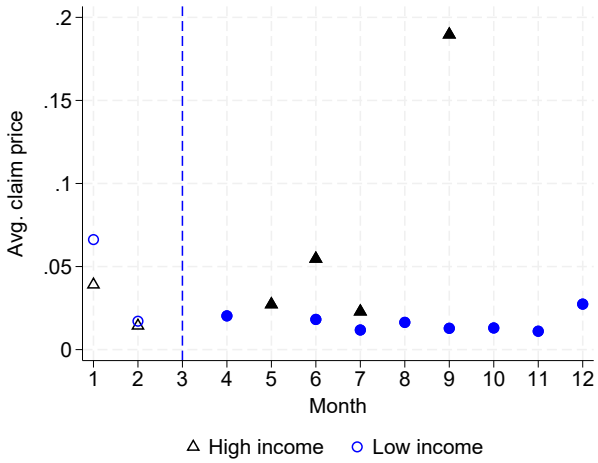
Figure 3 presents specific examples of my empirical approach in the analysis sample. Panel A depicts the time trend of average claim price for two women. Both women are enrolled with the same insurer (*Famisanar*), live in Bogotá, never received a chronic disease diagnosis, and had a hospitalization at the same clinic (*Hospital de San José*) that pushed them over the low OOP maximum. One of these women has high income, depicted in the black triangles, and the other one has low income, depicted in the blue circles. The blue dashed vertical line denotes the month in which the low-income woman reaches the low OOP maximum (the average claim price in this month is excluded for exposition). Comparing the solid black triangles and the solid blue circles after women reach the low OOP maximum reveals that the low-income woman on average consumes cheaper services than the high-income counterpart even though her OOP prices are zero and those of her counterpart are

positive.

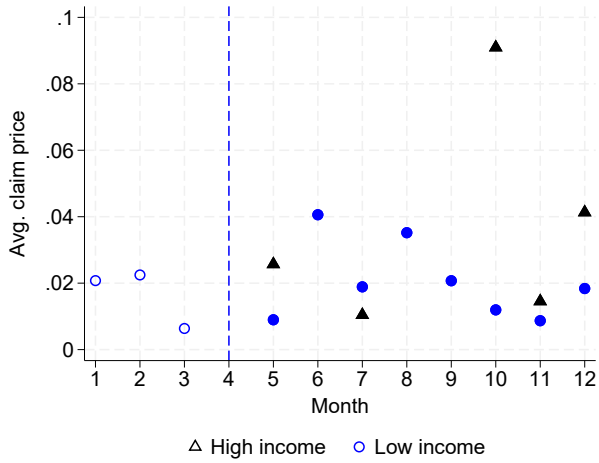
Panel B depicts another example of two men, enrolled with the same insurer (*Sanitas*), who never received a chronic disease diagnosis, live in Bogotá, had a hospitalization at the same clinic (*Clínica Colsanitas*) that pushed them over the low OOP limit, but one of these men is low-income. The figure also shows that on average the low-income man claims cheaper services after reaching the low OOP maximum despite facing zero OOP prices.

FIGURE 3: Example of Data Variation in Average Claim Prices by Income Group

(A) Women who have a hospitalization



(B) Men who have a hospitalization



Note: Panel A presents a scatter plot of average claim prices by month for a high-income woman in the black triangles and a low-income woman in the blue circles. The dashed vertical line represents the month in which the low-income woman reaches the low OOP maximum. Women in this panel are enrolled with the same insurer and had a hospitalization at the same clinic which pushed them over the low OOP maximum. Panel B presents a similar scatter plot for a high- and a low-income man who are enrolled with the same insurer and had a hospitalization at the same clinic that pushed them over the low OOP maximum.

My empirical strategy will make the comparisons in Figure 3 in a more systematic way to determine whether there are differences in healthcare demand after individuals (in different income brackets) reach the low OOP maximum. The regression specification is as follows:

$$y_{it} = \beta LP_{it} + \alpha\, T_i \cdot LP_{it} + \lambda S_{it} + \delta_i + \gamma_t + \varepsilon_{it} \tag{1}$$

where y_{it} is an outcome of individual i in month t, T_i is an indicator for whether the individual is low-income (makes less than 2 times the MMW), LP_{it} is an indicator for whether the individual reaches the low OOP maximum in month t, S_{it} is the consumer's OOP spending minus the low OOP maximum in month t, δ_i are individual fixed effects, and γ_t are month fixed effects. The coefficient of interest is α, which measures the difference in outcomes

between low- and high-income individuals after the low OOP maximum. At this point OOP prices are zero for the low-income consumer but are strictly positive for the high-income one. Thus, a finding that $\alpha < 0$ can be interpreted as the impact of zero OOP prices on outcomes, or conversely as the impact of the full price of care, which would be consistent with insurers engaging in utilization management.

4.1 Threats and Approaches

Identifying the price sensitivity of healthcare after the OOP maximum is a challenging exercise. For one, there is a classic selection bias problem because people who reach the OOP maximum and face zero OOP prices can be unobservably sicker and generally less responsive to prices compared to those who do not reach the maximum. This type of unobserved heterogeneity can lead a researcher to underestimate the price sensitivity. My empirical strategy helps alleviate this concern by comparing individuals between income groups conditional on reaching the low OOP maximum, without having to leverage comparisons between those who do and do not reach the maximum. Additionally, I analyze within-patient changes in outcomes, as unobserved responsiveness to prices may also vary between income groups.

Second, there may be several confounding bias problems. Changes in demand after reaching the low 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 can lead a researcher to overestimate the price sensitivity of healthcare. To address this concern, I focus on low- and high-income individuals who are similar before reaching the low OOP maximum, which can make differences in unobserved characteristics such as health status less salient. Specifically, I use the group of patients who never received a chronic disease diagnosis and who reached the low OOP maximum due to a sudden hospitalization. Table 3 and Appendix Figure 3 summarized the characteristics of low- and high-income individuals before reaching the low OOP maximum, showing that the two groups are comparable in terms of healthcare utilization trends. Appendix Table 1 also showed that income does not predict the events that push individuals over the low spending threshold.

Comparing income groups conditional on reaching the low OOP maximum also helps

account for potential mean reversion. This concern arises because patients who experience a sudden hospitalization may recover and require fewer services after discharge. Moreover, those with prolonged hospital stays may not incur additional claims during their hospitalization. As a result, healthcare utilization could mechanically decline after reaching the low OOP maximum and we would wrongly attribute the decline to utilization management. However, if mean reversion is similar across income groups who face the same sudden hospitalization, then comparing low- and high-income individuals who reach the low OOP maximum will help balance out these differences.

4.2 Main Results

Table 4 presents the main results using as outcome variables the total number of claims and total healthcare spending. In all specifications I exclude the month in which individuals reach the low OOP maximum. Focusing on column (1), findings show that low-income consumers make substantially fewer claims after reaching the low OOP maximum compared to high-income individuals. The sensitivity of healthcare demand to zero OOP prices remains evident even when excluding the months in which patients do not make claims, as shown in column (2). In this case, the number of claims decreases 42% relative to baseline.

Low-income consumers not only file fewer claims after reaching the spending threshold but also have lower healthcare spending compared to their higher-income counterparts. Column (3) shows that after reaching the low OOP maximum, low-income individuals are 210 thousand pesos ($111) cheaper than baseline. This pattern also persists when restricting the analysis to months in which individuals make at least one claim in column (4), where low-income consumers incur 38% lower spending. The reduction in total healthcare spending is a combination of low-income consumers having lower utilization but also claiming relatively cheaper services than their counterparts as seen in Appendix Table 3 which uses the average claim price as dependent variable.

Table 5 presents estimates using as outcomes the number of primary care, specialist care, and urgent care claims conditional on making at least one claim. The notable result in column (1) is that low-income consumers reduce their utilization of primary care by 26% after reaching the low OOP maximum relative to high-income consumers. This in turn translates

TABLE 4: DID on Healthcare Utilization and Spending

Variable	Total claims		Total spending	
	Main	Intensive margin	Main	Intensive margin
	(1)	(2)	(3)	(4)
Post low OOP max	-3.027***	-0.915	-0.638***	-0.596***
	(1.115)	(0.799)	(0.175)	(0.159)
Low income × Post low OOP max	-1.409***	-2.927***	-0.210**	-0.142*
	(0.412)	(0.861)	(0.084)	(0.075)
OOP Spending relative to low OOP max	7.117***	4.314***	1.362***	1.508***
	(1.957)	(1.525)	(0.329)	(0.328)
Constant	4.973***	8.068***	0.508***	0.704***
	(0.473)	(0.345)	(0.081)	(0.072)
Fixed effects				
Individual	✓	✓	✓	✓
Month	✓	✓	✓	✓
Observations	827857	367997	827857	367997
R-squared	0.153	0.256	0.136	0.274

Note: Table presents regression results using as outcome variables the total number of claims and total healthcare spending in millions of 2011 pesos. An observation is a person-month. All specifications exclude the month in which individuals reach the low OOP maximum. Columns (1) and (3) use the analysis sample of individuals who never received a chronic disease diagnosis and who had a hospitalization. Columns (2) and (4) exclude the months in which the total number of claims equals zero. Specifications include individual and month fixed effects. Standard errors in parenthesis are clustered at the individual level.

into significant declines in the number of specialist and urgent care claims in columns (2) and (3), respectively. Since patients in Colombia need a referral from a primary care physician to access specialized care, these results suggest that insurers have stronger incentives to enforce utilization management at the primary care level to limit downstream costs after patients reach the OOP maximum.

Robustness checks. I conduct several robustness checks on my main results to provide evidence that estimates represent causal effects. Appendix Table 4 compares low- and high-income individuals that are within a bandwidth of 0.2 times the MMW around the income cutoff, in the style of a differences-in-discontinuities design.[9] This sample restriction makes the two income groups more balanced in terms of pre-low OOP maximum characteristics (as seen in Appendix Table 2) while reducing the precision of the estimates. Results in the appendix show that even in this smaller sample, low-income patients consume significantly

[9]This bandwidth guarantees that income groups are comparable in terms of healthcare utilization trends before reaching the low OOP maximum as seen in Appendix Figure 3. This bandwidth is smaller than the optimal bandwidth using Calonico et al. (2014)'s methodology with a uniform kernel. However, results are very similar when using the optimal bandwidth. In this specification I drop high-income consumers making more than 5 times the MMW who have different cost-sharing rules.

TABLE 5: DID on Healthcare Services

Variable	Primary care claims (1)	Specialist care claims (2)	Urgent care claims (3)
Post low OOP max	-0.028	-0.019	-0.289***
	(0.028)	(0.024)	(0.056)
Low income × Post low OOP max	-0.223***	-0.085***	-0.198***
	(0.033)	(0.028)	(0.076)
OOP spending relative to low OOP max	0.127**	0.121***	0.274***
	(0.050)	(0.031)	(0.070)
Constant	0.856***	0.598***	0.943***
	(0.011)	(0.007)	(0.015)
Fixed effects			
Individual	✓	✓	✓
Month	✓	✓	✓
Observations	367997	367997	367997
R-squared	0.331	0.328	0.266

Note: Table presents regression results using as outcome variables the number of primary care, specialist care, and urgent care claims. An observation is a person-month. All specifications exclude the month in which individuals reach the low OOP maximum and exclude the months in which the number of claims is zero. Results use the analysis sample of individuals who never received a chronic disease diagnosis and who had a hospitalization and exclude the month in which individuals reach the low OOP maximum. Specifications include individual and month fixed effects. Standard errors in parenthesis are clustered at the individual level.

fewer services after reaching the low spending threshold compared to high-income consumers.

Appendix Table 5 estimates the main regression model excluding high-income individuals who ever reach the high OOP maximum, which would force the comparison to be between low-income consumers who have both spot and shadow prices equal to zero and high-income consumers for whom both prices are strictly positive throughout the year. Results in the appendix show that $\hat{\alpha}$ on the total number of claims and total healthcare spending are negative and significant. The fact that the magnitude of my estimate is similar across the different specifications suggests that consumers are highly myopic and tend to respond to the spot price of care.[10]

Appendix Tables 7 and 8 estimate the main regression model exclusively on dependents and on main contributors, respectively. Since cost-sharing rules are slightly different for each group, incentives to consume healthcare after reaching the low OOP maximum may

[10]I also corroborate this claim in Appendix Table 6 where I compare initial healthcare utilization between low- and high-income consumers by the month in which patients make their first claim, following Aron-Dine et al. (2015)'s methodology. Results in the appendix show that low- and high-income consumers do not differ in their initial utilization levels across the months in which they make their first claim, suggesting shadow prices play a small role in healthcare demand.

differ between the two. Results show that reductions in utilization and spending among low-income consumers who reach the low OOP maximum persist in these two samples. Finally, Appendix Table 9 presents placebo exercises using different income cutoffs to determine the comparison group, yielding noisy estimates as expected.

5 Interpretation as Utilization Management

Why does healthcare demand slope down when OOP prices are zero? In the second part of the analysis, I explore mechanisms for this result. By comparing low- and high-income individuals who are similar in several characteristics except for their OOP maximum—especially when they are within a small bandwidth around the income cutoff—, results in the previous subsection likely rule out several answers to this question such as mean reversion and changes in health status.

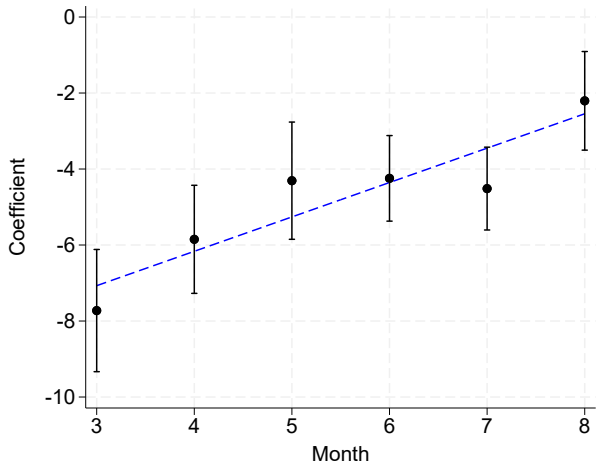
One potential answer is that low-income consumers are unaware that they have reached their OOP maximum and behave as if OOP prices are non-zero.[11] If this type of information friction disappears over time the more patients visit the doctor, then those who reach the low OOP maximum early in the year should consume more expensive services than those who reach it later in the year, as the former have had more time to interact with the healthcare system.

I explore the possibility of information frictions in explaining my results by estimating the main regression model conditional on consumers who reach the low OOP maximum in different months of the year. Figure 4 shows a pattern that is opposite to what we would expect in the presence of information frictions: low-income consumers who reach the low OOP maximum in March on average consume substantially fewer services that their higher-income counterparts compared to those who reach the spending threshold in August.

The decline in the number of claims when OOP prices drop to zero may also be driven by provider responses to payment structures. If insurers negotiate contracts that, for a

[11] The implications of information frictions for my results are similar to the implications of having measurement error in the OOP spending relative to the low OOP maximum, which would arise, for example, if there is manipulation income reports or if there are errors in assigning cost-sharing rules based on the average monthly income in a year.

FIGURE 4: Do Consumers Face Information Frictions?



Note: Figure shows coefficients and 95% confidence intervals on the interaction between low income and post-low OOP maximum in the main regression model. Each coefficient is estimated in a sample that is restricted to individuals who reach the low OOP maximum in each month (in addition to those that never reach the maximum). The dashed blue line corresponds to a linear fit across point estimates.

given service, either reimburse providers at lower rates or impose greater financial risk when treating low-income consumers who have exceeded their OOP maximum, then providers may have incentives to avoid these patients or deliver fewer services, explaining the patterns in Table 4.

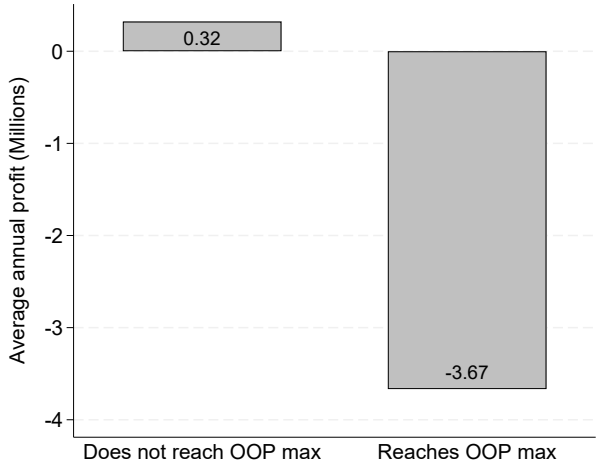
Appendix Table 10 uses the claims-level data to explore this hypothesis with two specifications. The first specification compares claim prices before and after patients reach the low OOP maximum, conditional on visiting the same provider, claiming the same service, and being enrolled with the same insurer. Here coefficients would only be identified from variation in prices across individuals, as desired. The second specification uses an indicator for whether the insurer reimburses the claim under a fee-for-service contract (relative to a capitation contract) as the outcome variable. Because fee-for-service contracts place the financial risk on the insurer, a significant reduction in the use of fee-for-service after patients reach the low OOP maximum would be consistent with providers facing greater financial risk. Results in the appendix show no evidence that provider responses to payment structures play a role in my findings.

Another plausible explanation for the observed patterns in the data is that healthcare demand is influenced by insurers' utilization management strategies. Once a patient reaches

their OOP maximum and price controls become ineffective, incentives to implement these strategies may intensify. This is particularly true for patients who reach their maximum early in the year, which helps explain the findings presented in Figure 4. In other words, the coefficient of interest can capture the differential effect on outcomes from having the insurer go from covering 88.5% to 100% of healthcare expenditures for the low-income individual versus having zero changes in expenditures for the high-income individual who still faces cost-sharing despite both experiencing a sudden health shock.

Is the 11.5 percentage point increase in insurers' costs among low-income consumers after the low OOP maximum large enough to drive the price sensitivity of healthcare demand? To better understand the magnitude of this change, Figure 5 presents the average annual insurer profit per enrollee by whether consumers reach their OOP maximum during the year in the full sample. This profit equals the government's risk-adjusted transfer minus the healthcare cost incurred by the insurer. Insurers face significant losses, of around 3.7 million pesos ($1.9K), from individuals who face zero OOP prices during the year, which stands in stark contrast to the average profit of 320 thousand pesos ($169) among those who face strictly positive OOP prices throughout the year.

FIGURE 5: Average Insurer Profit per Enrollee



Note: Figure presents the average annual profit per enrollee by whether the enrollee ever reached the OOP maximum in the year. Calculation uses the full sample and excludes patients with associated profit below the 1st percentile of the distribution. Profits per enrollee are equal to the risk-adjusted transfer from the government minus the healthcare cost incurred by the insurer. Profits are measured in 2011 pesos.

To identify utilization management as a potential mechanism for my results, the ideal

experiment would be to increase insurers' costs randomly across individuals while keeping consumers' OOP expenditures fixed. This experiment would exogenously increase incentives to engage in quantity controls while price controls are fixed. To approximate this ideal variation, I exploit a 2011 policy in which the Colombian government expanded the list of covered services in the national health plan but did not modify consumer cost-sharing.[12] During this period there were also no changes in insurance premiums which are always zero, no changes in eligibility since enrollment is always mandated for all the population, and no changes in market structure since there was no entry of health insurers in the contributory system.

Figure 6 illustrates my identification strategy. A low-income consumer faces a coinsurance rate or spot price equal to 11.5% up to the low OOP maximum when both spot and shadow prices fall to zero. Total OOP spending for the individual is represented in the blue shaded area. Insurer prices are depicted in the inverted right vertical axis. The cost to the insurer from covering this individual is represented in the orange shaded area. The 2011 policy expands the range of cumulative spending towards the right, increasing the insurer's total cost by the red shaded area, but leaving consumers' total OOP spending unchanged. The comparison of high- versus low-income consumers who face zero prices, before and after the expansion of benefits, would help identify the impact on healthcare demand from changes in insurer costs, since the cost of a high-income consumer would increase by a relatively smaller magnitude than for the low-income consumer.[13] The regression specification that implements this design is

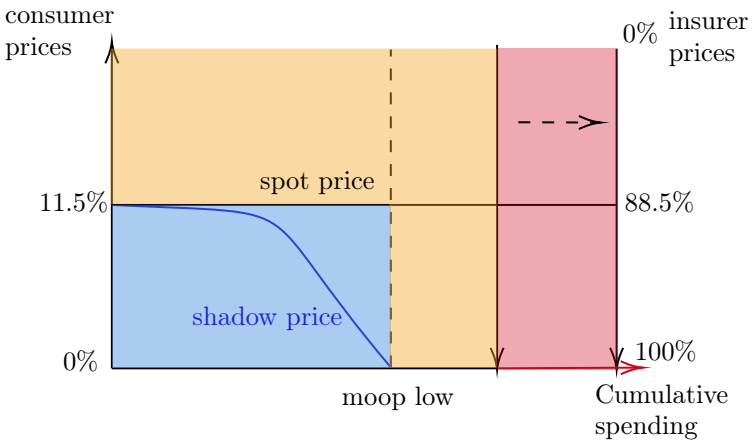
$$y_{it} = \beta LP_{it} + \alpha\ T_i \cdot LP_{it} + \theta\ T_i \cdot LP_{it} \cdot E_t + \lambda S_{it} + \delta_i + \gamma_t + \varepsilon_{it}$$

where E_t is a dummy for the post benefit expansion period and the rest of the variables are the same as in equation (1). The coefficient of interest is θ.

[12] In December 2011, the Colombian government unified the contributory and subsidized systems' insurance plans, which up until that point had different service coverage. The benefits package was expanded to cover more complex procedures such as open breast biopsy, laparoscopy ovary cystectomy, and colored doppler echocardiogram. The national prescription drug formulary was also expanded to include 63 additional drugs. See Law 1438 of 2011.

[13] This identification strategy is generally valid under the assumptions from the main regression model, that is, that the high- and low-income individuals are comparable and face the same sudden health shock.

FIGURE 6: Illustration of Identification Strategy for Insurer Gatekeeping



Note: Figure illustrates the identifying variation for the insurer gatekeeping mechanism. The figure depicts the spot and shadow prices of care for a low-income consumer measured in the left vertical axis and the insurer's marginal cost measured in the inverted right vertical axis. The blue area represents the consumer's total OOP spending. The orange area represents the insurer's total cost. And the red area represents the additional cost to the insurer from a policy that expands the list of covered benefits.

Table 6 presents the results using the total number of claims as outcome variable. All columns exclude the month in which individuals reach the low OOP maximum. Column (2) focuses on the months in which individuals make at least one claim and column (3) restricts to the months strictly after reaching the low OOP spending threshold. In columns (1) and (2) findings show that low-income consumers who reach the low OOP maximum after the expansion of benefits reduce their healthcare utilization by a greater magnitude than those who reach the spending threshold before the expansion of benefits. This result is consistent with an increase in insurers' incentives to engage in utilization management when individuals are expected to become more expensive. Column (3) shows that the expansion of benefits is related to an overall increase in the number of claims, which goes in line with findings in prior work (McNamara and Serna, 2022). However, low-income consumers file 26% fewer claims after reaching the low OOP maximum during the period of expansion of benefits relative to their counterparts.[14]

Robustness checks. Appendix Table 11 corroborates that results are due to quasi-

[14]Individual fixed effects in column (1) of Table 6 are perfectly collinear with the low-income indicator and the post benefit expansion indicator, which is why these variables are excluded from the output. In column (2) which focuses on the months after reaching the OOP maximum, the "Post low OOP max" variable is always 1 and the low income indicator is perfectly collinear with "Low income × Post low OOP max", thus these variables are also excluded from the output.

TABLE 6: Evidence of Insurer Gatekeeping on Total Number of Claims

Variable	Main	Any claim	Post low OOP max
	(1)	(2)	(3)
Post low OOP max	-3.030***	-0.911	—
	(1.116)	(0.802)	
Post expansion	—	—	1.381***
			(0.438)
Low income × Post low OOP max	-0.672	-2.081	—
	(0.775)	(1.327)	
Low income × Post low OOP max × Post expansion	-1.654	-1.879	-1.320**
	(1.136)	(1.321)	(0.625)
OOP spending relative to low OOP max	7.121***	4.306***	4.074**
	(1.958)	(1.529)	(1.618)
Constant	4.973***	8.067***	3.213***
	(0.473)	(0.346)	(0.743)
Fixed effects/Controls			
Individual	✓	✓	—
Month	✓	✓	✓
Socio-demographics	—	—	✓
Observations	827857	367997	39727
R-squared	0.153	0.256	0.058

Note: Table presents regression results using as outcome variable the total number of claims. An observation is a person-month. All specifications use the analysis sample excluding the month in which individuals reach the low OOP maximum. Columns (1) and (2) include individual and month fixed effects. Column (2) restricts to the months in which individuals make any claim. Column (3) keeps only the months after individuals reach the low OOP maximum and include month fixed effects and demographic controls (dummies for sex, low-income, age group, and being hospitalized). Standard errors in parenthesis are clustered at the individual level.

random changes in insurers' costs brought by the expansion of benefits. This table presents a placebo test assuming that the expansion happened in 2010 and finding that $\hat{\theta}$ is of the opposite sign and statistically insignificant.

6 Utilization Management Mechanism

So far, the analysis indicates that utilization management exists, with incentives for these practices largely influenced by patients' total healthcare cost. In Colombia, insurers can adopt utilization management by guiding patients to lower-cost, lower-quality providers or by offering a limited network of providers, as competition among insurers primarily revolves around their networks.

Anecdotally, there are several ways in which insurers can use the network for utilization management. One common approach requires hospitals to seek insurer approval before

admitting patients. For instance, if a patient arrives at the emergency room and requires hospitalization, the insurer may opt to transfer them to a different clinic, even covering the transportation costs. Another method involves the contractual agreements insurers establish with providers. In Cali, for example, insurers often contract with *Fundación Valle del Lili*—a prestigious teaching hospital—exclusively for patients with complex medical conditions such high-risk pregnant women. As a result, low-risk pregnant women may be directed to lower-tier hospitals for delivery, even if *Fundación Valle del Lili* is the closest facility and they would prefer to go there. In this example, utilization management is desirable as both women are likely to experience positive health outcomes postpartum but healthcare spending is lower than if the low-risk woman had delivered her baby at *Fundación Valle del Lili*.

The challenge with the exercise of showing that insurers nudge patients who reach their OOP maximum to cheaper providers is separating consumer preferences from insurer utilization management. If we see the patient visit a cheaper or lower-quality hospital, is it because they have an unobserved preference for this hospital? Or is it because the insurer directed them there?

To separate the effect of utilization management, 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:

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

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 l_{ih} is an indicator for whether patient i had previously visited hospital h—which controls for the potential bias arising from provider inertia.[15] 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, age groups, and having low income). Moreover, $\omega_i \sim N(0,1)$ captures unobserved heterogeneity in price sensitivity

[15] 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.

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 full admission prices in both states of the world because insurer utilization management is present along the entire distribution of healthcare expenditures. However, its relative importance on the decision of 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 price sensitivity and utilization management. States of the world differ on this source of price responsiveness. Before reaching the OOP maximum, consumer choices are influenced by prices and utilization management. After reaching the OOP maximum, hospital choices are influenced only by utilization management since the insurer must cover 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 consumer i's OOP spending up to but not including the hospital admission, oop_i is the OOP maximum, and ν_i is measurement error which may arise, for instance, from information frictions regarding patients or insurers being unaware of having reached the OOP maximum. I further parameterize $\nu_i \sim N(0, \sigma_\nu^2)$, which implies that the probability of each state of world is

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

Finally, I 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

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

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 as in the reduced-form evidence presented earlier. Price variation within hospital (across insurers) and coinsurance rate variation across patients are needed to identify these coefficients. However, this 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).[16] There is a separate reference price for hospital admissions by basic, intermediate, and intensive care as well as by the 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 their 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 3 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

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

when there is not enough variation in the coinsurance rates. Imposing that β_i is the same across states forces identification to come only from the state of the world after reaching the OOP maximum. The unobserved preference heterogeneity parameter σ_p is identified from observationally identical patients who have not reached their OOP maximum but choose different hospitals, and from variation in the choice set across consumers. Finally, for the probability of each state, 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 the analysis sample. I drop hospital admissions that happen during 2009 because lagged hospital choices (l_{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, following prior literature (Ho, 2006).

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.[17] This information comes from the 2018 population census of Colombia. With these 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 this census block-level information exists. These municipalities represent 75% of admissions. Appendix 2 describes the census data by reporting maps of the 4 largest municipalities in my sample with their census blocks and hospital geolocations.[18]

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

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

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

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 while others might only use age group. 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 2 describes 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 7 and first-stage regression results of admission prices on the instruments are reported in Appendix Table 12. Since the instruments are a proxy for the hospitals' marginal cost, I find that there is a strong 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 (about 1/2 of the mean) reduces the probability of choosing a hospital by 15%. After reaching the maximum, a 10,000 pesos increase in admission prices (about 3% of the mean) reduces the choice probability by 1.7%. Because OOP prices are zero after patients reach their OOP maximum, price sensitivity in this state of the world can be explained by insurer utilization management.

Interactions of prices with consumer demographics are in line with intuition and previous literature (e.g., Ho, 2006). For example, low-income consumers are more responsive to OOP prices and older individuals tend to be less price sensitive. Utilization management incentives captured by full admission prices, are stronger among older individuals who are potentially more expensive to the insurer. Price sensitivity is substantially heterogeneous across consumers as seen by the estimate of σ_p. I also 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 13%.

TABLE 7: Hospital Demand Model Estimates

		Hospital demand	
		coef	se
OOP price		-14.72	(0.997)
Price		-1.716	(0.172)
Distance		-0.135	(0.074)
Lag visit		2.726	(0.013)
σ_p		6.008	(0.488)
σ_ν		0.833	(0.015)
Interactions			
OOP price	Low income	-7.194	(1.113)
	Male	9.969	(0.532)
	Age at least 20	(ref)	
	Age 21-30	15.036	(1.621)
	Age 31-40	15.562	(2.153)
	Age 41-50	2.939	(2.886)
	Age 51-60	7.084	(1.599)
	Age 61-70	-11.166	(1.413)
	Age 71 or older	-11.983	(2.437)
Price	Low income	0.264	(0.437)
	Male	-0.616	(0.239)
	Age at least 20	(ref)	
	Age 21-30	0.142	(0.227)
	Age 31-40	-0.195	(0.225)
	Age 41-50	0.078	(0.148)
	Age 51-60	-0.117	(0.112)
	Age 61-70	-0.729	(0.210)
	Age 71 or older	-2.217	(0.112)
Observations		195,200	
Log-likelihood		-47970.5	

Note: Table shows estimates of the hospital demand model. Estimation uses the analysis sample of individuals who were never diagnosed with a chronic health condition and had a sudden hospitalization between 2010 and 2011. Estimation uses a control function approach to implement the price instrument. Full admission prices and OOP prices are measured in millions of 2011 pesos. Distance is measured in kilometers. Lag visit takes the value of one if the consumer visited the hospital in 2009. Standard errors reported in parenthesis are based on 100 bootstrap resamples.

6.1 Partial Equilibrium Analysis

Using my model estimates, I conduct a partial equilibrium analysis that reveals the relative importance of utilization management on access to inpatient care by setting $\beta_i = 0$. I recompute individuals' choice probabilities and present summary statistics of the following measures: consumer surplus, price of the chosen hospital, quality rank of the chosen hospital, and demand elasticity with respect to admission prices. Appendix 4 reports the expressions

to compute these measures.

TABLE 8: Partial Equilibrium Results

	Observed (1)	No util. management (2)
Panel A. All consumers		
Consumer surplus per enrollee	0.122 [0.003, 0.228]	0.160 [0.037, 0.295]
Price of chosen hospital	0.294 [0.200, 0.291]	0.333 [0.210, 0.307]
Quality rank of chosen hospital	37.47 [20.78, 49.85]	37.16 [21.92, 48.44]
Price elasticity	-0.615 [-0.818, -0.334]	-0.125 [-0.164, -0.074]
Panel B. Low-income consumers		
Consumer surplus per enrollee	0.096 [0.066, 0.119]	0.126 [0.087, 0.153]
Price of chosen hospital	0.290 [0.221, 0.285]	0.316 [0.247, 0.295]
Quality rank of chosen hospital	42.94 [27.32, 50.40]	42.26 [27.81, 49.22]
Price elasticity	-0.586 [-0.738, -0.318]	-0.110 [-0.150, -0.056]
Panel C. High-income consumers		
Consumer surplus per enrollee	0.109 [0.037, 0.174]	0.145 [0.071, 0.234]
Price of chosen hospital	0.292 [0.201, 0.285]	0.325 [0.216, 0.300]
Quality rank of chosen hospital	40.17 [26.28, 50.40]	39.68 [26.24, 49.22]
Price elasticity	-0.692 [-0.911, -0.353]	-0.104 [-0.169, -0.026]

Note: Table presents mean, and 25th and 75th percentiles in brackets of the distribution of outcomes in the observed scenario in column (1) and in the scenario without utilization management in column (2). Panel A computes summary statistics across all consumers, panel B among consumers with low income, and panel C among consumers with high income. Consumer surplus per enrollee and prices are measured in millions of 2011 pesos.

Table 8, Panel A presents the mean along with the 25th and 75th percentiles (in brackets) of each measure under both the observed scenario and the scenario without utilization management. The findings show that, in the absence of utilization management, consumers would opt for hospitals that are 13% more expensive than baseline, while also exhibiting a slight increase in average hospital quality rank. These results emphasize the classic trade-off of quantity controls: while utilization management serves as a crucial tool for containing healthcare spending, its implementation may reduce the frequency of patient visits to preferred hospitals or restrict access to inpatient care altogether. This trade-off is further substantiated by the finding that consumer surplus per capita would increase 30% in absence of utilization management.

Table 8, Panels B and C examine how the effects of eliminating utilization management vary by income level. Overall, removing utilization management increases consumer surplus across all income groups, with more significant gains observed among higher-income individuals. For instance, the findings indicate that high-income consumers would select

hospitals that are 11% more expensive than baseline, compared to a 9% increase for low-income individuals. Additionally, without utilization management, high-income consumers tend to choose hospitals of higher quality than those available to low-income patients. While these results do not take into account the general equilibrium effects of banning utilization management on negotiated admission prices, they effectively highlight the trade-offs and consequences associated with implementing quantity controls in a healthcare system.

7 Conclusions

This paper illustrates how insurers actively shape patients' medical care decisions through utilization management strategies, using the Colombian healthcare system as an empirical context. Initially, I estimate that healthcare demand declines with the full price of care, even after controlling for consumers' out-of-pocket expenses. This negative relationship is identified through the discontinuity in cost-sharing created by the out-of-pocket maximum, which shows that demand decreases even when out-of-pocket costs are zero. I present multiple pieces of evidence supporting the causal nature of this effect.

Building on this foundation, I further investigate why healthcare demand is influenced by the full price of care and find causal evidence consistent with utilization management. In Colombia, insurers can engage in utilization management by directing patients to lower-cost, lower-quality providers within their networks. I substantiate this mechanism using a model of hospital choice, where partial equilibrium simulations indicate that utilization management indeed encourages patients to select cheaper and less-preferred hospitals.

The finding that insurer utilization management leads to lower healthcare utilization and suboptimal patient choices is highly significant, especially when these practices have drawn considerable media scrutiny recently. To the extent that decreased healthcare utilization may negatively impact health outcomes, my findings support concerns that utilization management may compromise patient health while also demonstrating that it is an effective cost-containment strategy.

References

Alpert, A., Dykstra, S., and Jacobson, M. (2024). Hassle Costs versus Information: How Do Prescription Drug Monitoring Programs Reduce Opioid Prescribing? *American Economic Journal: Economic Policy*, 16(1):87–123.

Anderson, K. E., Darden, M., and Jain, A. (2022). Improving Prior Authorization in Medicare Advantage. *JAMA*, 328(15):1497–1498.

Aron-Dine, A., Einav, L., Finkelstein, A., and Cullen, M. (2015). Moral Hazard in Health Insurance: Do Dynamic Incentives Matter? *Review of Economics and Statistics*, 97(4):725–741.

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

Brot-Goldberg, Z. C., Burn, S., Layton, T., and Vabson, B. (2023). Rationing Medicine Through Bureaucracy: Authorization Restrictions in Medicare. 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.

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*.

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. National Bureau of Economic Research.

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.

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.

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.

Gaines, M. E., Auleta, A. D., and Berwick, D. M. (2020). Changing the Game of Prior Authorization: The Patient Perspective. *JAMA*, 323(8):705–706.

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.

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.

Ho, K. (2006). The Welfare Effects of Restricted Hospital Choice in the US Medical Care Market. *Journal of Applied Econometrics*, 21(7):1039–1079.

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.

Kyle, M. A. and Song, Z. (2023). The Consequences and Future of Prior-Authorization Reform. *The New England journal of medicine*, 389(4):291.

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.

McNamara, C. and Serna, N. (2022). The Impact of a National Formulary Expansion on Diabetics. *Health Economics*, 31(11):2311–2332.

Ofri, D. (2014). Adventures in Prior Authorization. *The New York Times*.

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

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.

Serna, N. (2024). Determinants of Provider Networks: Risk Selection vs. Fixed Costs.

Shigeoka, H. (2014). The Effect of Patient Cost Sharing on Utilization, Health, and Risk Protection. *American Economic Review*, 104(7):2152–84.

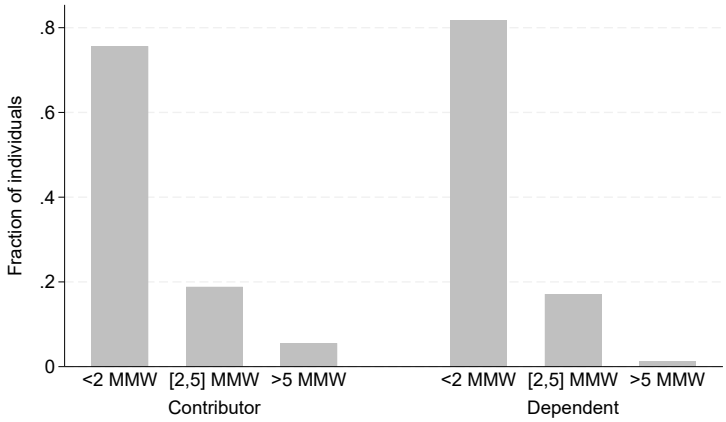
Span, P. (2024). When Prior Authorization Becomes a Medical Roadblock. *The New York Times*.

Stockton, A. (2024). What's My Life Worth? The Big Business of Denying Medical Care. *The New York Times*.

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

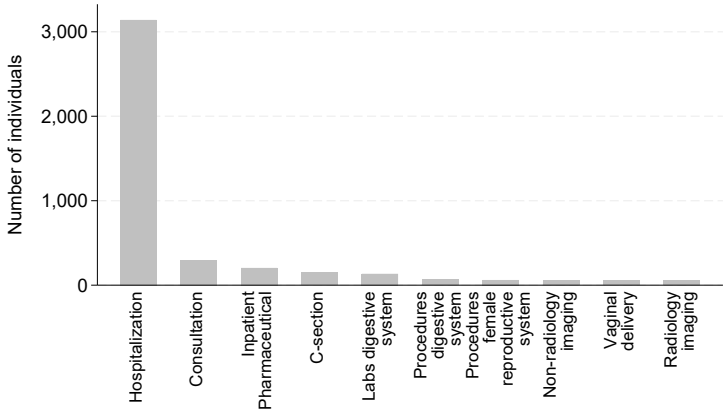
Appendix 1 Additional Descriptives and Results

Appendix Figure 1: Distribution of Individuals by Income Group



Note: Figure shows the fraction of individuals in the raw data by whether they are a contributor or a dependent and their predicted income level.

Appendix Figure 2: Most Expensive Types of Claims



Note: Figure shows the frequency of the top 10 most expensive types of services claimed by individuals in the week during which they reach their OOP maximum.

Appendix Table 1: Does Income Predict Health Shocks?

Variable	Any hospitalization
Lagged number of claims	0.0006***
	(0.0001)
Low income × Lagged number of claims	0.0001
	(0.0003)
Lagged spending	-0.0029***
	(0.0006)
Low income × Lagged spending	0.0001
	(0.0008)
Constant	0.0934***
	(0.0007)
Fixed effects	
Individual	✓
Month	✓
Observations	766271
R-squared	0.0268

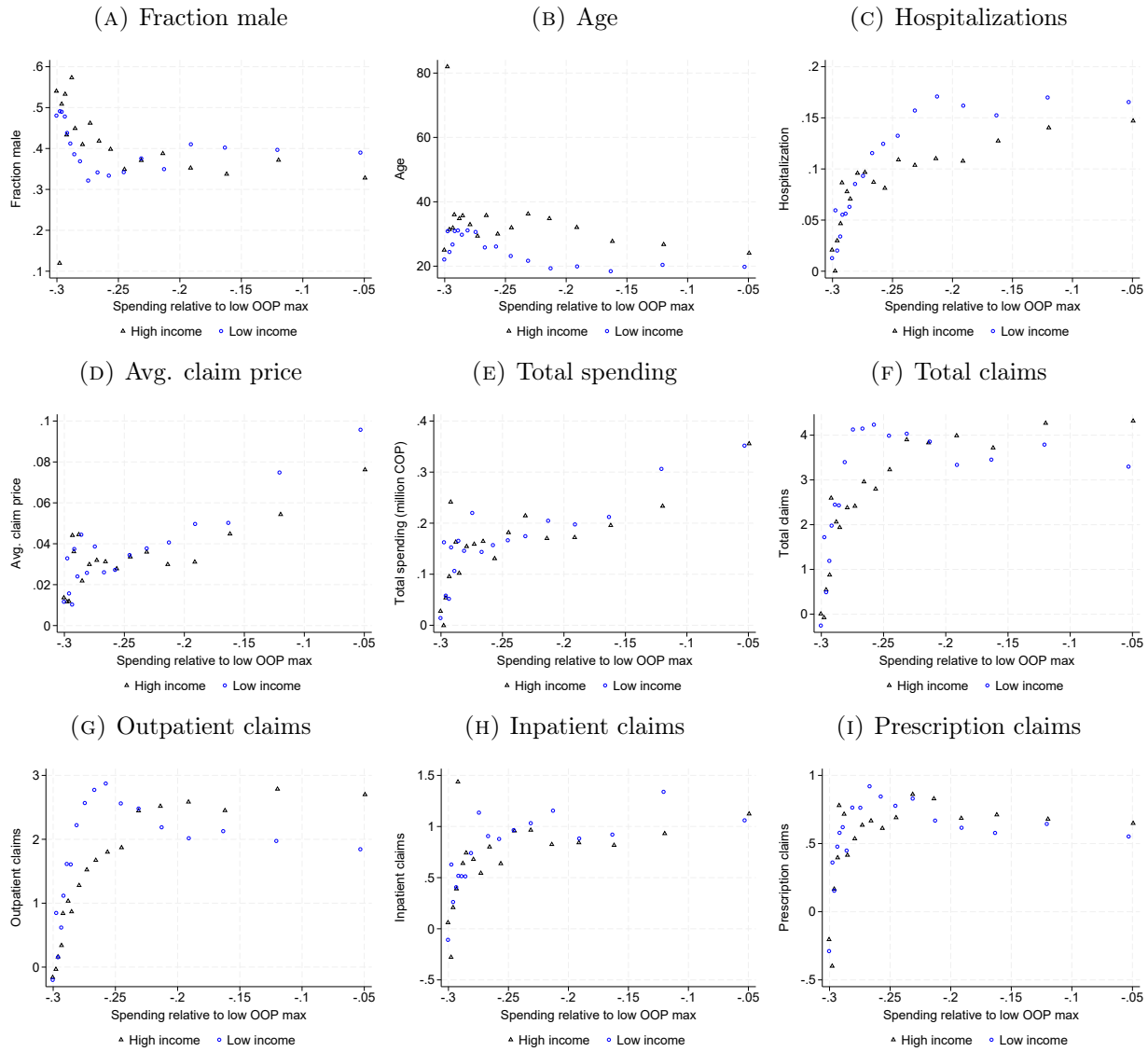
Note: Table presents regression results using as outcome variable an indicator for having any hospitalization. An observation is a person-month. Estimation uses the analysis sample of individuals who never received a chronic disease diagnosis and had a hospitalization. Specification includes individual and month fixed effects. Standard errors in parenthesis are clustered at the individual level.

Appendix Table 2: Summary Statistics By Income Group for Bandwidth Around Income Cutoff

Variable	High income	Low income
Male	0.43	0.41
	(0.50)	(0.49)
Age	30.94	25.21
	(22.04)	(21.89)
Any hospitalization	0.09	0.09
	(0.28)	(0.28)
Spending relative to low OOP max	-0.23	-0.25
	(0.08)	(0.06)
Average claim price	0.04	0.03
	(0.24)	(0.28)
Total spending	0.16	0.14
	(0.96)	(0.87)
Total number of claims	2.52	2.56
	(6.37)	(8.58)
Outpatient claims	1.49	1.55
	(3.25)	(3.98)
Inpatient claims	0.67	0.68
	(3.95)	(6.19)
Prescription claims	0.51	0.53
	(2.11)	(2.68)
Individuals × Months	18273	42512
Individuals	1692	3712

Note: Table presents mean and standard deviation in parenthesis of consumer characteristics conditional on the period before reaching the OOP maximum, on individuals who did not receive a chronic disease diagnosis before reaching their OOP maximum, and on individuals within a bandwidth of 0.2 times the MMW around the income cutoff of 2 times the MMW.

Appendix Figure 3: Parallel Trends Between Income Groups Before the Low OOP Maximum



Note: Figure presents bin-scatter plots of different variables with respect to spending relative to the low OOP maximum for high-income individuals in black and for low-income individuals in blue. The figure uses person-month data from the analysis sample restricted to the months before individuals reach the low OOP maximum. Average claim price and total healthcare spending are measured in millions of 2011 pesos.

Appendix Table 3: DID on Average Claim Price

Variable	Main (1)	Intensive margin (2)
Post low OOP max	-0.127	-0.197
	(0.099)	(0.157)
Low income × Post low OOP max	-0.019	-0.009
	(0.029)	(0.031)
OOP spending relative to low OOP max	0.254	0.424
	(0.174)	(0.312)
Constant	0.091**	0.155**
	(0.043)	(0.069)
Fixed effects		
Individual	✓	✓
Month	✓	✓
Observations	827857	367997
R-squared	0.132	0.248

Note: Table presents regression results using as outcome variable the average claim price in millions of 2011 pesos. An observation is a person-month. All specifications exclude the month in which individuals reach the low OOP maximum. Column (2) additionally excludes the months in which the total number of claims equals zero. Specifications include individual and month fixed effects. Standard errors in parenthesis are clustered at the individual level.

Appendix Table 4: DID on Measures of Healthcare Utilization Within Bandwidth Around Income Cutoff

Variable	Total claims	Total spending
Post low OOP max	-1.608***	-0.447**
	(0.501)	(0.176)
Low income × Post low OOP max	-1.813***	0.167
	(0.684)	(0.286)
OOP spending relative to low OOP max	1.250***	1.113***
	(0.301)	(0.374)
Constant	6.238***	0.558***
	(0.065)	(0.072)
Fixed effects		
Individual	✓	✓
Month	✓	✓
Observations	27328	27328
R-squared	0.299	0.255

Note: Table presents regression results using as outcome variables the total number of claims and total healthcare spending measured in millions of 2011 pesos. An observation is a person-month. All specifications exclude the month in which individuals reach the low OOP maximum and focus on individuals within a bandwidth of 0.361 times the MMW around the income cutoff of 2 times the MMW. This is the optimal bandwidth using Calonico et al. (2014)'s algorithm. All specifications include individual and month fixed effects. Standard errors in parenthesis are clustered at the individual level.

Appendix Table 5: DID Excluding High-Income People who Reach the High OOP Max

Variable	Total claims	Total spending
Post low OOP max	-2.977**	-0.465***
	(1.232)	(0.133)
Low income × Post low OOP max	-2.011***	-0.301***
	(0.691)	(0.100)
OOP spending relative to low OOP max	7.791**	1.254***
	(3.060)	(0.331)
Constant	5.144***	0.483***
	(0.750)	(0.082)
Fixed effects		
Individual	✓	✓
Month	✓	✓
Observations	824337	824337
R-squared	0.144	0.125

Note: Table presents regression results using as outcome variables the total number of claims and total healthcare spending measured in millions of 2011 pesos. An observation is a person-month. All specifications exclude the month in which individuals reach the low OOP maximum and exclude high-income consumer who ever reach the high OOP maximum. Estimation uses the analysis sample of individuals who never received a chronic disease diagnosis and who had a hospitalization, excluding high-income individuals who ever reach the OOP maximum. All specifications include individual and month fixed effects. Standard errors in parenthesis are clustered at the individual level.

Appendix Table 6: Relation Between Initial Medical Utilization and Claim Month

Variable	1-month claims	1-month spend	3-month claims	3-month spend
First month of claims	0.128***	0.022**	0.469***	0.081***
	(0.049)	(0.009)	(0.041)	(0.013)
Low income × First month of claims	-0.008	-0.001	-0.039	-0.009
	(0.049)	(0.009)	(0.042)	(0.012)
Constant	8.503***	0.654***	6.973***	0.394***
	(0.103)	(0.022)	(0.069)	(0.014)
Controls				
Demographics	✓	✓	✓	✓
Hospitalization	✓	✓	✓	✓
Observations	69645	69645	69645	69645
R-squared	0.126	0.096	0.031	0.047

Note: Table presents regression results using as outcome variable the total number of claims or total healthcare spending (measured in millions of 2011 pesos) during the first month or three months after the individual makes their first claim. Estimation uses the analysis sample and exclude months in which the total number of claims is zero. First claim month is the month in which the individual makes their first claim. An observation is an individual. All specifications control for patient sex, dummies for 15 age groups, and a dummy for hospitalization. Standard errors in parenthesis are clustered at the individual level.

Appendix Table 7: DID on Subsample of Enrollees who are Dependents

Variable	Total claims	Total spending
Post low OOP max	-3.321**	-0.763***
	(1.359)	(0.205)
Low income × Post low OOP max	-0.918*	-0.075
	(0.519)	(0.068)
OOP spending relative to low OOP max	7.047***	1.411***
	(2.056)	(0.336)
Constant	4.255***	0.422***
	(0.440)	(0.075)
Fixed effects		
Individual	✓	✓
Month	✓	✓
Observations	406217	406217
R-squared	0.140	0.238

Note: Table presents regression results using as outcome variables the total number of claims and total healthcare spending measured in millions of 2011 pesos. An observation is a person-month. All specifications exclude the month in which individuals reach the low OOP maximum and focus on individuals who are dependents. All specifications include individual and month fixed effects. Standard errors in parenthesis are clustered at the individual level.

Appendix Table 8: DID on Subsample of Enrollees who are Main Contributors

Variable	Total claims	Total spending
Post low OOP max	-3.766**	-0.130
	(1.784)	(0.089)
Low income × Post low OOP max	-0.767	-0.033
	(7.889)	(0.674)
OOP spending relative to low OOP max	10.512**	0.525**
	(4.498)	(0.214)
Constant	6.549***	0.374***
	(1.203)	(0.057)
Fixed effects		
Individual	✓	✓
Month	✓	✓
Observations	421640	421640
R-squared	0.165	0.105

Note: Table presents regression results using as outcome variables the total number of claims and total healthcare spending measured in millions of 2011 pesos. An observation is a person-month. All specifications exclude the month in which individuals reach the low OOP maximum and focus on individuals who are the main contributors. All specifications include individual and month fixed effects. Standard errors in parenthesis are clustered at the individual level.

Appendix Table 9: Placebo Tests on Income Cutoff

	2× MMW (Main) (1)	1.5×MMW (2)	4×MMW (3)
Post low OOP max	-3.027***	-5.346**	-2.682***
	(1.115)	(2.661)	(0.664)
Low income × Post low OOP max	-1.409***	1.240	-0.733
	(0.412)	(1.360)	(0.608)
OOP spending relative to low OOP max	7.117***	7.460**	9.413***
	(1.957)	(3.293)	(1.301)
Constant	4.973***	5.096***	5.436***
	(0.473)	(0.832)	(0.232)
Fixed effects			
Individual	✓	✓	✓
Month	✓	✓	✓
Observations	827857	827857	73958
R-squared	0.153	0.153	0.254

Note: Table presents regression results using as outcome variable the total number of claims. An observation is a person-month. All specifications use the analysis sample. Column (1) reports estimates from the main specification, which uses 2 times the monthly minimum wage (MMW) as cutoff. Column (2) reports estimates using 1.5 × MMW as a placebo income cutoff, limiting the sample to individuals between 0 and 2 × MMW to avoid comparisons across the actual threshold. Column (3) reports estimates using 4 × MMW as a placebo income cutoff, limiting the sample to individuals between 3 and 5 × MMW. Standard errors in parenthesis are clustered at the individual level.

Appendix Table 10: Evidence of Provider Response to Payments

Variable	Claim price (1)	FFS claim (2)
Post low OOP max	-0.034*	-0.010**
	(0.019)	(0.005)
Low income × Post low OOP max	0.007	-0.007
	(0.007)	(0.008)
OOP spending relative to low OOP max	0.057	0.002
	(0.036)	(0.003)
Constant	0.065***	0.627***
	(0.008)	(0.001)
Fixed effects/Controls		
Individual	✓	✓
Month	✓	✓
Service	✓	✓
Insurer	✓	✓
Provider	✓	✓
Observations	2835587	2835587
R-squared	0.303	0.718

Note: Table presents regression results using as outcome variables the claim price (in millions of 2011 pesos) and an indicator for whether the claim is reimbursed under a fee-for-service contract. Specifications use health claim data from individuals in the analysis sample and exclude claims filed during the month in which individuals reach the low OOP maximum. All specifications include individual, month, service, provider, and insurer fixed effects. Standard errors in parenthesis are clustered at the individual level.

Appendix Table 11: Placebo Tests on Benefit Expansion Period

Variable	Main	Main & Post OOP max	Placebo expansion	Placebo expansion & post OOP max
	(1)	(2)	(3)	(4)
Post low OOP max	0.192	—	-0.025	—
	(1.005)		(1.742)	
Post expansion	—	1.622***	—	1.291***
		(0.393)		(0.395)
Low income × Post low OOP max	-3.484***	-1.424***	-5.338***	-3.482***
	(0.699)	(0.328)	(1.299)	(0.941)
Low income × Post low OOP max × Post expansion	-1.638	-1.510**	2.640	2.243**
	(1.138)	(0.619)	(1.771)	(1.136)
OOP spending relative to low OOP max	6.721***	4.014**	7.209*	4.110
	(1.976)	(1.608)	(3.724)	(2.803)
Constant	4.854***	4.493***	4.578***	4.009***
	(0.476)	(0.634)	(0.887)	(1.069)
Fixed effects/Controls				
Individual	✓	—	✓	—
Month	✓	✓	✓	✓
Sociodemographics	—	✓	—	✓
Observations	831441	43308	527836	26616
R-squared	0.151	0.070	0.137	0.042

Note: Table presents regression results using as outcome variable the total number of claims. An observation is a person-month. All specifications use the analysis sample excluding the month in which individuals reach their OOP maximum. Columns (1) and (3) include individual and month fixed effects. Columns (2) and (4) keep only the months after individuals reach their OOP maximum and include month fixed effects and demographic controls (a dummy for sex, age group, and being hospitalized). Columns (3) and (4) exclude 2011 and assume the post benefit expansion period to be 2010 as a placebo exercise. Standard errors in parenthesis are clustered at the individual level.

Appendix 2 Census Tract Data and Admission Prices

While the claims data report the admission prices that each insurer negotiated with each hospital in its network, pricing units may vary between insurer-hospital pairs. For example, one pair may negotiate one admission price for women aged less than 20 and another price for women aged more than 45. To express negotiated admission prices in a single unit, I estimate the following regression separately for every insurer using the claims data:

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

where c is a claim, j is an insurer, h is a hospital, 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.

Then, to construct my population-weighted distance measure I use data from the 2018 Colombian census. These data report population density in each locality ("manzanas" for their Spanish name) within a municipality by age quintile. I limit my analysis sample to the 14 main capital cities in the country. Appendix Figure 4 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 3 Estimation Details

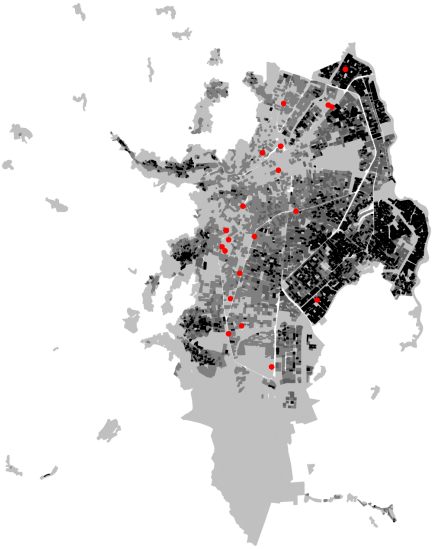
Control function. I estimate the model in Section 6 using Simulated Maximum Likelihood. I implement the instrumental variable approach using a control function (Petrin and Train, 2010). In the first stage I estimate a regression of the negotiated price between insurer j and hospital h on the price instrument (z_{jh}) and the hospital'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:

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

Appendix Figure 4: Hospital locations and census tracts

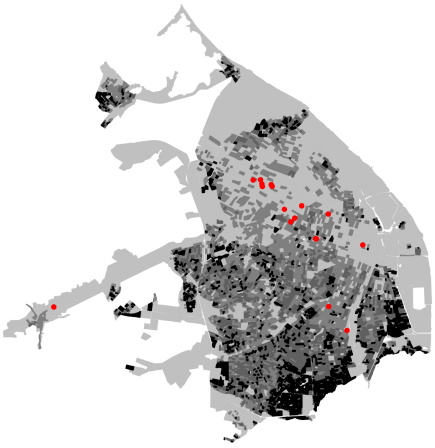
(A) Bogotá

(B) Cali



(C) Medellín

(D) Barranquilla



c

The residuals from this regression, $\tilde{p}_{jh} = p_{jh} - \hat{p}_{jh}$, are then added to the utility function in equation (2) as $\psi_i \tilde{p}_{jh}$, allowing for the same interactions with consumer demographics (a constant and dummies for sex, age groups, and having low income).

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

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

$\frac{\partial}{\partial\tilde{\theta}}\delta^s_{ijh}$ 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 transform them in estimation using a exponential function and modify the gradient accordingly.

I select the starting values by estimating the following auxiliary model using the control function approach:

$$u_{ijh} = \alpha_i r_i p_{jh} + \beta_i p_{jh} + \tau d_{ih} + \kappa l_{ih} + \xi_h + \varepsilon_{ijh}$$

I choose starting values for σ_ν and σ_p by estimating a simplified version of the main model. Specifically, I set all provider fixed effects $\xi_h = 0$ and assume no heterogeneity in price sensitivity $\alpha_i = \alpha, \beta_i = \beta$. All starting values for this estimation are set to zero. This estimation yields an estimate of $\log\sigma_\nu = -2.3, \log\sigma_p = 1.7$, which I use as the starting value for the main estimation.

Appendix 4 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

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

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 of the chosen alternative is given by $\sum_{h \in H_j} p_{jh} P_{ijh}$. 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, I report the average demand elasticity with respect to the admission price weighted by the choice probability which is given by:

$$\frac{\sum_i P_{ijh}\Big[\Big(\gamma_i \int (\alpha_i + \sigma_p \omega_i) r_i (1 - P^1_{ijh}) P^1_{ijh}\ d\phi(\omega) + (1-\gamma_i)\beta_i (1 - P^2_{ijh}) P^2_{ijh}\Big)(p_{jh}/P_{ijh})\Big]}{\sum_i P_{ijh}}$$

Appendix Table 12: First-Stage Regression of Admission Prices

Variable	Admission prices
Instrument	7.848^{***}
	(0.088)
Beds	0.012^{***}
	(0.0005)
Constant	-0.214^{***}
	(0.006)
Observations	195,200
R-squared	0.041
F Statistic	$4{,}177.6^{***}$

Note: Table reports OLS regression of admission prices on the instrument and the number of hospital beds. The instrument for an insurer-hospital pair jh is the average reference price in 2009 across all other hospitals $-h$ in j's network, weighted by their number of beds. Robust standard errors in parenthesis.