Cost Sharing Design in Public Health Insurance: Effects on Prices and Consumers

Natalia Serna*

Abstract

In this paper, I analyze the impact of cost sharing on consumer welfare and negotiated prices between insurers and hospitals using a model of hospital choice and Nash bargaining with data from the Colombian healthcare system. I leverage the non-linearity in cost sharing due to maximum out-of-pocket amounts to identify insurer steering from consumer sensitivity to prices. I find that consumers account for 37% of the effective demand elasticity, insurers for 61%, and insurer-hospital bargaining for the remaining 2%. Equilibrium prices are U-shaped with respect to the coinsurance rate and decreasing in the out-of-pocket limits. Consumer surplus is maximized at a coinsurance rate of 30%. But, these welfare effects are heterogeneous across consumers of different income groups.

Keywords: Cost sharing, Maximum out-of-pocket expenditures, Negotiated prices, Health insurance.

codes: I11, I13, I18, L13.

*Corresponding author. University of Wisconsin - Madison. 1180 Observatory Drive, office 7316. e-mail: nserna@wisc.edu. I am deeply grateful to the Colombian Ministry of Health for providing the data for this paper. I also want to thank Alan Sorensen, Ken Hendricks, Corina Mommaerts, Naoki Aizawa, participants to the 2020 Summer Research Fellowship presentations at UW Madison, and participants to the 2021 ASHEcon Conference, for all their useful comments and feedback. All errors and omissions are my own. The findings of this paper do not reflect the views of any of the institutions involved.

1 Introduction

Cost sharing in health insurance plans is an important tool to control the consumption of healthcare by making patients face a proportion of their costs. There is widespread evidence of how cost sharing affects demand for different types of health services (Chandra et al., 2010; Agarwal et al., 2018; Shigeoka, 2014; Serna, 2021), but less is known about its effect on supply, in particular on negotiated hospital prices. With governments increasingly relying on cost sharing as a cost containment mechanism in public health insurance systems, understanding hospital price responses can help regulators better design the cost sharing policies across different groups of enrollees or services. For example, if lowering coinsurance rates leads to non proportional increases in prices, then popular coinsurance designs such as value-based insurance can increase costs and decrease patient welfare. Accordingly, the purpose of this paper is to investigate how counterfactual cost sharing schedules affect social welfare and negotiated prices between insurers and hospitals through three possible channels: consumer sensitivity to prices, insurer steering, and insurer-hospital bargaining.

I study the effect of cost sharing on prices for a particular service, namely, hospital admissions. As of 2006, inpatient services represented 35% of total healthcare costs in OECD countries (Anderson et al., 2007). Being one of the most expensive and rapidly growing healthcare categories together with prescription drugs, hospital admissions have been subject to substantial cost sharing amounts. A recent Kaiser Family Foundation report shows, for example, that Medicare Advantage plans in the United States tailor the level and form of cost sharing to the length of hospital stay (Freed et al., 2020). But, there is little evidence on how hospital prices respond to these intricate cost sharing rules or if they respond at all when hospital market power is constrained by bargaining with insurance companies. I contribute to this literature by modelling demand for hospitals as a function of non-linear cost sharing rules and by modelling the bargaining game between insurers and hospitals, in the context of the Colombian healthcare system.

The main contribution of my paper is leveraging the discontinuity in coinsurance rates due to maximum out-of-pocket (OOP) amounts, to identify consumer- from insurer-induced demand elasticity. Insurer elasticity is the result of an underlying steering mechanism in which carriers can limit the set of hospitals certain patients have access to. Enrollee satisfaction surveys conducted by the Colombian Ministry of Health show that some insurers deny services for over 10% of their enrollees, making rationing of care an important steering mechanism in the country. By decomposing the elasticity into its three portions, I am able to show that, even in absence of bargaining and

under a tightly regulated market, insurers can constrain hospital prices. My paper also contributes to the literature of optimal design of health insurance plans by looking at the welfare effects of counterfactual cost sharing policies like uniform cost sharing. Tackling this question has been difficult in other countries because insurers compete along several dimensions of plan design, but the Colombian healthcare system characterized by strict government regulation eliminates that type of insurer competition.

In Colombia there is one national insurance plan provided by several private insurers or carriers. The government regulates copays, coinsurance rates, and maximum OOP amounts by making them functions of the enrollee's monthly income level and constant across insurers and hospitals. The government also sets premiums to zero to (allegedly) incentivize insurer competition on quality. Similar to the United States, insurers have discretion over negotiated prices and hospital networks, to the extent that they bargain prices with hospitals and agreement indicates inclusion to the network. This unique empirical setting helps alleviate two outstanding challenges to identification. The first challenge is the reverse causal effect of negotiated prices on cost sharing. If insurers are able to lower their coinsurance rates for hospitals with which they have agreed on low prices then my demand elasticity estimates would be biased downwards. The second challenge is the confounding effect of premiums on cost sharing. If insurers can respond to high hospital prices by increasing premiums, then the lower enrollment levels would bias the demand elasticity with respect to cost sharing upwards.

The strict regulation of cost sharing rules in Colombia means that to steer patients towards preferred providers and minimize costs, health insurers resort to mechanisms like rationing of care or provision of narrow hospital networks. Several papers have studied insurers' use of replacement threats to achieve lower prices and narrow networks during bilateral negotiations (Ho and Lee, 2019; Ghili, 2018; Liebman, 2018). In this paper, I focus on how limited access to hospitals for patients that reach their OOP limit can impact negotiated hospital prices. Intuitively, for this set of patients, any evidence of demand elasticity is going to be fully explained by the insurer, which allows me to separate consumer sensitivity to price from insurer steering. Even though it is forbidden by law, there is evidence from the Ministry of Health that insurers deny provision of certain services or access to certain hospitals (Velez, 2016), making Colombia one of the most litigious countries in Latin America regarding health insurance coverage (Lamprea and Garcia, 2016). In using the non-linearity in health insurance contracts as an identification strategy, I show that patients are not forward-looking nor consume healthcare strategically to reach the spending thresholds (in the lines

of Aaron-Dine et al. (2015)). So, healthcare utilization in Colombia is not necessarily dynamic

To answer the question of how cost sharing affects prices I use a structural approach because coinsurance rates have not changed since the establishment of Colombia's universal healthcare system in 1993 and there are intricate endogenous responses of insurers, hospitals, and consumers. I proceed by modelling the Colombian system as a two-stage game. First, insurers and hospitals engage in bilateral negotiations over prices for hospital admissions. Second, patients receive a health shock and choose a hospital within the network of their insurance company to receive treatment. I model the first stage of the game with a Nash-in-Nash bargaining framework under the usual assumptions of fixed enrollee pools and fixed networks, and estimate the parameters of the joint surplus function, namely the bargaining parameters, using GMM. Identification of the bargaining parameters relies on price variation across insurers, which generates variation in the split of the surplus.

In the second stage, patients make discrete choices over the set of hospitals made available by their insurer. Patient moral hazard is captured by the disutility of out-of-pocket coinsurance payments made before reaching the OOP limit. As is standard in the literature that uses a discrete choice framework to model demand for healthcare, moral hazard in my setting only affects the consumption of hospital admissions, but not of other healthcare services. I assume hospital demand follows a conditional logit specification, which I estimate using maximum likelihood. Consumer sensitivity to prices in this stage of game is identified from variation in hospital choice sets across patients in the same cost sharing tier that haven't reached their OOP thresholds. This parameter could be biased towards the null if individuals fully observe negotiated prices when making carrier choices and as a result selectively enroll insurers that have low prices for their preferred hospitals. Even though prices are not observable to the consumer when making carrier choices in Colombia, to address any potential endogeneity of this type I follow Prager (2020)'s strategy. I leverage strong carrier inertia to argue that, conditional on past choices, current choices of insurer and assignment of consumers to cost sharing tiers are as-if-random.

Insurer steering is identified from variation in choice sets across patients in the same cost sharing tier who are admitted to the hospital after reaching the OOP limit. This parameter could be biased towards the null if consumers experience strong hospital inertia or differ significantly with respect to patients that fail to reach the spending thresholds in terms of their ex-ante risk. To address these threats to identification, I control for previous choices of hospital in the consumer's utility function and allow the insurer steering parameter to vary with patient demographics and diagnoses

that fully capture observable differences in ex-ante risk.

The data for this paper is a panel of around 62 thousand individuals enrolled to the national insurance plan, who were admitted to the hospital between 2010 and 2011. I find that the (quantity-weighted) average effective price-elasticity of demand equals -0.2. 61% of this elasticity is explained by insurer steering, 37% by consumer sensitivity to prices, and 2% by insurer-hospital bargaining. My elasticity estimate is smaller (in absolute value) than other estimates in the literature (Gowrisankaran et al., 2015), which suggests that the strong regulation of premiums and plan characteristics has important effects on consumer and insurer price sensitivity. Deregulation could increase demand elasticity at the expense of exacerbating adverse selection. If carriers are allowed to compete in cost-sharing and premiums, a costly individual with private information on her health status will choose an insurance plan with relatively low premiums or low cost-sharing.

I use my model of demand and bargaining to conduct three counterfactual scenarios that better help understand the effect of each element of cost sharing on negotiated prices and welfare. First, I impose uniform coinsurance rates ranging from 0 to 100%, while holding the OOP limits fixed. Second, I impose uniform OOP limits, while holding the coinsurance rate schedule fixed. And third, I combine a uniform coinsurance rate with a uniform OOP limit. Results show that prices are U-shaped with respect to the coinsurance rate and decreasing in the OOP limits. Consumer surplus is maximized at a coinsurance rate of 30%, where negotiated prices are minimized. The optimal coinsurance rate is above the observed average rate because welfare gains of individuals that reach their OOP limit overcompensate the losses of those who fail to reach the expenditure thresholds and face a higher proportion of their healthcare costs. The distribution of these welfare effects also vary across income levels. I find that most of the welfare gains at a coinsurance rate of 30% are accrued by low income individuals.

In particular, results in the first counterfactual show that negotiated prices for hospital admissions would decrease 34% on average when the coinsurance rate increases from 0 to 30%. As the coinsurance rate increases beyond this point, prices increase around 20% because patients are more likely to reach their spending thresholds after which the insurer, that is the less elastic side of the market, has to cover the full cost of healthcare. Under the second counterfactual, my findings show that average prices are decreasing with the respect to the OOP limit. At low values of the spending threshold equal to 0.5 times the monthly minimum wage, an additional 20% of individuals reach their spending thresholds relative to the observed scenario, which reduces the elasticity of demand, and increases prices by 19%. At high values of the OOP limit equal to 1.5 times the monthly

minimum wage, less than 20% of patients reach the threshold, the elasticity of demand increases, and prices fall 5% relative to the observed average price.

This paper is related to a long strand of literature that investigates the effects of cost sharing on the consumption of health services, welfare, and prices (Kleinke, 2004; Busch et al., 2006; Hsu et al., 2006; Trivedi et al., 2008; Chandra et al., 2010; Thomson et al., 2013; Robinson and Brown, 2013; Choudhry et al., 2010; Brot-Goldberg et al., 2017; Serna, 2021). It is also related to the extensive work that uses bargaining models to explain the role of out-of-pocket prices on welfare and demand for certain health services, like imaging (Brown and Robinson, 2015), stents (Grennan, 2013), and hospital admissions (Gowrisankaran et al., 2015). Several studies have found that decreasing coinsurance rates results in lower hospital and drug prices because insurers can steer patients using narrow networks or drug formularies (Brown and Robinson, 2015; Duggan and Morton, 2010; Starc and Town, 2018; Lavetti and Simon, 2016). Other authors show that the insurers' bargaining leverage plays an important role in constraining hospital prices when coinsurance rates are set to zero (Gowrisankaran et al., 2015). I add to this literature by incorporating both the insurers' steering mechanism and their bargaining leverage in a model of hospital choice and Nash bargaining. Moreover, I contribute to the literature by quantifying not only the effects of coinsurance rates but of OOP limits on negotiated prices. Other work has focused on the effect of deductibles on welfare and healthcare consumption (Diaz-Campo, 2021).

In relation to the literature on optimal design of health insurance, I provide evidence of which forms and levels of cost sharing can be welfare maximizing. This is relevant not only for Colombia, but for countries where the transition to a unique non-means-tested health insurance plan is in the policy agenda (Baicker et al., 2013). My first counterfactual analysis shows that setting the coinsurance rate to a uniform 30% maximizes consumer welfare relative to a full insurance scenario, while holding OOP maximums fixed. Hospital profits fall 30% relative to full insurance because prices are minimized at 30% coinsurance, while insurer profits increase 10%. This finding is similar to that in Ho and Lee (2021), who use data from a large employer in the United States and allow healthcare consumption to vary with cost sharing. My results are not driven by individuals internalizing the dynamic incentives introduced by the spending thresholds as most papers in this literature that model dynamic moral hazard, but by insurance companies engaging in steering practices as patients reach their OOP maximums, in a setting where all other dimensions of plan design are regulated.

2 Background

The Colombian healthcare system is divided into two regimes called "Contributory" and "Subsidized". The contributory system is funded by the required contributions of its users and covers all formal employees, independent workers, and their beneficiaries, roughly 51% of the population. The subsidized system is fully funded by the government and covers all individuals who do not receive a monthly income and who are poor enough to qualify, roughly the remaining 49% of the population. The Colombian healthcare system is characterized by having almost universal health insurance coverage, with important variation in the number of uninsured across departments due to limited geographical access in peripheral areas.

Individuals in both systems can choose from a set of private national health insurers that are in charge of providing the national benefits plan. The national plan covers a list of more than 7,000 procedures, services, and devices, and 736 prescription medications as of 2012 (Decree 029 of 2011). The government sets premiums for the national plan to zero, but compensates private insurers with per-capita transfers risk-adjusted for sex, age, and municipality of residence. This ex-ante risk adjustment formula does not control for a patient's previous diagnoses, so it poorly compensates insurers for realized healthcare costs, and incentivizes them to engage in risk selection using the only elements of the national insurance plan over which they have discretion, namely, hospital networks and service prices. The government also compensates insurers at the end of every year for the following non-exhaustive list of diseases: cervical cancer, breast cancer, stomach cancer, colon cancer, prostate cancer, lymphoid leukemia, myeloid leukemia, hodgkin lymphoma, non-hodgkin lymphoma, epilepsy, rheumatoid arthritis, and HIV/AIDS. This compensating mechanism, known as the High Cost Account, takes money from insurers that were overcompensated by the ex-ante payments, and transfers those funds to insurers with a relatively high proportion of enrollees with chronic conditions in the list.

To provide services in the national benefits package, insurers form a network of providers by engaging in bilateral negotiations over prices and contracts. Patients are only covered by their insurer when they visit hospitals in the network, while out-of-network claims are not reimbursed. If the enrollee requires services not covered in the national plan, the insurer can decide whether to authorize provision and up to what percentage. Although denying services included in the national plan can lead to sanctions by the National Health Superintendency, several insurance companies do so, particularly for high-priced procedures or medications, by taking advantage of information

asymmetries relative to the patient. These asymmetries arise from insurer-hospital negotiated prices not being fully observable to the consumer when making carrier choices and from consumers being inattentive to the set of services covered in the national plan. Even though premiums are zero and cost sharing is standardized across insurers, preference heterogeneity for network breadth and cost heterogeneity across insurers, generates an asymmetric equilibrium in networks, where consumers enroll insurers on the basis of the proportion of hospitals in a market that the insurer includes in its network (Serna, 2022). In addition to the national benefits package, private carriers can offer supplementary health insurance plans where they have discretion over premiums and cost sharing rules, however, individuals can only purchase these plans after being enrolled through the national plan.

Insurer steering through mechanisms like rationing of care is common in Colombia. McNamara and Serna (2021), for example, document the extent of rationing of care in the subsample of enrollees with type I diabetes after a regulatory change that made these individuals relatively more expensive to the insurer. Appendix (A) provides additional evidence of rationing of care using enrollee satisfaction survey data from 2013 to 2016, conducted by the Ministry of Health. The survey asks individuals enrolled in the contributory healthcare system whether they have been denied services by their insurance company over the last year and the reasons why. Appendix figure (A1) shows that service denials vary across insurers and some companies deny services for over 10% of their enrollees. Appendix figure (A2) shows that, conditional on being a denied a service, the provider being out-of-network and the service not being included in the national plan are the main reasons for claim denials. For 7 out of the 13 insurers depicted in the figure, over 10% of enrollees indicated out-of-network providers as the main reason for service denials, which highlights the importance of narrow networks for patient welfare. In the context of hospital admissions, insurer rationing of care can be understood as insurers requiring stricter prior authorization at more expensive hospitals.

From 2009 to 2011, Colombia's contributory healthcare system had 23 private insurers, 14 of which covered over 97% of enrollees. My sample for estimation consists of a subset of 13 of these insurers, described in more detail in the next section. Table (1) presents the national market share for each insurer as of December 2011. The Colombian insurance market is highly concentrated, with the top three insurers accounting for 48% of enrollees. This heterogeneity in market shares is due mostly to differences in hospital networks, given that all other plan characteristics are regulated. The government regulates copays, coinsurance rates, and maximum OOP amounts, which are a function of the enrollee's monthly income level as seen in table (2) for 2010. These prices are indexed to

the monthly minimum wage (MMW), are uniform across providers and services, and have remained fixed since the establishment of Colombia's healthcare system in 1993. Unlike the Unites States, there are no deductibles, so copays and coinsurance rates apply at all moments before reaching the OOP maximum. After reaching the expenditure limit, insurers have to offer full coverage.

Individuals with monthly incomes less than 2 times the MMW, have a copayment of 1 USD, a coinsurance rate of 11.5% of the price per claim, and an OOP maximum per year equal to 57.5% of the MMW. Those with incomes between 2 and 5 times the MMW have a copayment equal to 4 USD, a coinsurance rate of 17.3%, and an OOP maximum per year equal to 230% of the MMW. Finally, people earning more than 5 times the MMW, have copays of roughly 11 USD, coinsurance rates of 23% per claim, and a maximum OOP expenditure per year of 460% of the MMW.[1] At the end of every year, all insurers in the country report reimbursed health claims to the Ministry of Health. The data for this paper comes from the reports made by all insurers to the regulator.

Table 1: National market shares in December 2011

Insurer	Market share
EPS001	1.74
EPS002	8.91
EPS003	4.02
EPS005	4.89
EPS008	4.04
EPS009	1.87
EPS010	7.44
EPS012	1.59
EPS013	20.93
EPS016	15.09
EPS017	7.22
EPS018	4.15
EPS023	3.15
EPS037	11.95

Table 2: Copay, coinsurance rate, and out-of-pocket maximum in the contributory system in 2010

Income level	Copay (COP)	Coinsurance rate	Out-of-Pocket maximum	
		Per claim	Per claim	Per year
$y < 2 \times MMW$	2,100	11.5%	28.7%	57.5%
$y \in [2, 5] \times MMW$	8,000	17.3%	115%	230%
$y > 5 \times MMW$	20,900	23.0%	230%	460%

Note: The MMW in 2010 equals 515,000 COP or roughly 271 USD. The coinsurance rates are percentages of claims cost, whereas the maximum OOP expenditures are percentages of the MMW.

[1]The average exchange rate during 2010 is 1,898 COP/USD and the monthly minimum wage is roughly 271 USD.

3 Data and descriptive evidence

I use data from the contributory healthcare system in Colombia to estimate hospital demand and price bargaining. My data is originally a panel of 487,358 enrollees and all their general acute care hospital admissions through the national plan from 2009 to 2011, a total of 850,886 admissions. This dataset was built by the Ministry of Health and contains individuals who did not switch their insurance company during the three years and who made at least one claim. This latter constraint implies that the sample of enrollees is in worse health condition compared to the population. So, to properly adjust hospital choice probabilities later in my model in section (4), I use aggregate enrollment and demographic data from the Ministry of Health to calculate the probability that a consumer is admitted to the hospital.

From the admissions data, I select the sample of patients who are aged 19 or older with non-missing income data ($N = 483,916$), and focus on hospitals that provide at least 30 admissions per year ($N = 142,530$). By constraining the hospital sample in this way, I end up dropping small clinics or health centers that are less likely to be chosen by a patient. This matters in my case because I recover hospital networks from observed claims. My sample of hospitals accounts for 33% of all admissions and for 20% of total annual healthcare costs. I also conduct robustness checks on my measure of networks later in the document.

Even though the sample is originally constrained to individuals who did not switch their insurance carrier over the three years, I drop admissions in 2009 because I can not guarantee that those claims are not associated to the consumer's first choice of carrier. A consumer who is enrolling for the first time might selectively enroll carriers that have negotiated lower prices for preferred hospitals, which would bias my elasticity estimates with respect to cost sharing towards the null. The switching patterns in my sample are consistent with population-wide switching. Using data from 2010 and 2011, Serna (2022) documents that switching rates across the population of enrollees to Colombia's contributory system is only 0.2%. After dropping 2009, my final dataset has 62,040 patients, 76,229 admissions, 13 insurers, and 178 hospitals.

For every patient, I observe basic demographic characteristics like sex, age, and municipality of residence. Unfortunately, I do not observe the patient's address to measure distance to each hospital. This is not an issue for the analysis because for my sample of patients who are relatively sicker than the population, distance to the hospital is a relatively less important predictor of hospital choice. For every claim, I observe date of provision, service or procedure (identified by a procedures code

known as CUPS by its Spanish acronym), service price, associated ICD-10 diagnosis code, provider identifier, and insurer. I categorize the ICD-10 diagnosis codes following Alfonso et al. (2013), resulting in the following long-term disease categories: genetic anomalies, asthma, arthritis, arthrosis, autoimmune disease, cancer, diabetes, cardiovascular disease, long-term pulmonary disease, renal disease, HIV-AIDS, tuberculosis, epilepsy, and transplant. I denote a department-year combination as a market. There are a total of 21 markets in my data.

Following the related literature (Gowrisankaran et al., 2015; Ho, 2006), I assume insurers and hospitals bargain over the "base price" of a hospital admission instead of a specific price per patient. Admission prices can vary across patients based on their hospital length-of-stay, but this variation is unobserved by insurance companies at the time of negotiations with hospitals. Insurers in Colombia reimburse hospitals with a per-day rate that can vary with the patient's list of diseases. So, to calculate the price of an admission, I add the prices across all claims associated to the admission and divide by the patient's length-of-stay to get a measure of price per hospital-day. Then, I calculate the "base price" as the average price per admission for each insurer-hospital pair in a market. For convenience, in the rest of the paper I will use the term "price" to refer to the price obtained from this methodology.

To determine whether a patient has reached the OOP threshold at the time of a hospital admission, I calculate total healthcare costs by adding prices across all claims (all inpatient, outpatient, and prescription claims) made right before each admission, separately for every year. This means that even if the patient reaches the spending limit *during* her hospital stay, the total OOP cost incurred right before the hospital admission is the one that matters for hospital choice.

Table (3) provides some summary statistics of the resulting data. The unit of observation is a hospital admission. The table shows that 45.6% of admissions are associated to males, the average age is 57.9 years and its standard deviation is 19.3 years. 75.5% of admissions are associated to patients living in metropolitan municipalities, and 82.4% to patients having incomes below 2 times the MMW. The average price of a hospital admission in the full sample equals $185.4 with a standard deviation of $172.4. Admissions in the full sample have average coinsurance payments equal to $14.6 and average copays equal to $2.1. Cardiovascular diseases, cancer, and two or more comorbidities are the most prevalent conditions in the data, followed by diabetes and renal disease. Hospitals in my data have an average number of beds equal to 190.8, an average number of rooms equal to 8.8, 22.1% of them own ambulances, and a little over half are private hospitals. 59.6% of admissions correspond to patients who have not reached their OOP limit and 40.4% to their counterparts.

Table 3: Summary statistics of final patient sample

	Full sample
Price (USD)	185.4 (172.4)
Coinsurance (USD)	14.6 (22.3)
Copay (USD)	2.1 (2.3)
Demographics	
Male (%)	45.6 (49.8)
Age	57.9 (19.3)
Location (%)	
Metropolitan	75.5 (43)
Adjacent	23.6 (42.4)
Peripheral	1.0 (9.7)
Income group (%)	
$<$ 2 x MMW	82.4 (38.1)
$[2, 5]$ x MMW	14.4 (35.1)
$>$ 5 x MMW	3.2 (17.7)
Diagnoses (%)	
Cancer	6.3 (24.2)
Cardiovascular	12.9 (33.5)
Diabetes	0.9 (9.6)
Renal	0.7 (8.1)
Other	5.2 (22.3)
>=2 diseases	53.7 (49.9)
Healthy	20.2 (40.1)
Hospital characteristics	
Beds	190.8 (137)
Rooms	8.8 (6.4)
Any ambulance (%)	22.1 (41.5)
Private (%)	51.3 (50.0)
All admissions	76,229
Admissions before OOP	45,432
Admissions after OOP	30,797
Patients	62,040
Hospitals	178
Insurers	13

Note: Mean and standard deviation (in parenthesis) of the main variables in the full sample. A unit of observation is a hospital admission.

Within the set of admissions by patients in the first, second, and third income categories, 45.7%, 16.8%, and 12.5% have reached their expenditure limit, respectively. These admissions will identify insurer steering in my demand model. A natural concern with this identification strategy is that individuals, particularly in the first income bracket, can behave strategically to reach their OOP maximum. To see if this is the case, figure (1) reports the proportion of hospital admissions by level of out-of-pocket spending relative to the income-specific OOP limits in the lines of Einav et al. (2016). For the group of patients with incomes below the $2 \times MMW$ threshold, there is no discontinuity in the likelihood of being admitted to the hospital at a relative expenditure of zero. The jump in the probability of admission happens \$100 before reaching the OOP limit, which suggests that patients in this income category are not forward-looking. For individuals in the second category (earning between 2 and 5 times the MMW), the likelihood of admission is maximal around

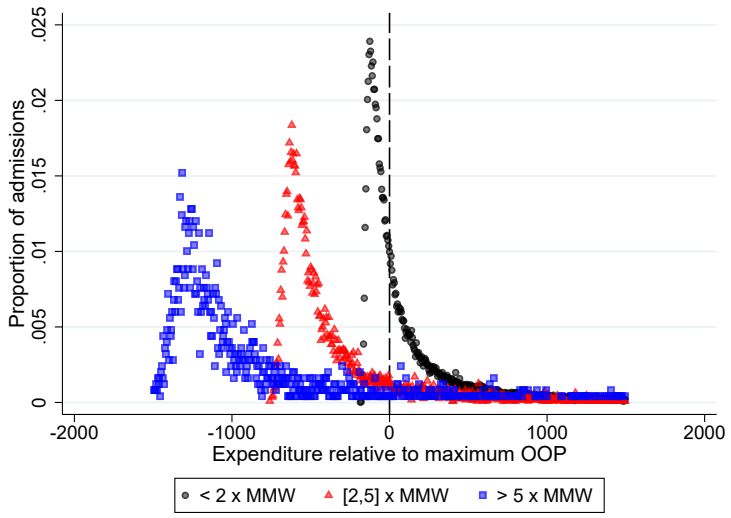


Figure 1: Proportion of admissions by level of relative out-of-pocket spending

Note: This figure presents the proportion of admissions by level of out-of-pocket spending relative to the maximum expenditure amount, conditional on income level. The relative expenditure is calculated as the difference between total healthcare costs up to the hospital admission and the allowed maximum expenditure in that income category. Black dots correspond to admissions by patients earning less 2 times the MMW, red triangles to admissions by patients with incomes between 2 and 5 times the MMW, and blue squares to admissions by patients with incomes above 5 times the MMW.

$530 before their expenditure limit. While, for enrollees in the highest income bracket, the jump in the probability of an admission happens around $1,090 before their OOP limit.

A second source of identification of consumer sensitivity to prices and insurer steering is the variation in hospital choice sets and prices across carriers. Table (4) presents several statistics of the number of in-network hospitals. The table shows considerable heterogeneity in network breadth across insurers and markets. EPS008 covers an average of 14.00 hospitals in a market, followed by EPS009 with 11.50, and EPS037 with 7.51. EPS003 has the least generous network, covering an average of 4.07 hospitals in a market. Because networks are not complete, I take them as fixed before insurers and hospitals engage in bilateral negotiations. Empirically, this means that insurers' disagreement payoffs will be underestimated relative to a situation where networks are endogenous. However, since there have been no relevant changes in insurers' networks over time for my final sample of hospitals, taking them as fixed is a natural assumption for my context.

In table (5), I explore the price variation across insurers and markets. There is important heterogeneity in negotiated prices across and within insurers. EPS013 has the lowest average negotiated price equal to $120.8 while EPS008 has an average price of $392.0. These two insurers are, respectively, in the lower and upper tail of the distribution of average number of in-network hospitals.

Table 4: Summary statistics of number of in-network hospitals

Insurer	Mean	SD
EPS001	5.07	3.63
EPS002	6.85	5.65
EPS003	4.07	2.58
EPS005	6.61	4.89
EPS008	14.00	4.69
EPS009	11.50	2.12
EPS012	5.67	3.51
EPS013	5.24	3.64
EPS016	7.28	5.60
EPS017	7.45	8.16
EPS018	4.92	3.73
EPS023	7.00	2.94
EPS037	7.51	6.68

57% of the variation in prices comes from differences across insurers, 39% from differences across hospitals, and 0.8% from differences across markets.

Table 5: Summary statistics of prices

Insurer	Mean	SD
EPS001	293.6	140.0
EPS002	145.4	55.0
EPS003	138.7	62.4
EPS005	122.4	47.9
EPS008	392.0	157.4
EPS009	475.0	471.3
EPS012	346.6	231.7
EPS013	120.8	50.0
EPS016	350.8	214.8
EPS017	126.9	41.3
EPS018	343.6	345.0
EPS023	129.1	45.3
EPS037	131.8	136.8

To investigate the relation between average negotiated prices and network breadth, I estimate the following equation via OLS:

$$\log(\overline{p}_{jt}) = \beta_0 + \beta_1 H_{jt} + \lambda_j + \eta_t + \varepsilon_{jt} \tag{1}$$

where $\overline{p}_{jt}$ is the average negotiated price of insurer j across the hospitals in market t, H_{jt} is the number of hospitals in the network of insurer j in market t, λ_j are insurer fixed effects, and η_t are market fixed effects.

Table 6: Network breadth and negotiated prices

	log(price)		
	(1)	(2)	(3)
Network	0.012**	0.010***	0.010*
	(0.006)	(0.003)	(0.006)
Fixed effects			
Insurer		X	X
Market			X
N	223	223	223
R^2	0.02	0.69	0.74

Note: OLS regression of the logarithm of average negotiated price by insurer and market on the number of in-network hospitals. Column (1) has no fixed effects. Column (2) includes insurer fixed effects. Column (3) includes insurer and market fixed effects. Robust standard errors in parenthesis. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

Results presented in table (6) show that there is a positive correlation between prices and networks in column (1) without fixed effects, in column (2) where insurer fixed effects are added, and in column (3) which additionally includes market fixed effects. This positive correlation is consistent with previous findings in the literature (Ho and Lee, 2017). Broad network carriers tend to negotiate higher prices because they can not use the threat of exclusion to exert a downward pressure on prices. It is also consistent with the stylized fact that broad network carriers have higher costs in equilibrium compared to narrow network insurers (Liebman, 2018). Carriers with broad networks also tend to have higher market shares in the total number of enrollees and in the total number of hospital admissions as seen in table (7). This table presents results of the following OLS regression:

$$s_{jt} = \beta_0 + \beta_1 H_{jt} + \lambda_j + \eta_t + \varepsilon_{jt} \tag{2}$$

where s_{jt} is the insurer market share in either the number of enrollees or number of admissions and the rest of variables are as before. The positive correlation between network breadth and insurer size in number of enrollees and number of admissions is indicative of consumers having strong preferences for broader networks. In figure (2), I further decompose this effect by presenting the coefficient on network breadth for separate regressions of the share of admissions by disease count and age group. The figure confirms the positive correlation between market share and network breadth for subsamples of enrollees, consistent with adverse selection in the Colombian insurance market. Although the coefficients are statistically equal across regressions by age group, there is a decreasing trend in the point estimate associated to network breadth that is suggestive of older

individuals having slightly lower preference for broader networks compared to younger enrollees.

Table 7: Network breadth and market share

	Share enrollees	Share admissions
Network	1.255***	1.253***
	(0.263)	(0.262)
Fixed effects		
Insurer	X	X
Market	X	X
N	223	223
R^2	0.64	0.64

Note: OLS regression of market share in number of enrollees and market share in number of admissions by insurer and market, on the number of in-network hospitals. Both regressions include insurer and market fixed effects. Robust standard errors in parenthesis. $^{***}p < 0.01,^{**}p < 0.05,^{*}p < 0.1$.

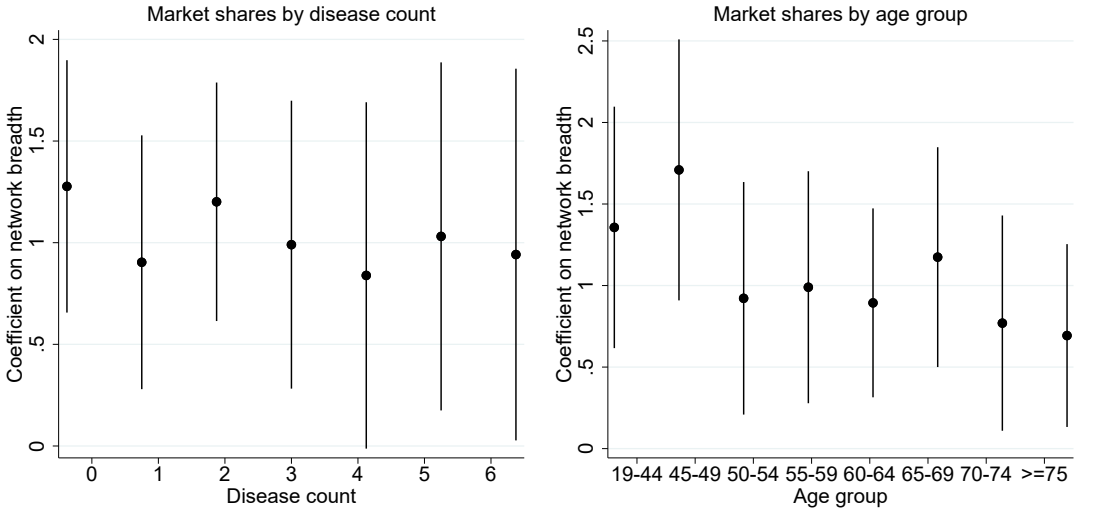


Figure 2: Heterogeneity by disease count and age group

Note: This figure presents the point estimate and 95% confidence interval on network breadth for separate OLS regressions of the share of admissions by disease count in the left panel, and of the share of admissions by age group in the right panel. Each regression includes insurer and market fixed effects.

4 Model

To answer the question of how reimbursement rates respond to cost sharing and whether the current cost sharing policy is optimal from the point of view of social welfare, I need a model of price formation and a model of hospital demand. The timing is as follows:

1. Insurers and hospitals bargain over the price of a hospital admission.

2. Patients receive a health shock and choose a hospital in the network of their insurer to receive treatment.

Since I do not observe insurer choice sets in each market, I can not model patients' decisions over insurance companies. So, to recover insurer demand for the first stage of the model, I follow the same strategy as in Gowrisankaran et al. (2015) and Prager and Tilipman (2020), using the enrollee's predicted willingness-to-pay for the network. I assume hospital networks are fixed during the bargaining game, which implies that my model and all counterfactual scenarios are estimated conditional on the patients' original choices of insurer. I can decompose demand elasticities into their consumer- and insurer-induced portions by leveraging variations in hospital choice sets across admissions that have zero coinsurance or happen after reaching the expenditure limit. Disregarding insurer steering would result in demand elasticities that are biased downwards because for patients who reach their OOP maximum we would wrongly predict that demand is perfectly inelastic. I further assume that patients are myopic about their future health status, so I rule out dynamic incentives by patients close to reaching their expenditure limit and focus on static discrete choices of hospitals.

4.1 Hospital demand

I start by describing the second stage of the model. I assume the choice of hospital depends not only on variables related to patients but also on variables related to insurers, since there is an underlying insurer steering mechanism that manifests in a reduced form way in the choice of hospital. The utility function of patient i for hospital h in the network of insurer j is given by:

$$v_{ijh} = \underbrace{\alpha_i^0 \kappa_i p_{jh} + \sum_{k=1}^{K} \sum_{l=1}^{L} \beta_{lk} x_{il} g_{hk} + \eta_h}_{\tilde{v}_{ijh}} + \alpha_i^1 (1 - \kappa_i) p_{jh} + \varepsilon_{ijh} \tag{3}$$

where,

$$\begin{aligned} \alpha_i^0 &= \overline{\alpha}^0 + X_i' \alpha^0 \\ \alpha_i^1 &= \overline{\alpha}^1 + X_i' \alpha^1 \end{aligned} \tag{4}$$

and $\tilde{v}$ denotes the part of utility due to patients, κ_i is the coinsurance rate for patient i that varies with income level, X_i are patient observable characteristics, g_h are hospital observable characteristics, and η_h is a hospital fixed-effect that captures unobserved hospital quality. All these variables also vary across markets, but to simplify notation I drop the market subscript. Finally, I assume ε_{ijh} follows a type-I extreme value distribution.

Identification. α_i^0 is identified from two types of variation: (i) the variation in hospital choice sets across patients in the same cost sharing tier before hitting their OOP maximum and, because my specification includes hospital fixed effects, (ii) from the variation in prices within hospitals. This price, however, can be endogenous if consumers selectively enroll carriers that have negotiated low prices with their preferred hospitals. To deal with this type of endogeneity, I rely on the fact that negotiated hospital prices are unobservable to consumers in Colombia when making insurer choices, and that consumers are generally inattentive with regard to the set of services covered by the national insurance plan. Moreover, since individuals in Colombia are generally myopic about their future healthcare costs and utilization, it is unlikely that their first insurer choice is made on the basis of how lenient the carrier is in permitting hospital choice once the enrollee reaches the OOP maximum.

To correct for possible selection into carriers I also follow Prager (2020)'s strategy. The author uses cost-sharing from past plan choices as an instrument for current cost-sharing, under the argument that if consumers have strong inertia over plans, then future or subsequent assignment to cost sharing rules are as-if-random. In my case, I don't observe the consumer's first carrier choice to use the associated cost sharing as an instrument for the coinsurance rate in my demand model. But, because my sample is constrained to enrollees who did not switch their insurance carrier over the three years, I need only drop 2009 from my estimation, since for this year I can not guarantee that the consumer's choice of carrier does not correspond to their first choice that potentially reflects selection. Thus, my main specification uses data from 2010 and 2011.

Another source of selection arises from hospitals negotiating higher prices with carriers where consumers have strong preferences for those hospitals. If this is the case, then my estimates of α_i^0 would be biased towards the null. I deal with this potential endogeneity in two ways. First, I explicitly allow α_i^0 to vary across patient demographics and diagnoses to capture the extent of selection on observables. This type of selection is relevant in my context because the government's risk-adjustment formulas poorly compensate for patient diagnoses. Second, I conduct a robustness check where I estimate hospital demand on the set of insurers that have below median negotiated

prices with star hospitals in each market. Then, I show that the predicted choice probabilities from this subsample after manually increasing prices, are similar to those estimated in the full sample for hospitals with above median prices. If so, the exercise will be suggestive of selection on unobservables not posing a concern in my setting. I describe this robustness check in more detail in the next section.

The coefficient on insurer steering α_i^1 is identified mainly from the variation in choice sets across patients that have reached their expenditure limit and face zero coinsurance, and from the variation in negotiated prices within hospitals. This coefficient could be biased towards zero if consumers tend to select cheaper hospitals for their first admission and continue to visit that hospital due to inertia or brand loyalty, which the model would interpret as an aversion to expensive hospitals. To isolate the effect of prices from the effect of provider inertia, I include an indicator for previous chosen hospitals in the consumer's utility function.

Another potential source of endogeneity is the fact that patients who reach their OOP maximum may differ significantly from patients that fail to reach the spending limits in terms of their ex-ante risk. If consumers that hit the OOP thresholds are riskier or sicker than their counterparts and are significantly less responsive to prices, then α_i^1 could be biased towards the null. To account for differences in ex-ante risk, I allow the insurer steering coefficient to vary across patient demographics and diagnoses. Patient observable characteristics included in these interaction terms are the ones used by the government to calculate the ex-ante risk-adjusted transfers, so they fully capture variation in consumer ex-ante risk that is observable to insurance companies. While it is possible that insurers change how stringent they are in their steering practices depending on the level of cost sharing, this is not a first order issue in my counterfactual analysis.

In second term on the right-side of equation (3), $\beta_{lk} g_{hk}$ represents the marginal utility of patients with traits x_{il} for hospitals with characteristics g_{hk}. These coefficients are identified from within-hospital variation in patient and admission characteristics. Identification of patient characteristics is achieved only from their interaction with hospital characteristics, because I assume there is no outside option.

Patient i and insurer j to which she is enrolled choose a hospital to maximize utility. If $v_{ijh} \geq v_{ijk}$, $\forall k \neq h$, then hospital h is chosen. The individual choice probability is $s_{ijh} = E[v_{ijh} \geq v_{ijk} | p_{jh}, \mathbf{x}_i, \kappa_i, \mathbf{g}_h]$, which takes the known logistic form after integrating out the distribution of ε_{ijh}:

$$s_{ijh} = \gamma^a_{j(i)l(i)} \frac{exp(\delta_{ijh})}{\sum_{k \in \mathcal{H}_j} exp(\delta_{ijh})} \tag{5}$$

Here $\delta_{ijh} = \tilde{v}_{ijh} + \alpha_1(1-\kappa_i)p_{jh}$, $\mathcal{H}_j$ is the set of hospitals in the network of insurer j, and $\gamma^a_{j(i)l(i)}$ is the probability that a consumer type l enrolled to insurer j is admitted to the hospital. I define a consumer type as a combination of sex, age group (19-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, >=75), and diagnosis (cancer only, cardiovascular only, diabetes only, renal only, other disease only, 2 or more diseases, no diseases). These probabilities are calculated non-parametrically with data on all enrollees to each insurer. Appendix (D) presents some summary statistics of the probability of a hospital admission by subsample of patients and insurer. Using the estimated coefficients from (3) and following McFadden (1996), the consumer's expected utility for the set of hospitals in the network of insurer j is:

$$W_{ij}(\mathcal{H}_j, p_{jh}, \kappa_i, \mathbf{g}_h) = \gamma^a_{j(i)l(i)} \log\left(\sum_{h\in\mathcal{H}_j} exp(\delta_{ijh})\right) \tag{6}$$

As in Capps et al. (2003), Town and Vistnes (2001), and Ho (2006), I refer to the difference $W_{ij}(\mathcal{H}_j, p_{jh}, \kappa_i, \mathbf{g}_h) - W_{ij}(\mathcal{H}_j \backslash h, p_{jh}, \kappa_i, \mathbf{g}_h)$ as the consumer's willingness-to-pay for hospital h.

4.2 Insurer-hospital bargaining

In the first stage of the model, insurers and hospitals bargain over the price of an admission to maximize their joint surplus. Both agents have complete information about patient characteristics, size of enrollee pools, and hospital characteristics and costs. The insurers' profit function is a weighted average of patient welfare and own profits:

$$\pi_j(\mathcal{H}_j, \{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j}, \kappa_i) = \tau \sum_{i\in N_j} \frac{1}{|\alpha^1_i|} W_{ij}(\mathcal{H}_j, \mathbf{p}_{jh}, \kappa_i, \mathbf{g}_h) - \sum_{i\in N_j}\sum_{h\in\mathcal{H}_j} (1-\kappa_i) p_{jh} s_{ijh}(\mathcal{H}_j, \mathbf{p}_{jh}, \kappa_i, \mathbf{g}_h) \tag{7}$$

where N_j denotes the set of patients enrolled to j. Insurers take into account patient welfare weighted by an altruism parameter τ (first term in the right-hand side of the equation) and incur in the cost of treatment net of the patient's coinsurance payments (second term on the right-hand side of the equation). Insurers weigh patient welfare with own profits because every year they undergo government evaluations of their service quality and enrollee satisfaction. Following Gowrisankaran et al. (2015), $\tau = 1$ implies that insurer and patient incentives are perfectly aligned, while $\tau < 1$ means that insurers care more about financial profits than consumer welfare. The first term in the right-hand side of equation (7) is divided by $|\alpha_1|$ to convert consumer expected utility into dollars *as perceived by the insurer*. By dividing the consumer's expected utility by $|\alpha^1_i|$ instead of $|\alpha^0_i|$, I am

effectively allowing insurers to perceive that the derivative of demand with respect to prices might be different from that of consumers. This is similar to Ho and Lee (2017) who include an elasticity scaling parameter in the first order condition of the Nash joint surplus function with respect to premiums. In my case, $|\alpha_i^1|$ reflects the true value of a dollar to the insurer and captures insurer's incentives to engage in different steering efforts depending on whether the patient has reached the OOP limit or not.[2]

Hospital profits are given by their markups times hospital demand as seen below:

$$\pi_h(\mathcal{J}_h, \{\mathbf{p}_{jh}\}_{j\in\mathcal{J}_h}) = \sum_{j\in\mathcal{J}_h}\sum_{i\in N_j}(p_{jh} - mc_{jh})s_{ijh}(\mathcal{H}_j, \mathbf{p}_{jh}, \kappa_i, \mathbf{g}_h) \tag{8}$$

$\mathcal{J}_h$ is the set of insurers that include hospital h in their network, mc_{jh} is the marginal cost to hospital h of admitting a patient enrolled to insurer j, and hospital demand is given by $q_{jh} = \sum_{i\in N_j} s_{ijh}$. The marginal cost varies across hospitals because of differences in technology, physical capital, and human capital. It also varies across insurers because different carriers might involve different administrative costs to the hospital. I assume marginal costs are constant to avoid externalities between the pricing rules of different insurer-hospital pairs and model them as a function of exogenous variables $\mathbf{v}$ (dummy variables for each hospital) and a structural error ψ_{jh} that is specific to the insurer-hospital pair.

$$mc_{jh} = \lambda\mathbf{v}_h + \psi_{jh} \tag{9}$$

Assuming insurers and hospitals engage in Nash bargaining, I follow Horn and Wolinsky (1988) to define the Nash-in-Nash equilibrium of the bargaining game where each pair of agents chooses contract prices conditional on the equilibrium choices of every other pair. The joint surplus of an insurer-hospital pair is a Cobb-Douglas function where the exponents represent the bargaining power of each agent:

$$\begin{aligned} S_{jh}(p_{jh_{h\in\mathcal{H}_j}}|\mathbf{p}_{jk_{k\in\mathcal{H}_j\setminus h}}) = &\Big(\pi_j(\mathcal{H}_j, \{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j}, \kappa_i) - \pi_j(\mathcal{H}_j\setminus h, \{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j\setminus h}, \kappa_i)\Big)^{\beta_j} \\ &\times (\pi_h(\mathcal{J}_h, \{\mathbf{p}_{jh}\}_{j\in\mathcal{J}_h}) - 0)^{1-\beta_j} \end{aligned} \tag{10}$$

For insurer j, the outside option during negotiations with hospital h is given by the equilibrium profits it would obtain from all other hospitals except h, represented by the term $\pi_j(\mathcal{H}_j\setminus h, \{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j\setminus h}, \kappa_i)$.

[2] If $|\alpha_i^0| > |\alpha_i^1|$, then dividing the consumer's expected utility by $|\alpha_i^1|$ would result in an estimate of τ that is smaller than if the expected utility were divided by $|\alpha_i^0|$. In that case, dividing by $|\alpha_i^1|$ makes it easier for the insurer to satisfy the incentive compatibility constraint in the joint Nash surplus maximization problem.

This outside option is exogenous given the assumption of no externalities. For hospital h, the disagreement payoff equals zero because patients who want to visit this hospital can not switch from j towards an insurer who includes h in its network given the assumption of fixed enrollee pools and empirical evidence of zero switching rates.

The problem of an insurer-hospital pair is:

$$\begin{aligned} & \max_{p_{jh}} \; S_{jh}(p_{jh_{h\in\mathcal{H}_j}}|\mathbf{p}_{jk_{k\in\mathcal{H}_j\backslash h}}) \\ s.t \;\; & \pi_j(\mathcal{H}_j,\{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j},\kappa_i) - \pi_j(\mathcal{H}_j\backslash h,\{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j\backslash h},\kappa_i) \geq 0 \\ & \pi_h(\mathcal{J}_h,\{\mathbf{p}_{jh}\}_{j\in\mathcal{J}_h}) \geq 0 \\ & \pi_j(\mathcal{H}_j,\{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j},\kappa_i) \geq 0 \end{aligned} \tag{11}$$

Maximizing the joint surplus function with respect to prices gives the following expression:

$$(1-\beta_j)\frac{q_{jh}+\sum_{k\in\mathcal{J}_h}\frac{\partial q_{jk}}{\partial p_{jh}}(p_{jk}-mc_{jk})}{\sum_{k\in\mathcal{J}_h}q_{jk}(p_{jk}-mc_{jk})} = \beta_j\frac{\frac{\partial \pi_j(\cdot)}{\partial p_{jh}}}{\pi_j(\cdot,\mathcal{H}_j)-\pi_j(\cdot,\mathcal{H}_j\backslash h)}$$

where $q_{jh}=\sum_{i\in N_j}s_{ijh}$. Notice that if $\beta_j=0$, the Nash surplus maximization problem collapses to one of hospital profit maximization, so the FOC denotes the usual Nash-Bertrand solution to the hospital's problem. In that case, prices equal marginal costs plus a markup that is a function only of quantities and the derivatives of hospital demand with respect to prices $\frac{\partial q_{jk}}{\partial p_{jh}}$. In the context of bargaining where $\beta_j>0$, the derivative of the joint surplus function with respect to prices involve not only the derivatives of hospital demand $\frac{\partial q_{jk}}{\partial p_{jh}}$, but also the derivatives of insurer demand $\frac{\partial \pi_j(\cdot)}{\partial p_{jh}}$. To the extent that insurer demand decreases with hospital prices, $\frac{\partial \pi_j(\cdot)}{\partial p_{jh}}$ will be a negative semi-definite matrix, which means that the price-elasticity of demand is greater in the Nash bargaining game than in Bertrand competition.

Let $\Upsilon=\frac{\beta}{1-\beta}\frac{C}{D}q$ with $C=\frac{\partial \pi_j(\cdot)}{\partial p_{jh}}$ and $D=\pi_j(\cdot,\mathcal{H}_j)-\pi_j(\cdot,\mathcal{H}_j\backslash h)$. Solving for equilibrium prices and rewriting in matrix notation yields:

$$\mathbf{p}=\mathbf{mc}-\left(\Omega^i+\Omega^j+\Upsilon\right)^{-1}\mathbf{q} \tag{12}$$

I denote the term $\Omega^i+\Omega^j+\Upsilon$ as the "effective" price-elasticity. Ω^i is the part of the effective elasticity explained by consumer sensitivity to prices, Ω^j by insurer steering, and Υ is the additional sensitivity to prices introduced by insurer demand in the bargaining process with hospitals. The

expression for each of these matrices is provided in appendix (B). If insurer steering matters, the effective demand elasticity will be greater in magnitude than in a context where insurers do not have mechanisms to steer demand other than premiums. Failure to account for insurer steering, results in demand elasticities that are biased downwards, and thus in either underestimation of the insurer's bargaining power or in negative marginal costs.

Identification. Equation (12) denotes the inverse mapping from the model to the primitive, mc_{jh}. Given that prices are observed and, conditional on β_j and τ, effective-price elasticities and quantities are all estimated in the model, the only unobservable in equation (12) is the marginal cost. After solving for marginal costs, the parameters in λ in equation (9) can be identified from an OLS regression of mc on $\mathbf{v}$ under the full rank condition. The cost shocks in this regression are endogenous because they are observed by hospitals before bargaining takes place, so they affect the choice of prices in the bargaining stage. For example, if the hospital makes an investment in capacity that increases its marginal costs then estimates for β_j will be biased towards 1. Or if the hospital is vertically integrated with the insurer, which lowers its marginal cost, then β_j will be biased towards 0.

Following Gowrisankaran et al. (2015) and Ho and Lee (2017), I use the predicted choice probabilities and the predicted willingness-to-pay for each hospital, that would result from setting prices equal to the average price in the market, as instruments for β_j. The intuition here is that since ψ_{jh} is mean-independent of observed and exogenous hospital characteristics, the predicted choice probabilities and predicted willingness-to-pay per hospital with market-level average prices are uncorrelated with ψ_{jh} but likely correlated with prices. I also include the hospital fixed effects in $\mathbf{v}$ in the instrument set.[3] Because I exploit the price variation across insurers to identify β_j, the remaining price variation across hospitals helps identify τ. For example, if enrollees to insurer j have a high willingness-to-pay for hospital h, this hospital will be able to bargain a relatively high price with j relative to other hospitals in the network. The extent of price variation across hospitals is captured by the relative importance of willingness-to-pay in the insurer profit function, which helps identify τ. I estimate the non-linear parameters, τ and β_j, using GMM under the assumption that $E[\psi|Z] = 0$, where Z is the matrix of instruments.

[3]Hospital fixed effects included both in the hospital demand model and in the marginal cost regression account for much of the endogeneity arising from consumer preferences for observed hospital characteristics, which guarantees the mean-independence of ψ_{jh}.

5 Estimates

5.1 Demand for hospitals

Table (8) shows the results of the hospital choice model. This model is estimated via maximum likelihood using a conditional logit specification with hospital fixed effects, normalizing the largest hospital in each market to zero. Without accounting for insurer steering, my estimates suggest that if a hospital increases the price of an admission by 10 USD, the probability of a patient choosing it decreases by 2.33% ($= \hat{\alpha}^0\overline{\kappa}_i$). After accounting for insurer steering, this percentage decrease is closer to 4.13% ($= \hat{\alpha}^0\overline{\kappa}_i + \hat{\alpha}^1(1-\overline{\kappa}_i)$). The effect of insurer steering on hospital demand is not muted by provider inertia, even though my estimates indicate that previous chosen hospitals are 2.65% more likely to be visited compared to a new provider.

I find that individuals with chronic diseases like cardiovascular disease, renal disease, or with more than two comorbidities, are less responsive to out-of-pocket coinsurance payments than patients without diagnoses. The coefficients associated to these diseases in the interaction with coinsurance payments are all positive and significant. Patients aged 65 or older are less responsive to out-of-pocket prices compared young patients, while males and females have a similar sensitivity to price. Old males are less likely to be steered by their insurance company compared to young females. Insurer steering is also less likely among the set of patients with chronic conditions, such as cancer and cardiovascular disease, than among healthy individuals.

The interactions between demographics and diagnoses with hospital characteristics show that males have stronger preferences for private hospitals with a higher number of beds and ambulances, compared to females. Patients aged at least 65 have stronger preferences for beds rather rooms, while the opposite is true for individuals aged at most 64. Preference heterogeneity across diagnoses shows that patients with cancer prefer hospitals with a higher number rooms, while those with cardiovascular diseases prefer hospitals with a higher number of beds. Individuals with more than two comorbidities have strong preferences for bigger hospitals and are 6% more likely to visit a private hospital than a public one. To better interpret demand results, appendix (E) presents the distribution of own- and cross-price elasticities before bargaining takes place and explores some potential sources of elasticity heterogeneity. The results in this appendix suggest that the price-elasticity of demand increases with the cost-sharing tier similar to findings in Einav et al. (2018). The elasticity is decreasing with age and sickness level, and increasing with hospital size.

To account for the potential endogeneity in coinsurance payments arising either (i) from hospitals

Table 8: Hospital demand

		Coefficient	Std. error
$\kappa_i p_{jh}$		-17.973***	(2.202)
$(1-\kappa_i)p_{jh}$		-2.070***	(0.313)
Previous provider		2.650***	(0.023)
Interactions			
$\kappa_i p_{jh}$	Male	1.668	(1.332)
	Age>=65	10.377***	(1.412)
	Cancer	0.742	(3.332)
	Cardio.	2.713	(2.634)
	Diabetes	-4.028	(8.313)
	Renal	18.006***	(6.754)
	Other	9.944**	(4.086)
	>=2 diagnoses	5.834***	(2.244)
	Healthy	(ref)	(ref)
$(1-\kappa_i)p_{jh}$	Male	0.503***	(0.150)
	Age>=65	0.029	(0.154)
	Cancer	1.187***	(0.389)
	Cardio.	1.317***	(0.346)
	Diabetes	0.070	(0.774)
	Renal	-1.418	(1.034)
	Other	-0.837	(0.566)
	>=2 diagnoses	1.188***	(0.300)
	Healthy	(ref)	(ref)
Beds	Male	0.033***	(0.008)
	Age>=65	0.086***	(0.009)
	Cancer	-0.061***	(0.018)
	Cardio.	0.061***	(0.015)
	Diabetes	0.046	(0.043)
	Renal	0.128***	(0.049)
	Other	0.070***	(0.020)
	>=2 diagnoses	0.057***	(0.012)
	Healthy	(ref)	(ref)
	Income	-0.072***	(0.010)
	After OOP	0.021**	(0.009)
Rooms	Male	-0.003	(0.002)
	Age>=65	-0.026***	(0.002)
	Cancer	0.039***	(0.004)
	Cardio.	-0.020***	(0.005)
	Diabetes	-0.007	(0.014)
	Renal	-0.007	(0.015)
	Other	0.002	(0.006)
	>=2 diagnoses	0.007**	(0.003)
	Healthy	(ref)	(ref)
	Income	0.017***	(0.002)
	After OOP	0.028***	(0.003)
Any ambulance	Male	0.113***	(0.024)
	Age>=65	0.034	(0.027)
	Cancer	-0.213	(0.052)
	Cardio.	0.014	(0.039)
	Diabetes	0.041	(0.110)
	Renal	-0.236*	(0.137)
	Other	-0.127**	(0.055)
	>=2 diagnoses	-0.009	(0.031)
	Healthy	(ref)	(ref)
	Income	-0.075***	(0.028)
	After OOP	-0.113***	(0.026)
Private	Male	0.055**	(0.022)
	Age	-0.177***	(0.025)
	Cancer	-0.096**	(0.048)
	Cardio.	0.100***	(0.038)
	Diabetes	-0.013	(0.110)
	Renal	-0.161	(0.128)
	Other	-0.043	(0.051)
	>=2 diagnoses	0.066**	(0.030)
	Healthy	(ref)	(ref)
	Income	0.079***	(0.023)
	After OOP	0.005	(0.024)
N		1,066,982	
Pseudo-R^2		0.22	

Note: Maximum likelihood estimation of patient demand for hospitals. Includes hospital fixed effects, normalizing the largest hospital in each market to zero. Prices are measured in thousands of USD. Number of beds is measured in hundreds. Robust standard errors reported in parenthesis. ***p<0.01,**p<0.05,*p<0.1

negotiating higher prices with carriers where consumers have strong preferences for those hospitals, or (ii) from consumers enrolling carriers that have negotiated low prices with their preferred hospitals, I conduct a robustness check where I estimate demand on the sample of carriers that have negotiated below median prices with star hospitals in each market. Appendix table (C2) presents the resulting demand estimates. If selection on unobserved consumer preferences is not an issue, then the choice probabilities predicted from this alternative specification after manually increasing prices, should be the same as the choice probabilities for star hospitals that have negotiated above median prices with insurers in my main demand specification. Appendix table (C3) presents a Kolmogorov-Smirnov test on the equality of the distribution of average choice probabilities in these samples. In all cases, the null hypothesis that the distributions are the same can not be rejected.

I conduct additional robustness checks on my definition of network, given that I recover hospital networks from observed claims. In appendix table (C1), I estimate demand defining the network relative to hospitals that have an ICU bed and relative to hospitals with at least one room. My main estimators associated to coinsurance payments and insurer steering are robust to these alternative network definitions.

5.2 Bargaining game

I estimate the bargaining model on data from the main markets in the country (Antioquia 05, Atlántico 08, Bogotá 11, Valle del Cauca 76) for computational simplicity. These markets represent 69% of all hospital admissions in the sample period. Table (9) shows the estimators for τ and β_j together with standard errors in parenthesis based on 100 bootstrap samples. The Nash bargaining protocol assumes that, conditional on hospital networks, every insurer-hospital pair obtains a positive surplus, otherwise the agreement breaks down. This provides a natural lower bound on τ that has insurers satisfy their incentive compatibility and rationality constraints from equation (11). The estimator for τ is statistically greater than one, which suggests that insurance companies place a weight of 31% $(= 1 - \hat{\tau}/(1+\hat{\tau}))$ on financial profits and 69% $(= \hat{\tau}/(1+\hat{\tau}))$ on enrollee welfare. With the exception of EPS008, EPS009, and EPS012, insurers extract most of the surplus in their interaction with hospitals. The bargaining parameters for all other insurers are significant and greater than 0.5.

In table (10), I present the resulting (quantity-weighted) average marginal costs and the average Lerner indexes, with their standard errors in parenthesis. Hospitals enjoy greater markups in markets with fewer competitors. In market 05, which has 224 insurer-hospital pairs, the average

Table 9: Bargaining parameters

	Estimate	SE
τ	2.207	(0.499)
βs		
EPS001	0.602	(0.039)
EPS002	0.813	(0.046)
EPS003	0.611	(0.041)
EPS005	0.633	(0.023)
EPS008	0.293	(0.133)
EPS009	0.461	(0.041)
EPS012	0.164	(0.057)
EPS013	0.702	(0.041)
EPS016	0.895	(0.139)
EPS017	0.605	(0.110)
EPS018	0.550	(0.060)
EPS023	0.767	(0.036)
EPS037	0.702	(0.049)

Note: This table shows the estimates of τ and β_j using GMM. Standard errors in parenthesis are based on 100 bootstrap samples of admissions within insurer-hospital-market.

Lerner index equals 12.48%. While in market 08, which has 88 insurer-hospital pairs, the average Lerner index is 43.08%. I find that all insurers and over 94% of hospitals in each market satisfy their incentive compatibility and rationality constraints given the estimators for τ and β_j.

Table 10: Average marginal costs and Lerner indexes

Market	Price	Marginal cost	Lerner index	Total pairs	Hospitals w/ surplus>0	Insurers w/ surplus>0
05	181.37	158.74 (23.02)	12.48 (17.09)	224	94.6%	100%
08	127.57	72.62 (11.05)	43.08 (9.12)	88	97.7%	100%
11	179.00	141.08 (39.45)	21.18 (9.87)	304	98.7%	100%
76	200.53	156.73 (24.49)	21.84 (15.89)	129	95.3%	100%

Note: This table reports the (quantity-weighted) average price, average marginal cost, average Lerner index, total number of insurer-hospital pairs, and percentage of hospitals and insurers with positive Nash surplus in the four main markets. Standard errors in parenthesis are based on 100 bootstrap samples of admissions at the insurer-hospital-market level.

5.3 Elasticity decomposition

Using the estimated profit parameters, table (11) decomposes the effective price elasticity into its three components Ω^i, Ω^j and Υ, and reports their averages and standard errors (in parenthesis) across admissions in the full sample, and in the sub-samples of admissions before and after reaching the OOP maximum. The effective elasticity is greater in the sub-sample of admissions that happen before reaching the expenditure limit than after, because consumers –the most price sensitive side

of the market– explain a higher proportion of the elasticity of demand than insurers. My estimate of the effective elasticity in column (2) is smaller than the one in Gowrisankaran et al. (2015) who find an average ranging from 1.7 to 4.6. This suggests that the strict regulation in Colombia can make demand relatively more inelastic than if insurers were allowed to compete in premiums or set plan characteristics.

In the full sample, the percentage of the effective elasticity explained by insurer steering (61.25%) is greater than the portion explained by consumer sensitivity to prices (37.04%), followed by the part due to insurer-hospital bargaining (1.72%). Before reaching the OOP maximum, this relation reverses, with consumers explaining 52.95% of the effective elasticity and insurers accounting for 45.41%. In the sub-sample of admissions that happen after the OOP limit, insurers explain 97.95% of the effective elasticity, while bargaining accounts for only 2.05%. This effect of coinsurance rate discontinuities on hospital demand will be important for the counterfactuals where I measure welfare associated to different cost-sharing policies. The optimal coinsurance rate schedule will depend on which side of the market –consumers or insurers– has a more elastic demand.

Table 11: Own-price elasticity decomposition per market

Sample	Actual (1)	Effective (2)	Consumer (3)	Insurer (4)	Bargaining (5)
Full	-0.197 (0.005)	-0.200 (0.005)	37.04 (0.36)	61.25 (0.65)	1.72 (0.61)
Before OOP maximum	-0.296 (0.004)	-0.300 (0.004)	52.95 (0.44)	45.41 (0.42)	1.64 (0.17)
After OOP maximum	-0.129 (0.022)	-0.134 (0.022)	0.00	97.95 (0.25)	2.05 (0.25)

Note: Column 1 reports the (quantity-weighted) average actual own-price elasticity per sample given by $\Omega_i + \Omega_j$. Column 2 reports the (quantity-weighted) average effective own-price elasticity given by $\Omega_i + \Omega_j + \Upsilon$. Columns (3) to (5) show the percentage of the effective own-price elasticity explained by consumer sensitivity to prices, insurer steering, and insurer-hospital bargaining, respectively. Standard errors in parenthesis based on 100 bootstrap samples of admissions at the insurer-hospital-market level.

6 The Effect of Cost Sharing on Equilibrium Prices

In this section, I simulate the impact of alternative cost sharing rules on equilibrium prices and decompose the effect into its portions explained by consumer sensitivity to prices, insurer steering, and insurer-hospital bargaining. For the counterfactual exercises, I assume that hospital networks, enrollee pools, and the parameters of the utility and profit functions are fixed. To calculate the proportion of patients that reach their expenditure threshold in the counterfactuals, I further assume that healthcare utilization and prices for all other services and procedures, except hospital admissions, remain fixed. Therefore, up to hospital admissions, total healthcare costs will be equal to observed costs and the OOP expenditure will depend on the counterfactual cost sharing rules.

Finally, I assume that capacity constraints for hospitals are not binding. All my counterfactual exercises are computed with data from the four main markets in the country during 2011.

6.1 From zero to full coinsurance

I start by describing a counterfactual setting where coinsurance rates are equal to zero. When patients face zero costs of their healthcare treatment, demand for hospitals becomes more inelastic and patients over-demand admissions relative to the base scenario. In this case, the FOC of the joint surplus maximization problem is:

$$(1-\beta_j)(\pi_j(\mathcal{H}_j, p^*_{jh}) - \pi_j(\mathcal{H}_j \backslash h, p^*_{jh})) = \beta_j \pi_h(\mathcal{J}_h, p^*_{jh})$$

Therefore, equilibrium prices will set hospital profits equal to the hospital's marginal value to the insurer as in Gowrisankaran et al. (2015). In particular, the counterfactual prices $\tilde{\mathbf{p}}$ that solve the FOC take the following form:

$$\tilde{\mathbf{p}} = \mathbf{mc} - \left(\tilde{\Omega}^j(\tilde{\mathbf{p}}) + \tilde{\Upsilon}(\tilde{\mathbf{p}})\right)^{-1} \tilde{\mathbf{q}}(\tilde{\mathbf{p}}) \tag{13}$$

Here $\tilde{\Omega}^j$ is the counterfactual matrix of demand elasticities due to insurer steering, $\tilde{\Upsilon}$ is the counterfactual matrix of derivatives of insurer profits (defined previously), and $\tilde{\mathbf{q}}$ is the counterfactual demand. Equation (13) shows that even though Ω^i equals zero, insurer steering and hospital bargaining generate a non-zero effective demand elasticity. Insurer sensitivity to prices will constrain price increases under zero coinsurance relative to a situation without steering, and insurer-hospital bargaining will constrain prices relative to the Nash-Bertrand solution. I compute the new equilibrium outcomes using an iterative procedure in prices until convergence up to a tolerance level.

In the polar case where coinsurance rates are equal to 1, my specification of insurer profits plays an important role in the resulting equilibrium prices. Although insurers' total costs equal zero before patients reach their spending limits, the full coinsurance policy affects their marginal profits through its effect on patient marginal utility and through the rate at which patients reach the OOP maximum. In this case, the FOC of the joint surplus maximization problem is given by:

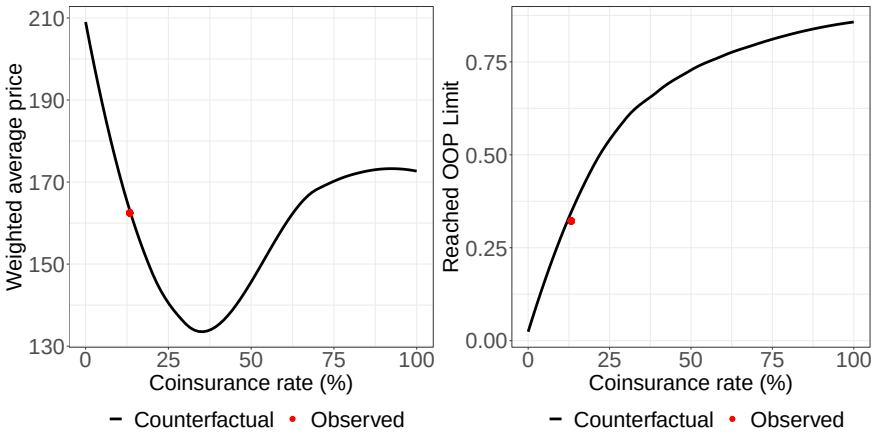
$$(1-\beta_j)\frac{q_{jh} + \sum_{j\in\mathcal{J}_h} \frac{\partial q_{jh}}{\partial p^*_{jh}}[p^*_{jh} - mc_{jh}]}{\pi_h(\mathcal{J}_h, p^*_{jh})} = \beta_j \frac{\frac{\partial W}{\partial p^*_{jh}}}{\pi_j(\mathcal{H}_j, p^*_{jh}) - \pi_j(\mathcal{H}_j \backslash h, p^*_{jh})}$$

Hence, equilibrium prices will set the marginal profit to hospitals equal to the marginal pa-

tient's willingness-to-pay. Notice that if insurers placed zero weight on patient welfare and cared entirely about minimizing costs, the full coinsurance scenario would imply that there is no insurer intermediation between patients and hospitals, and as a result equilibrium prices would be equal to Nash-Bertrand prices. The prices that solve the FOC are given by:

$$\tilde{\mathbf{p}} = \mathbf{mc} - \left(\tilde{\Omega}^{i}(\tilde{\mathbf{p}}) + \tilde{\Omega}^{j}(\tilde{\mathbf{p}}) + \tilde{\Upsilon}(\tilde{\mathbf{p}})\right)^{-1} \tilde{\mathbf{q}}(\tilde{\mathbf{p}})$$

where $\tilde{\Omega}^{i}$ is the counterfactual matrix of demand elasticities due to consumers. Insurer steering is non-zero because of the non-linearity in insurance contracts. As the coinsurance rate increases while holding the maximum OOP expenditure fixed, the more likely is a patient to reach her expenditure limit. So, as long as insurers are less responsive to price compared to consumers, we should see prices increase with marginal increases in the coinsurance rate at high levels of this rate.[4]



(a) Average price (b) Proportion that reach the OOP limit

Figure 3: Average price and proportion that reach the OOP limit under counterfactual coinsurance rates

The solid black line in panel (a) of figure (3) presents the (quantity-weighted) average equilibrium price under *uniform* coinsurance rates. The red dot corresponds to the observed average price. In this counterfactual exercise, I eliminate the observed three-tier system for coinsurance rates

[4]This result could be driven in part by the assumption that healthcare utilization for services other than hospital admissions does not change with the coinsurance rate. While this assumption is standard in discrete hospital demand and bargaining models, there is recent evidence suggesting that consumers adjust their medical spending to changes in the cost-sharing schedule (Marone and Sabety, 2021; Ho and Lee, 2021). In my counterfactual with zero coinsurance, this type of consumer moral hazard would imply that changes in insurers' total costs are going to be underestimated relative to a scenario where healthcare utilization is allowed to adjust with cost-sharing. In my counterfactual with a coinsurance rate of 100%, consumer moral hazard implies that insurers' total costs are going to be overestimated. My conclusions regarding optimal cost-sharing while holding healthcare utilization fixed are, nonetheless, similar to those obtained in Ho and Lee (2021).

and assign the same percentage to all patients, while holding fixed their income-indexed OOP maximums. The average equilibrium price is U-shaped in the coinsurance rate, with the inflection point happening at 30% coinsurance. First, average prices decrease 34% when the coinsurance rate increases from 0 to 30%. Then, prices increase 20% when coinsurance rates go from 30% to 100%. This pattern is consistent with the intuition outlined before: as the coinsurance rate increases, demand becomes more elastic because consumers face a higher proportion of their healthcare costs. But, as coinsurance rates continue to rise, consumers are also more likely to hit their OOP maximum after which the insurer, that is the less elastic side of the market, has to cover the full cost of healthcare. Panel (b) of the figure shows that, in fact, the fraction of patients who reach their spending limit is increasing and concave with respect to the coinsurance rate. This fraction changes from 0 to 0.60 as the coinsurance rate increases from 0 to 30%, and then from 0.60 to 0.78 as the coinsurance rate increases further to 100%. The elasticity patterns associated to these counterfactual prices are presented in figure (4). The elasticity curve is also U-shaped in the coinsurance rate.

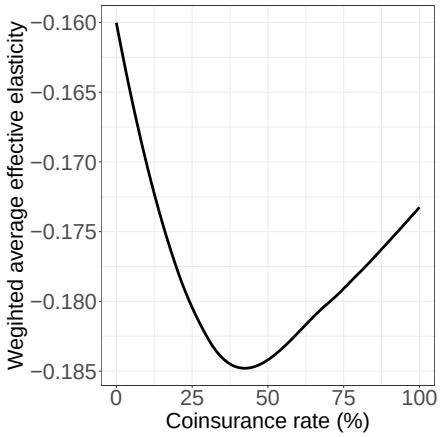
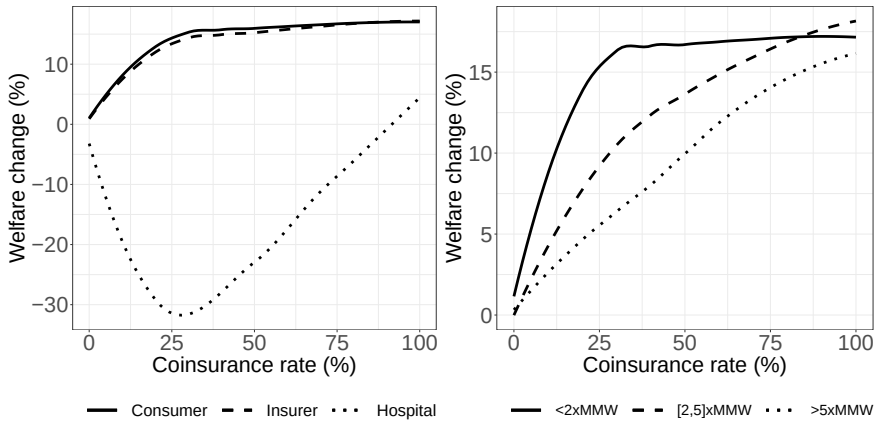


Figure 4: Average effective price elasticity under counterfactual coinsurance rates



(a) Total welfare changes

(b) Consumer welfare changes by income group

Figure 5: Welfare changes under counterfactual coinsurance rates

Panel (a) of figure (5) presents the percentage change in consumer welfare, insurer profits, and hospital profits, from these counterfactual exercises relative to the baseline scenario of full insurance, to facilitate comparisons with the existing literature on optimal cost-sharing. Consumer welfare is calculated as $\sum_i \frac{1}{|\alpha_i^0|} W_i(\mathcal{H}_j, \mathbf{p}_{jh}, \kappa_i, \mathbf{g}_h)$, insurer profits as $\sum_j \pi_j(\mathcal{H}_j, \{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j}, \kappa_i)$, and hospital profits as $\sum_h \pi_h(\mathcal{J}_h, \{\mathbf{p}_{jh}\}_{j\in\mathcal{J}_h})$. Hospital surplus follows the same pattern as the average equilibrium price. When the coinsurance rate equals 30%, demand is more elastic than baseline and total hospital profits decrease 30%. At a coinsurance rate of 100% when demand is relatively inelastic, hospital profits increase 5% relative to the full insurance scenario.

Consumer surplus is concave and greater than the full insurance scenario at every value of the coinsurance rate. Consumer welfare is maximized at 30% coinsurance, because even though individuals face a higher proportion of their healthcare costs, they also reach their spending limits at a higher rate than baseline. Beyond 30% coinsurance, welfare losses of individuals who face higher OOP costs and fail to reach the spending thresholds, partially compensate welfare gains of those who reach the OOP limit. As a result, consumer welfare is relatively flat at values of the coinsurance rate above 30%.

My results highlight the importance of controlling for discontinuities in OOP prices when analyzing welfare in health insurance markets. Failure to account for these discontinuities, would lead a researcher to predict that an increase in the coinsurance rate indistinctly lowers consumer surplus. This result is not driven by individuals being forward-looking nor internalizing the dynamic

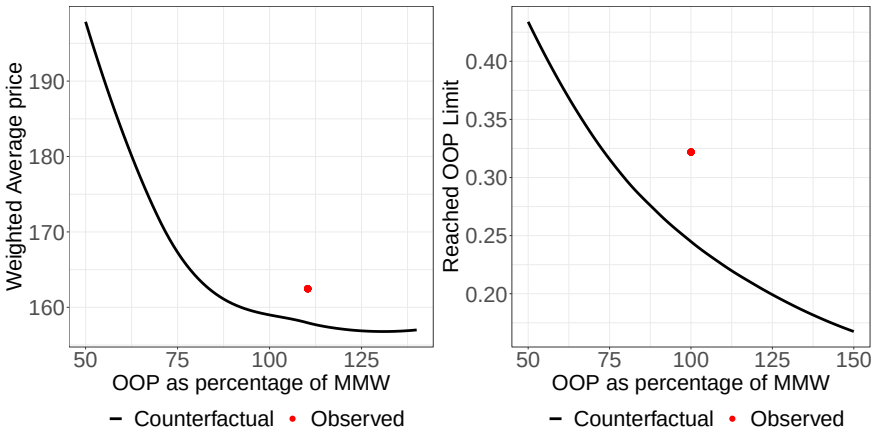
incentives introduced by the spending thresholds, but by insurance companies engaging in steering practices after patients reach their OOP limit.

My findings also suggest that a policymaker should take into account the distributional welfare effects of cost-sharing policies. Modifying elements of the cost-sharing system can have different impacts on the financial default risk due to healthcare expenses for consumers of different income levels. To get at the heterogeneous welfare effects, in panel (b) of figure (5), I present the change in average consumer welfare relative to baseline conditional on the enrollee's income level. Most of the welfare gains at 30% coinsurance are accrued by individuals in the low income tier, followed by those in the middle and high income categories. At coinsurance rates above 60%, the pattern is reversed, with consumers in the middle income level experiencing most of the welfare gains.

6.2 Counterfactual Out-of-Pocket Limits

In this subsection, I study the effects of maximum OOP expenditures on negotiated prices while holding the observed tiered coinsurance rate schedule fixed. I estimate several counterfactual scenarios where I impose uniformity in the OOP limits by setting them equal to a MMW times factors ranging from 0.5 to 1.5. Panel (a) of figure (6) depicts the resulting (quantity-weighted) average equilibrium price in the solid black line and the observed average price in the red dot. The counterfactual price is decreasing and convex with respect to the OOP limit, but prices are overall less responsive to the spending thresholds than to the coinsurance rates. The average price decreases 19% when the OOP threshold increases from 0.5×MMW to 1×MMW. Prices decrease by an additional 1% when the threshold changes from 1×MMW to 1.5×MMW. The price pattern is similar to that of the proportion of individuals that reach the spending thresholds in counterfactual, presented in panel (b) of the figure. This proportion falls by 30 percentage points when the threshold increases from 0.5×MMW to 1.5×MMW. Failure to reach the OOP limits translates into higher future out-of-pocket costs for consumers, which increases (in absolute value) the effective demand elasticity as seen in figure (7).

Panel (a) of figure (8) presents the welfare change for consumers, insurers, and hospitals under these counterfactual OOP limits relative to the full insurance scenario. Consumer surplus, insurer profits, and hospital profits are all decreasing and convex with respect to the OOP limit. Consumer welfare increases 12.5% relative to full insurance at an OOP limit of 0.5×MMW, but these welfare gains fall to 5% when the spending threshold is set to 1.5×MMW. Even though hospital profits increase almost 5% when the OOP limit equals 0.5×MMW, it presents no significant variations



(a) Average price (b) Proportion that reach the OOP limit

Figure 6: Average price and proportion that reach the OOP limit under counterfactual OOP limits

relative to the full insurance scenario at values of the spending threshold above 1×MMW.

In panel (b) of figure (8), I present average welfare changes relative to full insurance for consumers in different income categories. This figure shows that high income individuals experience the highest welfare gains at every value of the spending threshold, followed by those in the middle and low income tiers. At a value of the OOP limit equal to 0.5×MMW, consumers earning more than 5 times the MMW experience welfare gains almost twice larger than those with incomes below 2 times the MMW. This is because at low values of the spending thresholds, individuals in the high income tier, who have the highest coinsurance rate, reach the spending limits at a faster rate compared to those in the first to second income tiers. Moreover, differences in welfare between income groups remain constant with changes in the OOP threshold.

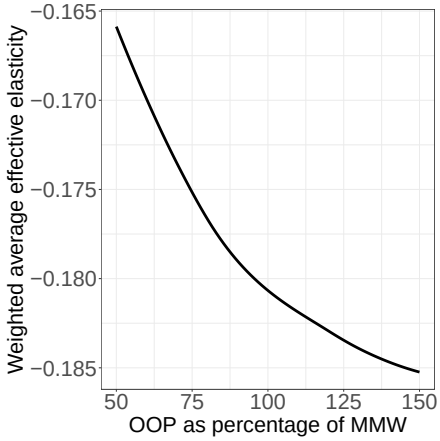
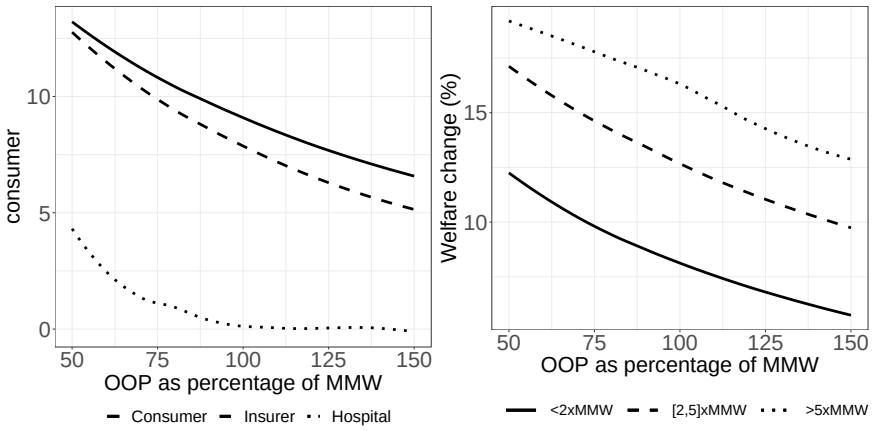


Figure 7: Average effective price elasticity under counterfactual OOP limits



(a) Total welfare changes

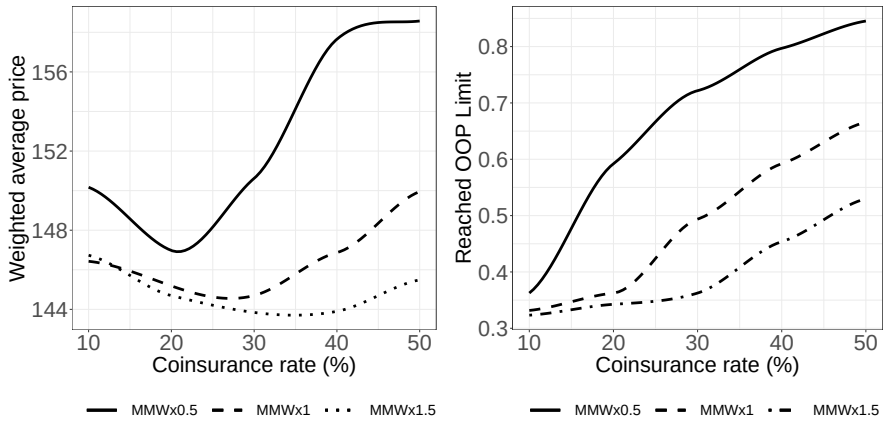
(b) Consumer welfare changes by income group

Figure 8: Welfare changes under counterfactual OOP limits

6.3 Simultaneous Changes to Coinsurance Rates and Out-of-Pocket Limits

In the last set of counterfactual analyses, I implement simultaneous changes to the coinsurance rate and the maximum OOP expenditure. I impose a uniform coinsurance rate ranging from 10 to 50% and a uniform OOP limit equal to the MMW times factors ranging from 0.5 to 1.5. Panel (a) of figure (9) shows the resulting (quantity-weighted) average counterfactual price from these exercises. The x-axis denotes the coinsurance rate, the y-axis is the average price, and each line corresponds to a different OOP limit. Results show that coinsurance rates determine variations along the price curve, while OOP limits determine overall price levels. When the OOP limit equals 0.5×MMW, a higher proportion of patients reach their expenditure thresholds as seen in panel (b) of the figure and demand is relatively more inelastic compared to the observed scenario (as seen in figure 10). As a result, equilibrium prices increase, and the higher the value of the coinsurance rate, the greater is the price increase compared to the observed cost sharing system. As the uniform OOP limit increases to 1×MMW, the proportion of patients that reach full insurance coverage at every value of the coinsurance rate decreases. Demand becomes more elastic because consumers face higher future out-of-pocket costs. So, equilibrium prices decrease following a U-shaped pattern with respect to the coinsurance rate.

Figure (11) presents the welfare change for consumers, insurers, and hospitals in each counterfactual exercise relative to the full insurance scenario. Consumer welfare and insurer profits are concave in the coinsurance rate and decrease monotonically with the OOP limit. At a spending threshold of 0.5×MMW, consumer welfare increases between 12.5% and 17.5% relative to full in-



(a) Average prices (b) Proportion that reach OOP limit

Figure 9: Average price and proportion that reach the OOP limit under counterfactual coinsurance rates and OOP limits

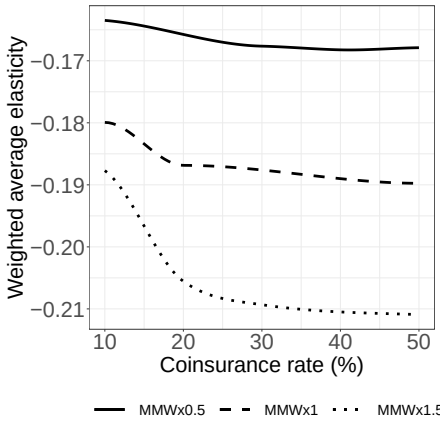


Figure 10: Average effective price elasticity under counterfactual coinsurance rates and OOP limits

surance when the coinsurance rate changes from 10% to 50%. As the spending threshold rises to 1.5×MMW, the change in consumer surplus varies between 5% and 12.5%. Similarly, insurer profits increase between 10% and 17.5% relative to the full insurance scenario when the OOP limit equals 0.5×MMW. But, at the highest value of the spending threshold, the increase in insurer profits varies between 5% and 10% with the coinsurance rate. Though hospital profits are also decreasing in the OOP limit, they are convex in the coinsurance rate as opposed to consumer welfare and insurer profits. Hospital surplus changes from -25% to over 40% relative to the full insurance scenario as the coinsurance rate increases from 10% to 50% and the spending threshold is set at 1.5×MMW. The percentage change in hospital profits varies between -20% and 0% when the OOP limit equals 0.5×MMW.

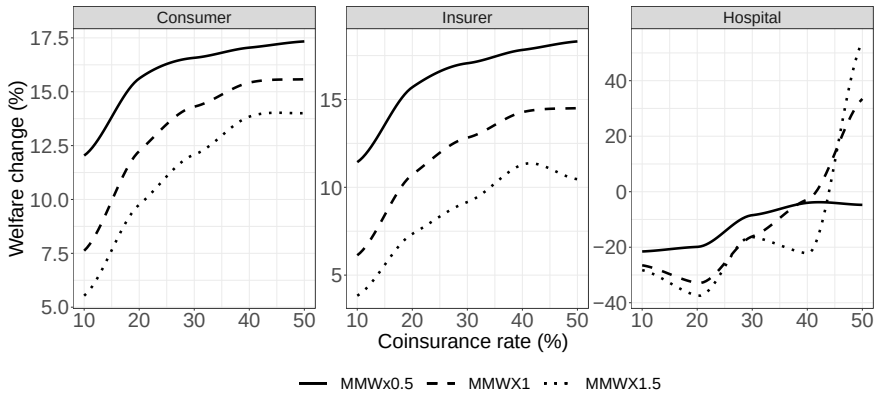


Figure 11: Total welfare changes under counterfactual coinsurance rates and OOP limits

The substantial variation in total consumer surplus with respect to the coinsurance rate and the OOP limit is coupled with significant heterogeneity across individuals in different income categories. Figure (12) shows that total welfare gains at an OOP limit of 0.5×MMW are mostly accrued by high income individuals. But, differences in welfare are decreasing with income: at a value of the spending threshold equal to 0.5×MMW, the difference in welfare between the low and middle income groups is greater than the difference between the middle and high income tiers. Consumer welfare in the high income tier increases between 15% and 20% as the coinsurance changes from 10% to 50%, while welfare in the low income tier varies between 11% and 17%. As before, the coinsurance rate determines the shape and variations along the welfare curve for all income groups, but the OOP limit determines the overall welfare level. Welfare curves for all income categories move downwards as the OOP limit rises.

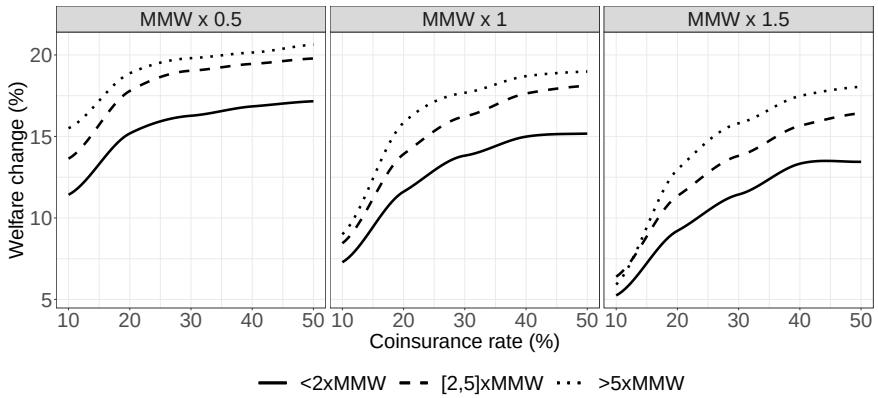


Figure 12: Consumer welfare changes by income group under counterfactual coinsurance rates and OOP limits

7 Conclusions

In this paper, I quantify the effect of counterfactual cost sharing rules (coinsurance rates and maximum out-of-pocket expenditures) on insurer-hospital negotiated prices and social welfare. My empirical setting is the Colombian healthcare system where coinsurance rates and maximum out-of-pocket amounts are indexed to the enrollee's monthly income level and determined exogenously by the government. I model the Colombian healthcare market as a two stage game. First, insurers and hospitals engage in bilateral negotiations over the price of a hospital admission following a Nash bargaining protocol. Second, consumers receive a health shock and choose a hospital in the network of their insurer to receive treatment. Insurers affect hospital demand not only through their hospital networks but also through steering claims towards certain preferred providers. I identify the effect of these steering incentives on hospital demand by using the discontinuity in coinsurance rates due to out-of-pocket maximums. Insurer steering is an additional source of price elasticity. After a patient reaches her out-of-pocket spending threshold, she faces zero prices and the insurer provides full coverage, so any sensitivity of demand to price is going to be solely explained by the insurer. This allows me to decompose the effective demand elasticity into three components: consumer sensitivity to prices, insurer steering, and insurer-hospital bargaining.

When holding OOP limits fixed, my findings show that average equilibrium prices are U-shaped with respect to the coinsurance rate, with the inflection point happening at 30% coinsurance. When holding coinsurance rates fixed, I find that prices are decreasing and convex with respect to the OOP limit. I find that consumer welfare is concave in the coinsurance rate and maximized at 30% coinsurance. Consumer welfare also decreases monotonically with the spending threshold.

The results in this paper have two implications for the design of health insurance programs: first, as noted by Einav et al. (2018) and Pauly and Blavin (2008), coinsurance rates should be higher for services or procedures with a relatively more elastic demand. This makes the case for insurance plans like value-based insurance to be potentially welfare enhancing. Second, the welfare effects of uniform cost sharing systems will depend on the relative weight of consumers and insurers in the elasticity of demand. A policymaker should also take into account the distributional welfare effects and understand how uniform cost sharing policies affect consumers' financial default risk due to healthcare expenses.

References

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

Agarwal, R., Gupta, A., and Fendrick, M. (2018). Value-Based Insurance Design Improves Medication Adherence Without An Increase in Total Health Care Spending. *Health Affairs*, 37(7):1057–1064.

Alfonso, E., Riascos, A., and Romero, M. (2013). The performance of risk adjustment models in Colombia competitive health insurance market.

Anderson, G., Frogner, B., and Reinhardt, U. (2007). Health Spending In OECD Countries in 2004: An Update. *Health Affairs*, 26(5):1481–1489.

Baicker, K., Sheppard, M., and Skinner, J. (2013). Public financing of the medicare program will make its uniform structure increasingly costly to sustain. *Health Affairs*, 32(5):882–890.

Brot-Goldberg, Z., Chandra, A., Handel, B., and Kolstad, J. (2017). What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics. *The Quaterly Journal of Economics*, 132(3):1261–1318.

Brown, T. and Robinson, J. (2015). Reference pricing with endogenous or exogenous payment limits: Impacts on insurer and consumer spending. *Health Economics*, 25(6):740–749.

Busch, S., Barry, C., Vegso, S., Sindelar, J., and Cullen, M. (2006). Effects of a cost-sharing exemption on use of preventive services at one large employer. *Health Affairs*, 25(6):1529–1536.

Capps, C., Dranove, D., and Satterthwaite, M. (2003). Competiton and market power in option demand markets. *RAND Journal of Economics*, 34(4):737–763.

Chandra, A., Gruber, J., and McKnight, R. (2010). Patient Cost-Sharing and Hospitalization Offsets in the Elderly. *American Economic Review*, 100(1):193–213.

Choudhry, N., Fischer, M., Avorn, J., Schneeweiss, S., Solomon, D., Berman, C., Jan, S., Liu, J., Lii, J., Brookhart, A., Mahoney, J., and Shrank, W. (2010). At Pitney Bowes, Value-Based Insurance Design Cut Copayments And Increased Drug Adherence. *Health Affairs*, 29(11):1995–2001.

Diaz-Campo, C. (2021). Dynamic moral hazard in nonlinear health insurance contracts. *Manuscript*.

Duggan, M. and Morton, F. (2010). The effect of medicare part d on pharmaceutical prices and utilization. *American Economic Review*, 100(1):590–607.

Einav, L., Finkelstein, A., and Polyakova, M. (2018). Patient Cost-Sharing and Hospitalization Offsets in the Elderly. *American Economic Journal: Economic Policy*, 10(3):122–153.

Einav, L., Finkelstein, A., and Shrimpf, P. (2016). Reprint of: Bunching at the kink: Implications for spending responses to health insurance contracts. *Journal of Public Economics*, 171:117–130.

Freed, M., Fuglesten, J., Damico, A., and Neuman, T. (2020). Medicare Advantage in 2021: Premiums, Cost sharing, Out-of-Pocket Limits and Supplemental Benefits. https://www.kff.org/medicare/issue-brief/medicare-advantage-in-2021-premiums-cost-sharing-out-of-pocket-limits-and-supplemental-benefits/. Last checked on Jul 20, 2021.

Ghili, S. (2018). Network Formation and Bargaining in Vertical Markets: The case of Narrow Networks in Health Insurance. *Unpublished*.

Gowrisankaran, G., Nevo, A., and Town, R. (2015). Mergers when Prices are Negotiated: Evidence from the Hospital Industry. *American Economic Review*, 105(1):172–203.

Grennan, M. (2013). Price Discrimination and Bargaining: Empirical Evidence from Medical Devices. *American Economic Review*, 103(1):145–177.

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

Ho, K. and Lee, R. (2017). Insurer Competition in Health Care Markets. *Econometrica*, 85(2):379–419.

Ho, K. and Lee, R. (2019). Equilibrium Provider Networks: Bargaining and Exclusion in Health Care Markets. *American Economic Review*, 109(2):473–522.

Ho, K. and Lee, R. (2021). Health Insurance Menu Design for Large Employers. *NBER Working Paper*, (27868).

Horn, H. and Wolinsky, A. (1988). Bilateral Monopolies and Incentives for Merger. *RAND Journal of Economics*, 19(3):408–419.

Hsu, J., Price, M., Huang, J., Brand, R., Fung, V., Hui, R., Fireman, B., Newhouse, J., and Selby, N. (2006). Unintended consequences of caps on medicare drug benefits. *New England Journal of Medicine*, 354(22):2349–2359.

Kleinke, J. (2004). Access Versus Excess: Value-Based Cost Sharing For Prescription Drugs. *Health Affairs*, 23(1):34–47.

Lamprea, E. and Garcia, J. (2016). Closing the gap between formal and material health care coverage in colombia. *Health and Human Rights*, 18(2):49–65.

Lavetti, K. and Simon, K. (2016). Strategic Formulary Design in Medicare Part D Plans. NBER Working Paper 22338.

Liebman, E. (2018). Bargaining in markets with exclusion: An analysis of health insurance networks.

Marone, V. and Sabety, A. (2021). When Should There Be Vertical Choice in Health Insurance Markets? *American Economic Review*, 112(1):304–342.

McFadden, D. (1996). Computing Willingness-to-Pay in Random Utility Models. *University of California at Berkeley, Econometrics Laboratory Software Archive, Working Papers*.

McNamara, C. and Serna, N. (2021). The Impact of a National Formulary Expansion on Diabetics.

Pauly, M. and Blavin, F. (2008). Moral hazard in insurance, value-based cost sharing, and the benefits of blissful ignorance. *Journal of Health Economics*, 27(6):1407–1417.

Prager, E. (2020). Health Care Demand Under Simple Prices: Evidence From Tiered Hospital Networks. *American Economic Journal: Applied Economics*, 14(4):196–223.

Prager, E. and Tilipman, N. (2020). Regulating Out-of-Network Hospital Payments: Disagreement Payoffs, Negotiated Prices, and Access.

Robinson, J. and Brown, T. (2013). Increases in consumer cost sharing redirect patient volumes and reduce hospital prices for orthopedic surgery. *Health Affairs*, 32(8):1392–1397.

Serna, N. (2021). Cost sharing and the demand for health services in a regulated market. *Health Economics*, pages 1–17.

Serna, N. (2022). Risk Selection Through Hospital Networks.

Shigeoka, H. (2014). The Effect of Patient Cost Sharing on Utilization, Health, and Risk Protection. *American Economic Review*, 104(7):2152–2184.

Starc, A. and Town, R. (2018). Externalities and Benefit Design in Health Insurance. NBER Working Paper 21783.

Thomson, S., Schang, L., and Chernew, M. (2013). Value-based cost sharing in the united states and elsewhere can increase patients' use of high-value goods and services. *Health Affairs*, 32(4):704–712.

Town, R. and Vistnes, G. (2001). Hospital competition in HMO networks. *Journal of Health Economics*, 20:733–753.

Trivedi, A., Rakowski, W., and Ayanian, J. (2008). Effect of cost sharing on screening mammography in medicare health plans. *New England Journal of Medicine*, 358(4):375–383.

Velez, C. (2016). *La salud en Colombia: Pasado, presente y futuro de un sistema en crisis*. Penguin Random House.

Appendix A Evidence of insurer steering

Every year, the Colombian Ministry of Health conducts a survey of enrollee satisfaction with their insurance company that is representative at the insurer level. Among others, the survey asks enrollees whether their insurer has denied any service in last year, and the reasons why the insurer denied the claim. Using survey responses for these questions from 2013 to 2016, for every insurer I plot the percentage of enrollees that responded affirmatively to the question of any service denial in figure (A1). The results here are suggestive of insurer steering being common and playing an important role in the Colombian health insurance market. Then, conditional on a service being denied, I plot the percentage of enrollees who indicated that their insurer denied the claim because the provider was out-of-network in panel (a) of figure (A2), or because the service was not included in the national plan in panel (b) of the figure. The results in figure (A2) indicate that narrow networks is one of the main reasons of dissatisfaction with an insurance company, and that consumers are generally inattentive about the services included in the national plan, even though these are public.

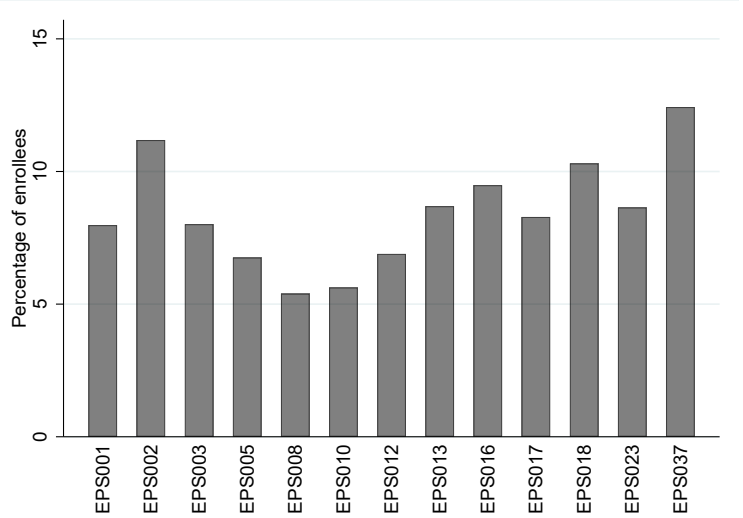
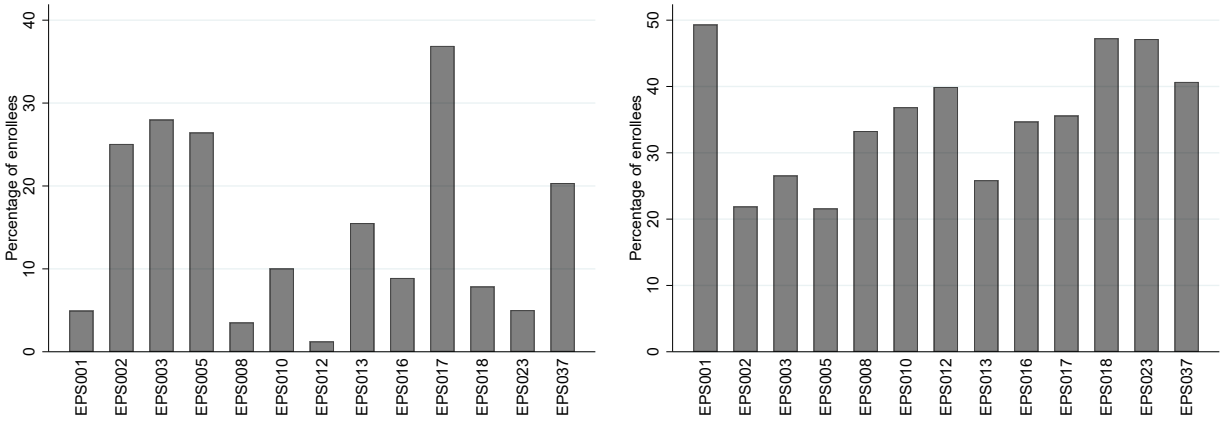


Figure A1: Percentage of enrollees that report service denials



(a) Out-of-network provider

(b) Service not in plan

Figure A2: Reason for service denials

Appendix B Matrices of elasticity decomposition

In this appendix, I present the expressions for the matrices describing the elasticity decomposition. Consumer sensitivity to prices is captured by:

$$\Omega^i = \left\{ \begin{array}{cc} \sum_{i \in N_j} \alpha_i^0 \kappa_i s_{ijh}(1 - s_{ijk}) & \text{if} \quad j = k \\ \sum_{i \in N_j} \alpha_i^0 \kappa_i s_{ijh} s_{ijk} & \text{if} \quad j \neq k \end{array} \right\}$$

Insurer steering is given by:

$$\Omega^j = \left\{ \begin{array}{cc} \sum_{i \in N_j} \alpha_i^1 (1 - \kappa_i) s_{ijh}(1 - s_{ijk}) & \text{if} \quad j = k \\ \sum_{i \in N_j} \alpha_i^1 (1 - \kappa_i) s_{ijh} s_{ijk} & \text{if} \quad j \neq k \end{array} \right\}$$

And the bargaining protocol enters the elasticity of demand through the following matrix:

$$\Upsilon = \left(\frac{\beta}{(1-\beta)D} \right) \sum_{i \in N_j} \frac{\tau}{|\alpha_i^1|} s_{ijh}(\alpha_i^0 \kappa_i + \alpha_i^1 (1 - \kappa_i)) - \sum_{i \in N_j} \sum_{h \in \mathcal{H}_j} (1 - \kappa_i) s_{ijh}$$
$$+ \sum_{i \in N_j} \sum_{h \in \mathcal{H}_j} \left\{ \begin{array}{cc} (1 - \kappa_i) p_{jh} s_{ijh}(1 - s_{ijk})(\alpha_i^0 \kappa_i + \alpha_i^1 (1 - \kappa_i)) & \text{if} \quad j = k \\ (1 - \kappa_i) p_{jh} s_{ijh} s_{ijk}(\alpha_i^0 \kappa_i + \alpha_i^1 (1 - \kappa_i)) & \text{if} \quad j \neq k \end{array} \right\}$$

Table C1: Robustness checks on hospital demand estimation

		Main	Alternative network	
			w/ ICU	w/ Rooms
$\kappa_i p_{jh}$		-17.973***	-15.947***	-18.951***
		(2.202)	(2.222)	(2.244)
$(1-\kappa_i)p_{jh}$		-2.070***	-2.155***	-2.034***
		(0.313)	(0.308)	(0.321)
Previous provider		2.650***	2.610***	2.647***
		(0.023)	(0.025)	(0.024)
Interactions				
$\kappa_i p_{jh}$	Male	1.668	2.642**	1.720
		(1.332)	(1.239)	(1.347)
	Age>=65	10.377***	6.187***	10.630***
		(1.412)	(1.254)	(1.425)
	Cancer	0.742	2.979	0.712
		(3.332)	(3.436)	(3.371)
	Cardio.	2.713	4.603*	2.957
		(2.634)	(2.472)	(2.682)
	Diabetes	-4.028	8.150	-3.579
		(8.313)	(9.805)	(8.329)
	Renal	18.006***	13.615**	17.173**
		(6.754)	(6.150)	(6.718)
	Other	9.944**	10.129**	10.427**
		(4.086)	(4.339)	(4.163)
	>=2 diagnoses	5.834***	7.143***	6.155***
		(2.244)	(2.193)	(0.479)
$(1-\kappa_i)p_{jh}$	Male	0.503***	0.405***	0.152
		(0.150)	(0.137)	(0.109)
	Age>=65	0.029	-0.160	0.030
		(0.154)	(0.138)	(0.154)
	Cancer	1.187***	0.825**	1.230***
		(0.389)	(0.399)	(0.393)
	Cardio.	1.317***	1.202***	1.337***
		(0.346)	(0.323)	(0.354)
	Diabetes	0.070	-1.220	0.061
		(0.774)	(1.076)	(0.784)
	Renal	-1.418	-0.619	-1.216
		(1.034)	(0.951)	(1.028)
	Other	-0.837	-0.698	-0.853
		(0.566)	(0.615)	(0.579)
	>=2 diagnoses	1.188***	1.120***	1.223***
		(0.300)	(0.291)	(0.307)
N		1,066,982	796,879	1,031,916
Pseudo-R^2		0.22	0.20	0.22

Note: Robustness checks on hospital demand with alternative network measures. All models include all the interaction terms between coinsurance payments, steering, and hospital characteristics with patient demographics and diagnoses. The models also include hospital fixed effects, normalizing the largest hospital in each market to zero. Robust standard errors in parenthesis. ***p<0.01,**p<0.05,*p<0.1

Appendix C Robustness checks on demand

I provide robustness checks on my demand model in this appendix. Table (C1) presents the main demand results and robustness checks on alternative network definitions. I construct the consumer's hospital choice set in a market by considering only hospitals that have ICU beds, and by considering hospitals with at least one room. The main specification uses all hospitals with the Ministry's certification for provision of inpatient services. All the models include the same variables as my main specification but only the average effects of coinsurance payments, insurer steering, and their interactions with consumer demographics and diagnoses are reported for exposition.

Table C2: Star hospital demand

		Coefficient	Std. Error
$\kappa_i p_{jh}$		-3.741	(3.948)
$(1-\kappa_i)p_{jh}$		-1.973***	(0.565)
Previous provider		3.144***	(0.029)
Interactions			
$\kappa_i p_{jh}$	Male	0.681	(2.358)
	Age>=65	0.748	(2.538)
	Cancer	-1.938	(5.500)
	Cardio.	10.231**	(4.867)
	Diabetes	9.489	(13.648)
	Renal	14.170	(8.938)
	Other	8.468	(6.574)
	>=2 diagnoses	6.076	(4.002)
	Healthy	(ref)	(ref)
$(1-\kappa_i)p_{jh}$	Male	0.281	(0.278)
	Age>=65	-1.689***	(0.285)
	Cancer	1.596**	(0.654)
	Cardio.	-1.049	(0.662)
	Diabetes	-2.942	(2.061)
	Renal	-2.079	(1.455)
	Other	-0.832	(0.827)
	>=2 diagnoses	0.118	(0.561)
	Healthy	(ref)	(ref)
N		532,170	
Pseudo-R^2		0.16	

Note: Hospital demand on subsample of plans that have negotiated below median prices with star hospitals. The model includes all the interaction terms from the main specification. Robust standard errors reported. ***p<0.01,**p<0.05,*p<0.1

In table (C2) I estimate hospital demand in the subsample of insurers that have negotiated below median prices with star hospitals. I compute the median price across all insurer-hospital pairs in each market. The table includes all the variables of my main specification, but reports only the main coefficients for exposition. This robustness check gets at the potential endogeneity problem arising from unobserved consumer preferences. If selection on unobserved preferences is not

an issue, then the choice probabilities for star hospitals predicted after manually increasing prices using the estimates of table (C2), should be similar to the choice probabilities of star hospitals in the full sample that have negotiated above median prices with insurers. In table (C3), using the Kolmogorov-Smirnov (K-S) statistic, I compare the distribution of average choice probabilities for star hospitals that have above median prices, with three predicted distributions using the sample and estimates in (C2): (i) the distribution resulting from setting prices equal to observed prices, (ii) the distribution resulting from setting prices equal to observed prices plus one standard deviation, and (iii) the distribution obtained from setting prices equal to observed prices plus two standard deviations. Table (C3) reports the K-S statistic and p-value associated to the null hypothesis of equality of distributions.

Table C3: Test of equality of distribution of choice probabilities for star hospitals

Price	K-S statistic	p-value
Observed	0.192	0.096
Observed + 1 sd	0.181	0.133
Observed + 2 sd	0.171	0.180

Appendix D Probability of a hospital admission

Table (D1) shows the mean and standard deviation of the probability of a hospital admission $\gamma^a_{j(i)l(i)}$ for subsamples of patients. Table (D2) reports summary statistics of this probability separately for each insurer.

Table D1: Summary of probability of hospital admission by demographics

Demographic	Mean	SD
Female	0.135	0.212
Male	0.145	0.221
Age<65	0.133	0.214
Age>=65	0.156	0.220
Healthy	0.031	0.083
Chronic	0.184	0.237

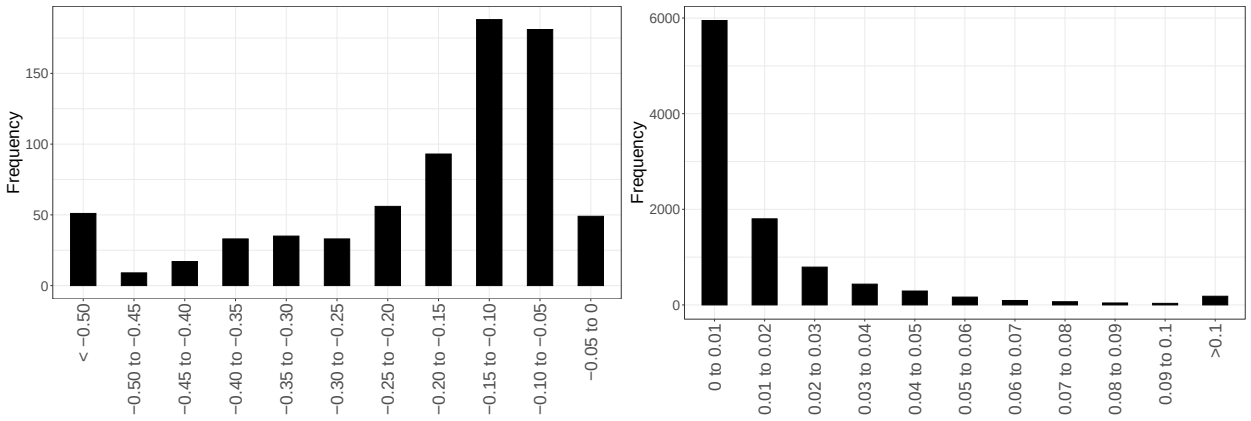
Table D2: Summary of probability of hospital admission by insurer

Insurer	Mean	SD
EPS001	0.100	0.198
EPS002	0.084	0.153
EPS003	0.281	0.267
EPS005	0.150	0.237
EPS008	0.134	0.239
EPS009	0.030	0.113
EPS012	0.047	0.130
EPS013	0.299	0.229
EPS016	0.104	0.145
EPS017	0.121	0.232
EPS018	0.078	0.179
EPS023	0.293	0.305
EPS037	0.128	0.181

Appendix E Description of demand elasticities

Figure (E1) shows the distribution of price elasticities before bargaining takes place, $\Omega^i + \Omega^j$. Panel (a) shows the distribution of own-price elasticities, interpreted as the percentage change in hospital demand following a 1% increase in its base price. Results show that 6.8% of the estimated elasticities are greater than -0.5 (in absolute value). The average own-price elasticity is -0.197 and its standard deviation equals 0.173. Panel (b) presents the distribution of cross-price elasticities or the percentage change in hospital demand when the price of other hospitals in their network increases by 1%. Almost all of these elasticities are concentrated between 0 and 0.02. The average cross-price elasticity is 0.015 and its standard deviation equals 0.028. 1.8% of the estimated cross-price elasticities are greater than 0.1, so in general there is little substitution between hospitals.

I inspect some potential sources of elasticity heterogeneity in table (E1). The table presents the mean of own- and cross-price elasticities before bargaining. I interpret these results in absolute value. The table shows that own-price elasticities are increasing with the coinsurance rate. This pattern is consistent with Einav et al. (2018) who find that private insurers set higher coinsurance rates for drugs with more elastic demand. The actual own-price elasticity goes from an average of -0.159 to -0.308 when we move from the low to the high cost sharing tier. Demand elasticities also vary significantly across patient and hospital characteristics. The average own-price elasticity is decreasing with age, and higher for healthy patients than for individuals with any chronic disease. Hospitals with more rooms have a more elastic demand compared to smaller hospitals. Cross-price elasticities exhibit similar patterns as own-price elasticities.



(a) Own-price elasticities

(b) Cross-price elasticities

Figure E1: Distribution of price elasticities

To rationalize the relation between coinsurance rates and demand elasticities, suppose insurers have discretion over coinsurance rates and these rates are continuous. By the envelope theorem, the partial derivative of the maximal joint surplus with respect to the coinsurance rate is given by:

$$\frac{\partial S^*_{jh}}{\partial \kappa_i} = \underbrace{\frac{\beta_j}{\pi_j(\mathcal{H}_j, p^*_{jh}) - \pi_j(\mathcal{H}_j \backslash h, p^*_{jh})} \left[- \sum_{i \in N_j} \sum_{h \in \mathcal{H}_j} ((1-\kappa_i)\Omega - 1) p^*_{jh} s_{ijh} - \frac{\tau(\alpha_0 - \alpha_1)}{|\alpha_1|} \sum_{i \in N_j} p^*_{jh} s_{ijh} \right]}_{\text{Change in insurer profits}}$$
$$+ \underbrace{\frac{(1-\beta_j)}{\pi_h(\mathcal{J}_h, p^*_{jh})} \left[\sum_{j \in \mathcal{J}_h} \sum_{i \in N_j} (p^*_{jh} - mc_{jh}) s_{ijh} \Omega \right]}_{\text{Change in hospital profits}} \tag{14}$$

where Ω is the matrix of elasticities. The first and second terms on the right-hand side of the equation represent the change in maximal insurer profits and the change maximal hospital profits due to a change in coinsurance rates, respectively. These expressions show that an increase in coinsurance rates has three effects: (i) decreases the marginal cost of coverage, (ii) decreases patient willingness-to-pay for insurers, and (iii) decreases hospital demand. If all hospitals are identical and Ω is a diagonal matrix of own-price elasticities, then the more elastic is the demand for a particular hospital, the larger is the first term in brackets for the change in insurer profits and, thus, the more positive is the change in joint surplus. This suggests that in a Nash surplus maximizing cost sharing scheme, coinsurance rates should be higher for groups of patients or hospitals with a relatively elastic demand.

Table E1: Conditional price elasticities

		Own-price	Cross-price
Cost sharing tier	Low	-0.159	0.017
	Medium	-0.220	0.022
	High	-0.308	0.029
Age	19-44	-0.290	0.029
	45-64	-0.208	0.022
	>=65	-0.091	0.010
Diagnoses	Chronic	-0.151	0.016
	Healthy	-0.434	0.041
Beds	<100	-0.165	0.018
	100-200	-0.163	0.021
	>200	-0.188	0.018
Rooms	<5	-0.153	0.020
	5-8	-0.169	0.021
	>8	-0.200	0.015

Note: This table presents the average own- and cross-price elasticities before bargaining takes place, conditional on cost sharing tiers, patient characteristics, and hospital characteristics.