Designing Cost Sharing for Regulated Health Insurance: Effects on Prices and Consumers

Natalia Serna

University of Wisconsin - Madison

(Draft, do not cite)

June 29, 2021

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 from consumer sensitivity to prices. I find that consumers and insurers account for 43% and 47% of the effective demand elasticity, respectively, while the bargaining protocol accounts for only 11%. Equilibrium prices are U-shaped with respect to the coinsurance rate, but they do not respond significantly to changes in the out-of-pocket limits. Results show that consumer surplus is directly proportional to the number of patients that reach the out-of-pocket limit under counterfactual cost sharing rules, but the distributional welfare effects matter for policy making.

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

codes: I11, I13, I18, L13.

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 (Agarwal et al., 2018; Shigeoka, 2014; Serna, 2021), but less is known about its effect on supply, in particular on negotiated hospital prices. It is important to understand how service prices vary with cost sharing because if lowering coinsurance rates leads to non proportional increases in prices, then popular coinsurance designs such as value-based insurance can actually have a negative impact on patient welfare. Accordingly, the purpose of this paper is twofold. The first is investigating how counterfactual changes to the cost sharing schedule affect negotiated prices between insurers and hospitals through three possible channels: consumer sensitivity to prices, insurer sensitivity to prices, and the bargaining protocol. The second is quantifying changes in social welfare from a counterfactual cost sharing policy in which the regulator assigns a uniform coinsurance rate and a uniform maximum out-of-pocket threshold.

The main contribution of my paper is leveraging the discontinuity in coinsurance rates due to maximum expenditure 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. By decomposing the elasticity into its three portions I am able to show that, even in absence of bargaining, insurers can constrain hospital prices. My paper also contributes to the literature of optimal design of health insurance plans a recent example of which is Ho and Lee (2021), by looking at the effects of a uniform cost sharing policy on social welfare. Tackling this question has been difficult in contexts like US because insurers compete in premiums and plans. But the Colombian healthcare system, where my data comes from, gives me a unique way of measuring the welfare associated to counterfactual cost sharing schedules thanks to the strict government regulation of health insurance.

Quantifying the effect of cost sharing on negotiated prices and separating this effect into its consumer- and insurer- explained portions are not straight-forward exercises either. On the one hand, negotiated prices can have a reverse causal effect on cost sharing, for example, insurers can lower coinsurance rates for hospitals with which they have agreed on relatively low prices. This suggests that to identify the direct effect of cost sharing on prices, exogenous changes to coinsurance rates are needed; but, these are often nonexistent in contexts where insurance companies have discretion over plan design such as Health Insurance Exchanges in the US. On the other hand, insurers can respond to high hospital prices by increasing premiums. The confounding effect of a lower enrollment level due to higher premiums can lead to biased estimates of the effect of cost sharing on prices. I take advantage of Colombia's unique empirical setting to abstract away from these potential problems to identification.

Colombia's healthcare system is characterized by strong government regulation of consumption and provision of health services and health insurance. From the demand side of the market, the government determines copays, coinsurance rates, and maximum out-of-pocket (OOP) expenditures, as functions of the enrollee's

monthly income, but they are constant across hospitals, insurers, and services. From the insurance side of the market, the government determines premiums and a basic universal plan that guarantees coverage of a list of around 7,000 procedures and more than 700 medications. Similar to the US, 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. I will describe the Colombian system in more detail in section (2). The data for this paper is a panel of around 158 thousand enrollees who were admitted to the hospital between 2009 and 2011 under the national basic plan.

The strict regulation of health insurance in Colombia means that carriers have no price mechanisms to induce demand elasticity. So they resort to non-price mechanisms such as ex-post rationing of care or provision of narrow hospital networks, to steer patients to preferred providers and minimize costs. Several papers have studied insurers' use of replacement threats to achieve lower prices during bilateral negotiations (Ho and Lee, 2019; Ghili, 2018; Liebman, 2018). In this paper, I focus on the former mechanism. Even though it is forbidden by law, there is anecdotal 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 the Latin American region regarding health insurance coverage lawsuits (Lamprea and Garcia, 2016). A recent paper documents these rationing efforts in the sample of patients with diabetes following a formulary expansion in Colombia (McNamara and Serna, 2021). To separate consumer- from insurer-induced demand elasticity, it would be ideal to have data on the claims that insurance companies deny in order to estimate their sensitivity to prices and use the data on reimbursed claims to recover the consumers' side. In absence of data on rationing of care, another way to do so is to find claims where the coinsurance rate equals zero for observed reimbursed claims. I use, then, the claims by patients who have reached their maximum OOP expenditure during the year, because any evidence of sensitivity to prices in that case is going to be fully explained by the insurer. Although in this paper I am not concerned with modelling the rationing mechanism, I am able to capture its effect on demand in a reduced-form way by leveraging the non-linearity in insurance contracts introduced by these spending thresholds. Aaron-Dine et al. (2015) stress that initial healthcare utilization responds significantly to the OOP limits because patients are forward-looking or consume healthcare strategically to reach the spending thresholds, which could undermine my identification strategy. Nonetheless, I show descriptive evidence that patients in Colombia are not forward-looking but utilization still responds to the expenditure limits because of insurer sensitivity to prices.

To answer the question of how cost sharing affects prices I need a structural approach because coinsurance rates have not changed since the establishment of Colombia's universal healthcare system in 1993 and there are 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 using a Nash-in-Nash bargaining framework under the usual assumptions of fixed enrollee pools and fixed networks. Hospital demand in the

second stage is obtained from patient's discrete choices over the set of hospitals made available by their insurer.

This paper is related to a long strand of literature that investigates the effects of cost sharing on the consumption of health services (Kleinke, 2004; Busch et al., 2006; Trivedi et al., 2008; Chandra et al., 2010; Thomson et al., 2013; Choudhry et al., 2010; Brot-Goldberg et al., 2017; Serna, 2021) and its impact on welfare and prices (Robinson and Brown, 2013; Hsu et al., 2006). Several studies have found that decreasing coinsurance rates results in lower hospital and drug prices because insurers can steer patients using drug formularies (Brown and Robinson, 2015; Duggan and Morton, 2010; Starc and Town, 2018; Lavetti and Simon, 2016). Other authors show that insurers' bargaining leverage plays an important role in constraining price increases 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, and by studying the effect not only of coinsurance rates but of OOP thresholds on negotiated prices.

In the context of an exogenous system for coinsurance rates, copays, and OOP maximums, indexed to the enrollee's income in Colombia, I find that the (quantity-weighted) average effective price-elasticity of demand equals -0.15. 46.6% of this elasticity is explained by insurer steering, 42.5% by consumer sensitivity to prices, and 11.0% by the bargaining protocol. My estimated elasticity is smaller (in absolute value) than the one found by Gowrisankaran et al. (2015) in the context of the US, which suggests that the strong regulation of premiums and plans has an important effect on price sensitivity by consumers and insurers. Deregulation can increase demand elasticity at the expense of increasing consumer adverse selection. Given my estimates, I conduct three counterfactual scenarios to understand the effect of each element of cost sharing on negotiated prices and welfare. First, I impose uniform cost sharing percentages ranging from 0 to 100%, while holding the OOP maximums 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 maximum. My results show that prices are U-shaped with respect to the coinsurance rate, but they remain mostly constant under counterfactual OOP limits. Consumer surplus is almost directly proportional to the number of patients that reach their OOP spending limits under counterfactual cost sharing rules but the distribution of these welfare changes varies with the enrollee's income level.

In particular, results in the first counterfactual show that negotiated prices for hospital admissions would increase 9.6% on average relative to the observed scenario when the coinsurance rate is set to zero. At 30% coinsurance, prices fall 3.2%, but as the coinsurance rate increases beyond this point, patients are more likely to reach their spending thresholds and face zero costs of healthcare. So at 100% coinsurance, prices actually increase 3.2%. The effective price elasticity in the case of zero coinsurance can be decomposed into an 85% due to insurer steering and a 15% due to the bargaining protocol. Under the second counterfactual, my findings show that although prices remain mostly constant, the price-elasticity of demand is U-shaped with respect to the OOP limit, which suggests that supply is highly elastic in this scenario. Consumer surplus

is convex and decreasing with respect to the OOP limits as is the proportion of patients that reach the spending thresholds.

Related to the literature on optimal design of coinsurance rate schedules and health insurance, Einav et al. (2018) find that private insurers set a higher cost sharing for drugs with more elastic demand in order to trade-off adverse selection and moral hazard while increasing their bargaining leverage with drug manufacturers. I contribute to this literature by addressing the question of whether uniform cost sharing schemes are optimal in social insurance and whether there is a design of coinsurance rates that improves social welfare. This question is not only relevant for Colombia but speaks particularly to existing uniform coinsurance rate plans like Medicare Part B in the US and debates of whether a unique non-means-tested Medicare plan is optimal from the point of view of the consumer (Baicker et al., 2013). My first counterfactual analysis shows that moving from the current three-tier system to one where the coinsurance rate is unique and below the observed average coinsurance, decreases consumer surplus because prices increase and fewer patients reach the OOP spending limits, so they end up facing relatively higher OOP costs. For example, if the coinsurance rate is set to 10% for all individuals, average hospital profits would increase 5%, insurer profits would decrease 1%, and consumer surplus would decrease 5%, relative to the observed scenario. However, if the coinsurance rate is fixed at 25%, hospital surplus would decrease 11% and consumer surplus would increase 10% given that more than 60% of patients reach their OOP limit. I also show how these changes in consumer welfare are distributed across patients conditional on their income level.

I conclude by estimating a counterfactual scenario where I impose uniformity in both the coinsurance rate and the OOP limit. I find that the coinsurance rate determines movements along the price curve while the OOP limit determines the absolute price level. At low values of the OOP threshold, prices are increasing and convex with respect to the coinsurance rate, but at high values of the threshold, the price curve is decreasing and convex. Consumer surplus is always increasing and concave with respect to the coinsurance rate, but the surplus function moves downward as the OOP limit is increased.

2 Background

The Colombian healthcare market is divided into two systems called "Contributory" and "Subsidized". The contributory system is funded by the required contributions of users and it covers all formal employees, independent workers, and their beneficiaries, roughly 51% of the population. The subsidized system is 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%. There is almost universal health insurance coverage in the country, 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, known as EPSs for their Spanish acronym. Insurers receive per-capita transfers risk-adjusted by age, sex, and location, which

are determined by the government, while at the same time premiums are set to zero. Every health insurer has to provide a standard benefits package, known as POS by its Spanish acronym, also established by the regulator. The POS covers a list of more than 7,000 procedures, services, and devices, and 736 prescription medications as of 2012 (Decree 029 of 2011). Insurers can offer complementary plans where they have discretion over premiums and cost sharing rules, but individuals can only purchase them after being enrolled through the POS. To provide services covered in the national basic plan, 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 within the network and out-of-network claims are not reimbursed. In cases where the enrollee requires services not covered by the POS, the insurer can decide whether to authorize provision and up to what percentage. Although denying services included in the POS 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. This type of ex-post rationing of healthcare is a non-price steering mechanism.

The cost sharing schedule in the national basic plan is also determined by the government. Copays, coinsurance rates, and maximum OOP expenditures, are functions of the enrollee's monthly income level in the contributory system as shown in table (1) for 2010. These prices are indexed to the monthly minimum wage (MMW), are uniform across providers and services, and in percentage terms have remained fixed since the establishment of Colombia's healthcare system in 1993. Unlike countries such as US, 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 100% 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 during the year equal to 57.5% times 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% times the MMW. Finally, people earning more than 5 times the MMW, have copays of roughly 11 USD, coinsurance rates of 23% per claim, and maximum OOP expenditure of 460% times the MMW.[1] Since there have been no changes to the cost sharing rules so far, I rely on a structural approach to answer what would happen to welfare under different cost sharing arrangements. 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.

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

Table 1: Copay, coinsurance rate, and out-of-pocket maximum in the contributory system in 2010

Income level	Copay	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.

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 POS 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. I select the sample of patients who are aged 18 or older ($N = 483,916$), and who were admitted to one of the largest hospitals in the country ($N = 177,879$). I obtain the list of all institutions equipped to provide inpatient care from the Ministry's Special Registry of Healthcare Providers and drop those whose number of beds falls below the 75th percentile of the distribution of beds in each department or state.[2] My sample of hospitals accounts for 33% of admissions and for an average of 20% of annual total healthcare costs. After dropping individuals with zero income and length-of-stay greater than 20 days, the final dataset has 95,368 patients, 122,461 admissions, 12 insurers, and 182 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, but I include distance from the hospital to the centroid of the municipality where it is located as a covariate in my hospital demand model. 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 also categorize municipalities as metropolitan, adjacent, and peripheral, following the definitions of the Ministry of Health (Comisión de Regulación en Salud, 2010). I denote a department-year combination as a market because price variation across departments is more important than variation

[2]The Special Registry of Healthcare Providers can be accessed through: https://prestadores.minsalud.gov.co/habilitacion/

within departments and across municipalities. There are a total of 33 departments in the country.

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. 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 base price obtained from this methodology.

Table 2: Summary statistics of final patient sample

Variable	Full sample	Before OOP limit	After OOP limit
	(1)	(2)	(3)
Price	241.3 (192.3)	230.6 (175.3)	257.7 (214.6)
Coinsurance	18.4 (24.3)	30.4 (24.7)	—
Copay	2.0 (2.2)	2.3 (2.5)	—
Demographics			
Male (%)	46.9 (49.9)	44.6 (49.7)	50.3 (50)
Age	58.3 (18.7)	55.6 (19.1)	62.3 (17.3)
Age group (%)			
19-44	26 (43.9)	31.7 (46.5)	17.2 (37.8)
45-64	31.4 (46.4)	31.3 (46.4)	31.5 (46.5)
$\geq$65	42.6 (49.4)	36.9 (48.3)	51.3 (50)
Location (%)			
Metropolitan	62.6 (48.4)	60.1 (49)	66.3 (47.3)
Adjacent	36.1 (48)	38.7 (48.7)	32.2 (46.7)
Peripheral	1.3 (11.4)	1.2 (10.8)	1.5 (12.2)
Income group (%)			
$< 2 \times MMW$	82.8 (37.7)	76.1 (42.7)	93.2 (25.2)
$[2, 5] \times MMW$	14.0 (34.7)	19.4 (39.5)	5.8 (23.5)
$> 5 \times MMW$	3.2 (17.5)	4.6 (20.9)	1.0 (9.8)
Diagnoses (%)			
Cancer	23.8 (42.6)	20.7 (40.5)	28.5 (45.1)
Cardiovascular disease	60.8 (48.8)	52.8 (49.9)	73.0 (44.4)
Diabetes	18.5 (38.8)	15.3 (36)	23.3 (42.3)
Renal disease	14.5 (35.2)	10.2 (30.3)	21.2 (40.9)
Other	45.2 (49.8)	38.2 (48.6)	55.8 (49.7)
Hospital characteristics			
Beds	222.8 (148.3)	218.4 (147.4)	229.5 (149.4)
Rooms	11.4 (8.3)	11.1 (8.1)	11.9 (8.6)
Any ambulance (%)	33.0 (47)	36.9 (48.2)	27.0 (44.4)
Private (%)	74.0 (43.9)	72.0 (44.9)	77.0 (42.1)
Distance	18.6 (59.4)	18.6 (75.2)	18.6 (16.8)
Admissions	122,461	73,962	48,499
Patients	95,368	67,318	35,691
Hospitals	182	181	176
Insurers	12	12	12

Note: Mean and standard deviation (in parenthesis) of the main variables for the full sample in column (1), the subsample of admissions before the out-of-pocket limit in column (2), and the subsample of admissions after the expenditure limit is reached in column (3). A unit of observation in this table is a hospital admission.

Table (2) provides some summary statistics of the resulting data. The unit of observation is a hospital admission. Column (1) reports the descriptives for the full sample of admissions, column (2) for admissions by patients who have not reached their maximum OOP expenditure, and column (3) corresponds to admissions

by patients who have already reached their expenditure limit. In the full sample, 47% of admissions are associated to males, the average age is 58 years and its standard deviation is 19 years. The majority of the sample, 63%, lives in metropolitan municipalities, but have income below 2 times the MMW, roughly 83%. The average base price of a hospital admission is $241 with a standard deviation of $192. Admissions in the full sample have average coinsurance payments equal to $18 and average copays equal to $2. Cardiovascular diseases and cancer are the most prevalent conditions in the data, followed by diabetes and renal disease. Hospitals in the full sample have an average number of beds equal to 223, an average number of rooms equal to 11, and 33% of them own ambulances. 60% of admissions correspond to patients who have not reached their OOP limit and 40% to their counterparts. The average price per admission in column (3) is $27 higher than the average price in column (2), and so is the average risk-adjusted transfer from the government. In general, admissions in column (3) are associated to relatively sicker, older, and poorer patients than admissions in column (2).

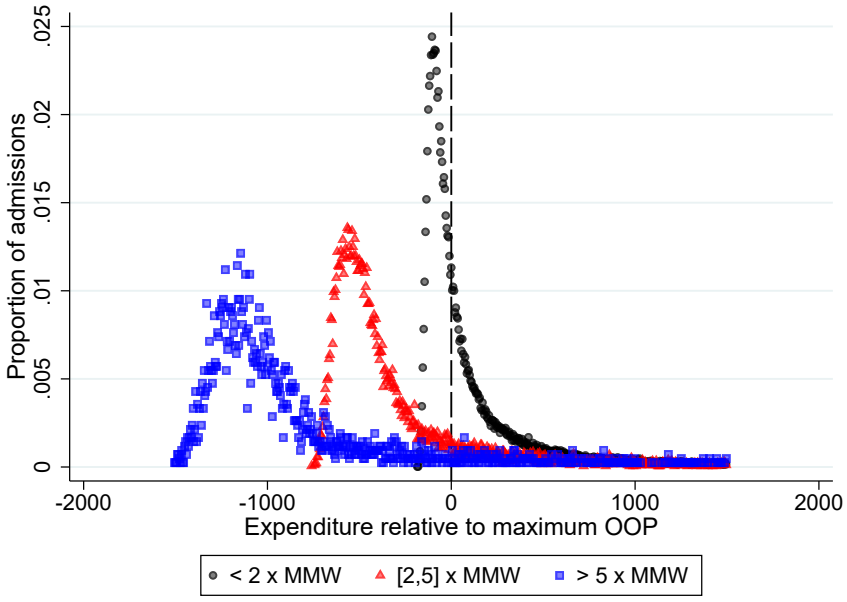


Figure 1: Proportion of admissions by level of relative out-of-pocket spending

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.

Within the set of admissions by patients in the first, second, and third income categories, 45, 16, and 12% have reached their expenditure limit, respectively. These admissions will identify the insurers' sensitivity to prices. A natural concern of my identification strategy is that individuals, particularly in the first income level, behave strategically to reach their OOP maximum, in which case both my estimates of consumer and insurer elasticity would be biased upwards. 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 income 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 $105 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 $565 before their expenditure limit is reached. While, for enrollees in the highest income category, there is no significant change in the likelihood of admission around $1400 before their expenditure limit.

Table 3: Distribution of percentage of in-network hospitals

	Before OOP Limit				After OOP Limit			
Insurer	Mean	SD	P25	P75	Mean	SD	P25	P75
EPS001	34.6	13.9	25.0	40.0	34.1	18.0	21.4	40.0
EPS002	63.6	21.1	50.0	83.3	51.6	18.4	37.5	66.7
EPS003	26.0	13.1	16.7	33.3	24.0	7.7	20.0	31.0
EPS005	65.4	18.8	50.0	75.0	52.3	23.1	35.7	66.7
EPS008	46.6	25.2	23.6	71.4	61.0	20.2	37.5	75.0
EPS009	31.7	2.4	30.0	33.3	38.5	9.0	32.1	44.8
EPS013	59.4	24.0	38.8	75.0	53.8	20.9	37.5	66.7
EPS016	67.7	21.9	50.0	83.3	67.2	20.4	57.1	75.0
EPS017	49.6	29.2	30.0	77.8	44.1	29.2	22.2	75.0
EPS018	60.2	35.2	25.0	100.0	50.7	25.5	30.0	66.7
EPS023	19.3	7.1	18.2	20.0	23.7	6.4	20.0	31.0
EPS037	82.2	18.0	66.7	100.0	76.7	19.8	60.0	100.0

Note: This table presents statistics of the distribution of percentage of in-network hospitals across markets, for each insurer and type of admission. I define the admission type relative to reaching the out-of-pocket spending limit. There are a total of 99 markets (33 departments in 3 years).

Table (3) presents several statistics of the distribution of the percentage of in-network hospitals across markets for each insurance company and type of admission (before or after reaching the spending limit). I define hospital choice sets as being specific to the type of admission, in order to identify the effect of coinsurance rates and insurer steering on demand. The main identifying variation is in choice sets across patients in the same income category and across admissions. The table shows considerable heterogeneity in network breadth along the relevant dimensions. For patients that haven't reached the spending limit, EPS037 covers an average of 82% of hospitals in a market, followed by EPS005 with 65%, and EPS016 with 68%. EPS023 has the least generous network, covering an average of 19% of hospitals. For admissions by patients that have reached the spending limit, EPS037 provides access to 77% of hospitals on average, followed by EPS016 with 67%, and EPS008 with 61%. Because networks are not complete, I take them as fixed before insurers and hospitals engage in bilateral negotiations á la Nash-in-Nash. 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, I believe

taking them as fixed is a natural assumption for my context.

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 how prices are formed and a model of hospital demand for two reasons: first, since the establishment of Colombia's healthcare system there have been no changes to the coinsurance percentages to allow identification of the impact on prices from a reduced-form perspective and, second, the endogenous responses of insurers and hospitals to the cost sharing rules requires modelling their bargaining protocol. The timing of the model 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 as a measure of demand. 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 do not incorporate a stage of insurer design of plans, because my data consists of a sample of claims made through the national basic plan determined by the regulator. 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:

$$v_{ijh} = \underbrace{\alpha_0 \kappa_i p_{jh} + \sum_{k=1}^{K} \sum_{l=1}^{L} \beta_{lk} x_{il} g_{hk} + \eta_h}_{\bar{v}_{ijh}} + \alpha_1 (1 - \kappa_i) p_{jh} + \varepsilon_{ijh} \tag{1}$$

where $\tilde{v}$ denotes the part of utility due to patients, κ_i is the coinsurance rate for patient i, $\mathbf{x}_i$ are patient observable characteristics, $\mathbf{g}_h$ are hospital observable characteristics, and η_h is a hospital fixed-effect. All these variables also vary across markets, but for notational simplicity I drop the market subscript. I assume this decision is observed by the econometrician up to an error term ε_{ijh} that follows an extreme value type-I distribution. The insurance company observes ε simultaneously with the patient but after bargaining has taken place, so in the bargaining stage the insurer is unable to condition on these shocks when deciding on prices with hospitals.

Identification. α_0 is identified from the variation in hospital choice sets across patients with the same coinsurance rate before hitting their OOP maximum. α_1 is identified from the variation in choice sets across patients that have reached their expenditure limit and face zero coinsurance. $\beta_{lk}g_{hk}$ represents the marginal utility of patients with traits x_{il} for hospitals with characteristics g_{hk}. Because I focus on acute care hospital admissions, there is no need of an outside option, so to identify the effect of patient characteristics on hospital choice they need to be interacted with hospital characteristics.

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}:

$$s_{ijh} = \frac{exp(\delta_{ijh})}{\sum_{k \in \mathcal{H}_j} exp(\delta_{ijh})} \tag{2}$$

Here $\delta_{ijh} = \tilde{v}_{ijh} + \alpha_1(1 - \kappa_i)p_{jh}$ and $\mathcal{H}_j$ is the set of hospitals in the network of insurer j. Using the estimated coefficients from (1) and following McFadden (1996), the consumer's expected utility for the set of hospitals in the network of insurer j is:

$$W_{ij}(\mathcal{H}_j, p_{jh}, \kappa_i, \mathbf{g}_h) = \log\left(\sum_{h \in \mathcal{H}_j} exp(\delta_{ijh})\right) \tag{3}$$

As in Capps et al. (2003); Town and Vistnes (2001); 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 base price of an admission in order 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:

$$\pi_j(\mathcal{H}_j, \{\mathbf{p}_{jh}\}_{h \in \mathcal{H}_j}, \kappa_i) = \frac{\tau_j}{|\alpha_1|} \sum_{i \in N_j} W_{ij}(\mathcal{H}_j, \mathbf{p}_{jh}, \kappa_i, \mathbf{g}_h) - \sum_{i \in N_j} \sum_{h \in \mathcal{H}_j} (1 - \kappa_i) p_{jh} s_{ijh}(\mathcal{H}_j, \mathbf{p}_{jh}, \kappa_i, \mathbf{g}_h) \tag{4}$$

The insurer takes into account patient welfare weighted by an altruism parameter τ and incurs in the cost of treatment net of the coinsurance payments by the patient. Insurers take N_j as fixed because they can not choose premiums and they 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 insurer and patient incentives are perfectly aligned, while $\tau < 1$ means that insurers care more about financial profits than consumer welfare. The third term at the right-hand side of equation (4) is divided by $|\alpha_1|$ to convert consumer expected utility into dollars as perceived by the insurer.

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

$$\pi_h(\mathcal{J}_h, \{\mathbf{p}_{jh}\}_{j\in\mathcal{J}_h}) = \sum_{j\in\mathcal{J}_h}\sum_{i\in N_j}(p_{jh} - mc_{jh})s_{ijh}(\mathcal{H}_j, \mathbf{p}_{jh}, \kappa_i, \mathbf{g}_h) \tag{5}$$

In this equation, $\mathcal{J}_h$ is the set of insurers that include h in their network, mc_{jh} is the marginal cost of admission at hospital h from a patient enrolled with insurer j, and hospital demand is $q_{jh} = \sum_{i\in N_j} s_{ijh}$. The marginal cost varies across hospitals because of differences in technology, physical capital, and human capital. It also varies across insurers within a hospital because of differences in administrative processes and unobserved risk selection by insurance carriers that is independent of the choice of hospital. I assume marginal costs are constant to avoid interactions between the pricing rules of different insurer-hospital pairs. Therefore, marginal costs can be modelled as a function of exogenous variables $\mathbf{v}$ and an econometric random shock ψ_{jh} that is specific to the pair.

$$mc_{jh} = \lambda \mathbf{v}_h + \psi_{jh} \tag{6}$$

I assume insurers and hospitals follow a Nash bargaining protocol to determine prices. I follow Horn and Wolinsky (1988) for defining the Nash-in-Nash equilibrium of the bargaining game, so each pair of agents chooses contract prices conditional on the equilibrium choices of every other pair. The joint surplus of a pair is a Cobb-Douglas function where the exponents represent the bargaining power of each agent:

$$\begin{aligned} S_{jh}(p_{jh_{h\in\mathcal{H}_j}}|\mathbf{p}_{jk_{k\in\mathcal{H}_j\setminus h}}) = &\left(\pi_j(\mathcal{H}_j, \{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j}, \kappa_i) - \pi_j(\mathcal{H}_j\setminus h, \{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j\setminus h}, \kappa_i)\right)^{\beta_j} \\ &\times \left(\pi_h(\mathcal{J}_h, \{\mathbf{p}_{jh}\}_{j\in\mathcal{J}_h}) - 0\right)^{1-\beta_j} \end{aligned} \tag{7}$$

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 $\pi_j(\mathcal{H}_j\setminus h, \{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j\setminus h}, \kappa_i)$. This outside option is exogenous because I assume there are no externalities between hospitals nor insurers. 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.

The problem of an insurer-hospital pair is:

$$\begin{aligned} &\max_{p_{jh}} \; S_{jh}(p_{jh_{h\in\mathcal{H}_j}}|\mathbf{p}_{jk_{k\in\mathcal{H}_j\setminus h}}) \\ s.t \;\; &\pi_j(\mathcal{H}_j,\{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j},\kappa_i) - \pi_j(\mathcal{H}_j\setminus h,\{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j\setminus h},\kappa_i) \geq 0 \\ &\pi_h(\mathcal{J}_h,\{\mathbf{p}_{jh}\}_{j\in\mathcal{J}_h}) \geq 0 \\ &\pi_j(\mathcal{H}_j,\{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j},\kappa_i) \geq 0 \end{aligned} \tag{8}$$

Let $\Upsilon = \frac{\beta}{1-\beta}\frac{C}{D}q$ with $C = \frac{\partial \pi_j(\cdot)}{\partial p_{jh}}$ and $D = \pi_j(\cdot,\mathcal{H}_j) - \pi_j(\cdot,\mathcal{H}_j\setminus h)$. Maximizing the joint surplus function with respect to prices and rearranging terms gives the following expression for equilibrium prices in matrix notation:

$$\mathbf{p} = \mathbf{mc} - \Big(\Omega^i + \Omega^j + \Upsilon\Big)^{-1}\mathbf{q} \tag{9}$$

I denote the term $\Omega^i + \Omega^j + \Upsilon$ as the "effective" price-elasticity. Ω^i is the part of the effective elasticity that is explained by consumer sensitivity to prices, Ω^j by insurer steering, and Υ captures the effect of the bargaining protocol. The latter matrix is generally negative semidefinite, which means that with bargaining, markups are smaller than if hospitals were competing á la Nash-Bertrand. The expression for these matrices is provided below:

$$\Omega^i = \left\{ \begin{array}{cc} \alpha_0 \sum_{i\in N_j} \kappa_i s_{ijh}(1-s_{ijk}) & \text{if} \quad j=k \\ \alpha_0 \sum_{i\in N_j} \kappa_i s_{ijh} s_{ijk} & \text{if} \quad j\neq k \end{array} \right\}$$

$$\Omega^j = \left\{ \begin{array}{cc} \alpha_1 \sum_{i\in N_j} (1-\kappa_i) s_{ijh}(1-s_{ijk}) & \text{if} \quad j=k \\ \alpha_1 \sum_{i\in N_j} (1-\kappa_i) s_{ijh} s_{ijk} & \text{if} \quad j\neq k \end{array} \right\}$$

$$\begin{aligned} \Upsilon = &\left(\frac{\beta}{(1-\beta)D}\right)\frac{\tau_j}{|\alpha_1|}\sum_{i\in N_j} s_{ijh}(\alpha_0\kappa_i + \alpha_1(1-\kappa_i)) - \sum_{i\in N_j}\sum_{h\in\mathcal{H}_j}(1-\kappa_i)s_{ijh} \\ &+ \sum_{i\in N_j}\sum_{h\in\mathcal{H}_j}\left\{ \begin{array}{cc} (1-\kappa_i)p_{jh}s_{ijh}(1-s_{ijk})(\alpha_0\kappa_i+\alpha_1(1-\kappa_i)) & \text{if} \quad j=k \\ (1-\kappa_i)p_{jh}s_{ijh}s_{ijk}(\alpha_0\kappa_i+\alpha_1(1-\kappa_i)) & \text{if} \quad j\neq k \end{array} \right\} \end{aligned}$$

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. Both of these are inconsistent with the empirical setting since the former implies that insurers absorb all the surplus from hospitals, while the latter suggests that zero prices can be optimal. But the strict concavity of the surplus function gives an interior solution in prices.

Identification. Equation (9) denotes the inverse mapping from the model to the primitive, mc_{jh}. After

recovering the marginal costs, the parameters in λ 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. 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. I also include the exogenous variables from $\mathbf{v}$ in the set of instruments. τ is identified from variation in profits across insurers and markets. The non-linear parameters (τ and β_j) are estimated using 2-step GMM under the assumption that $E[\psi|Z] = 0$, where Z is the matrix of instruments.

5 Estimates

5.1 Demand for hospitals

Table (4) shows the results of the hospital choice model. This stage is estimated via maximum likelihood using a conditional logit specification with hospital fixed effects. In my fixed effects I normalize the largest hospital in each market to zero. I find that if a hospital increases the base price of an admission by 1 USD, the probability of a patient choosing it decreases by 0.6% and the probability of an insurer authorizing the claim decreases by 0.06%. From the interactions we can see that males have stronger preferences for private hospitals with ambulances compared to females. Age is an important predictor of hospital choice, with older patients having stronger preferences for beds rather rooms, relative to younger patients. Results show that patients with cancer prefer private hospitals with a larger number rooms, while those with cardiovascular diseases prefer hospitals with more beds. Individuals with renal diseases do not have strong preferences for private hospitals but the larger the number of rooms the more likely it is for them to choose the hospital. The interaction between distance from the hospital to the municipality centroid and the indicator for whether the admission happens after reaching the OOP limit, suggests that insurer steering can make patients travel longer distances to seek care.

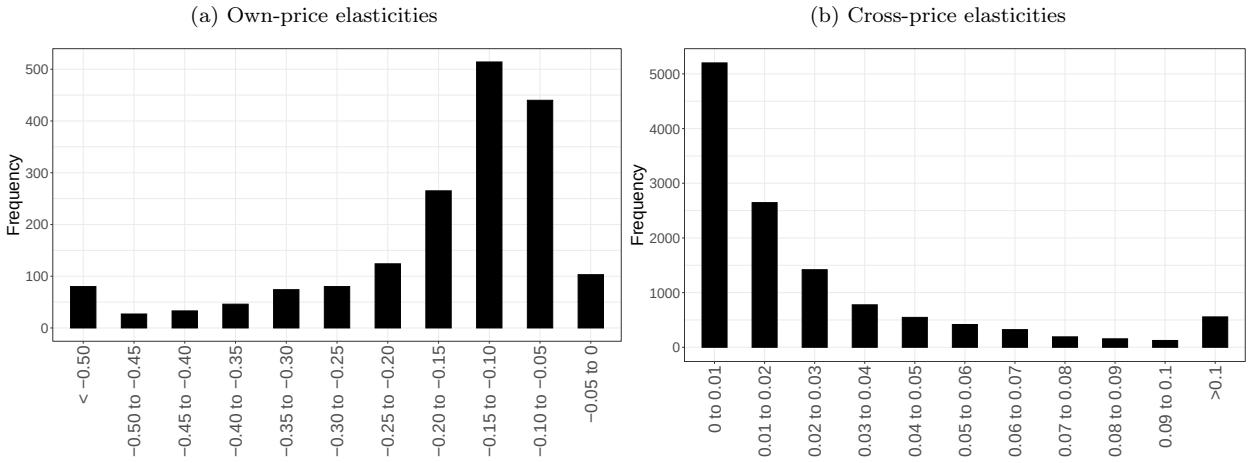
To better interpret demand results, figure (2) shows the distribution of price elasticities $\Omega^i + \Omega^j$. Panel (a) shows the distribution of own-price elasticities, interpreted as the percentage change in hospital demand following a 1% increase in its base price. Results show that 4.4% of the estimated elasticities are greater than 0.5 (in absolute value). The average own-price elasticity is -0.182 and its standard deviation equals 0.163. 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 0.01. The average cross-price elasticity is 0.026 and its standard deviation equals 0.042. 4.5% of the estimated cross-price elasticities are greater than 0.1, so in general there is little substitution between hospitals.

Table 4: Hospital demand

		Coefficient	SE
$\kappa_i p_{jht}$		-6.158***	(0.408)
$(1-\kappa_i)p_{jht}$		-0.583***	(0.040)
Beds ×	Age	0.004***	(0.000)
	Income	-0.099***	(0.006)
	Cancer	-0.086***	(0.007)
	Cardiovascular	0.044***	(0.007)
	Renal	0.038***	(0.009)
	Other diagnosis	0.030***	(0.005)
Rooms ×	Age	-0.086***	(0.004)
	Income	0.940***	(0.125)
	Cancer	3.344***	(0.137)
	Diabetes	-0.768***	(0.135)
	Cardiovascular	-1.702***	(0.146)
	Renal	1.608***	(0.184)
	Steering	1.983***	(0.099)
Any ambulances ×	Age	-0.003***	(0.001)
	Male	0.108***	(0.022)
	Income	-0.112***	(0.025)
	Cancer	-0.087***	(0.026)
	Cardiovascular	0.067***	(0.025)
	Renal	-0.039	(0.027)
	Other diagnosis	0.058**	(0.023)
Private ×	Age	-0.011***	(0.001)
	Male	0.048*	(0.025)
	Income	0.321***	(0.040)
	Cancer	0.088***	(0.030)
	Diabetes	-0.095***	(0.027)
	Cardiovascular	0.110***	(0.029)
	Other diagnosis	0.137***	(0.026)
Distance ×	Steering	0.009***	(0.002)
Observations		1,141,532	
Pseudo R^2		0.18	

Note: Maximum likelihood estimation of joint patient demand and insurer demand for hospitals. Includes hospital fixed effects, normalizing the largest hospital in each market to zero. Prices are measured in thousands of USD. The "distance" variable measures the distance in kilometers from the hospitals to the centroid of the municipality. The number of beds and rooms are measured in hundreds. Robust standard errors in parenthesis. ***p<0.01,**p<0.05,*p<0.1

Figure 2: Distribution of price elasticities



I inspect some potential sources of heterogeneity in elasticities in table (5). The table presents the mean, standard deviation, 25th and 75th percentiles of the own- and cross-price elasticity across admissions that happen before reaching the expenditure limit. The table shows that the average and standard deviation of own-price elasticities are increasing (in absolute value) with income, and thus 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 own-price elasticity goes from an average of -0.168 to -0.265 when we move from the low to the high income level. Demand elasticities also vary significantly across hospital characteristics, but not across other patient characteristics like location, age, or diagnoses. The average own-price elasticity increases (in absolute value) from -0.157 to -0.248 and the average cross-price elasticity from 0.019 to 0.040, as the number of beds goes from less than 125 to more than 250. Similar patterns are observed when we move from hospitals with less than 7.5 rooms to hospitals with more than 15 rooms.

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

$$\frac{\partial S^*_{jh}}{\partial \kappa_i} = \frac{\beta_j}{\pi_j(\mathcal{H}_j, p^*_{jh}) - \pi_j(\mathcal{H}_j \backslash h, p^*_{jh})} \left[- \sum_{i \in N_j} \sum_{h \in \mathcal{H}_j} ((1-\kappa_i)\Omega - 1) p^*_{jh} s_{ijh} - \frac{\tau(\alpha_0 - \alpha_1)}{|\alpha_1|} \sum_{i \in N_j} p^*_{jh} s_{ijh} \right]$$
$$+ \frac{(1-\beta_j)}{\pi_h(\mathcal{J}_h, p^*_{jh})} \left[\sum_{j \in \mathcal{J}_h} \sum_{i \in N_j} (p^*_{jh} - mc_{jh}) s_{ijh} \Omega \right] \tag{10}$$

where Ω is the matrix of elasticities. The first and second terms at the right-hand side of the equation represent the change in maximal insurer profits and maximal hospital profits due to a change in coinsurance rates, respectively. These expressions are very intuitive: if coinsurance rates are increased, maximal insurer profits increase due to a lower marginal cost of coverage and decrease due to a lower patient willingness-to-

pay, while maximal hospital profits always decrease due to a lower demand. Now suppose all hospitals are identical. In that case, Ω is a diagonal matrix of own-price elasticities with zeros off-diagonal. The more elastic is demand for a particular hospital, the larger is the first term in brackets in the expression for the change in insurer profits, therefore the more positive is the change in surplus. This shows that in an optimal cost sharing scheme, coinsurance rates should be higher for groups of patients or hospitals with a relatively elastic demand.

Table 5: Conditional price elasticities

	Own-price elasticities				Cross-price elasticities			
	Mean	SD	P25	P75	Mean	SD	P25	P75
Income level								
$< 2 \times MMW$	-0.168	0.160	-0.189	-0.085	0.024	0.039	0.005	0.028
$[2, 5] \times MMW$	-0.216	0.206	-0.243	-0.110	0.030	0.050	0.006	0.035
$> 5 \times MMW$	-0.265	0.243	-0.309	-0.136	0.035	0.057	0.006	0.040
Location								
Metropolitan	-0.195	0.184	-0.223	-0.095	0.026	0.043	0.005	0.029
Adjacent	-0.182	0.170	-0.206	-0.093	0.026	0.042	0.005	0.029
Peripheral	-0.200	0.170	-0.233	-0.109	0.020	0.032	0.004	0.021
Diagnoses								
Chronic	-0.184	0.175	-0.206	-0.093	0.027	0.043	0.005	0.030
Not chronic	-0.183	0.173	-0.206	-0.092	0.026	0.042	0.005	0.030
Age group								
19-44	-0.182	0.173	-0.206	-0.092	0.026	0.043	0.005	0.030
45-64	-0.187	0.178	-0.209	-0.094	0.027	0.044	0.005	0.031
$\geq$65	-0.184	0.167	-0.207	-0.094	0.026	0.042	0.005	0.030
Beds								
< 125	-0.157	0.146	-0.174	-0.087	0.019	0.036	0.002	0.022
$[125, 250]$	-0.173	0.148	-0.202	-0.092	0.026	0.029	0.009	0.032
> 250	-0.248	0.234	-0.301	-0.122	0.040	0.054	0.010	0.054
Rooms								
< 7.5	-0.157	0.144	-0.174	-0.087	0.025	0.041	0.004	0.029
$[7.5, 15]$	-0.205	0.167	-0.256	-0.104	0.024	0.032	0.007	0.027
> 15	-0.260	0.276	-0.332	-0.125	0.040	0.052	0.010	0.052

Note: This table presents several statistics of the distribution of the estimated own- and cross-price elasticities conditional on patient and hospital characteristics, for the subsample of admissions that happen before reaching the out-of-pocket limit.

5.2 Bargaining game

I estimate the bargaining model on data from the main departments in the country (Antioquia, Atlántico, Bogotá, and Valle del Cauca) for computational simplicity. These departments represent 67% of admissions in the full dataset. Table (6) shows the estimators for τ and β_j together with their 95% confidence intervals in brackets based on 100 bootstrap resamples. 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 (8). The estimator for τ is statistically greater than one, which suggests that insurance companies place a weight of 37% on financial profits and 63% on enrollee welfare. All the estimators for the bargaining power of insurance companies are significant and, with the exception

of EPS016, greater than 0.5, so insurers extract most of the surplus generated in their interaction with hospitals.

Table 6: Bargaining parameters

	Estimate	CI
τ	1.682	[1.442,1.682]
β's		
EPS001	0.776	[0.604, 0.986]
EPS002	0.690	[0.446, 0.804]
EPS003	0.709	[0.530, 0.752]
EPS005	0.681	[0.531, 0.822]
EPS013	0.706	[0.455, 0.797]
EPS016	0.477	[0.197, 0.489]
EPS017	0.709	[0.468, 0.731]
EPS023	0.765	[0.543, 0.835]
EPS037	0.598	[0.249, 0.703]
EPS009	0.661	[0.571, 0.872]
EPS018	0.753	[0.520, 0.800]
EPS008	0.947	[0.908, 1.000]

Note: This table shows the estimates of τ and β_j using 2-step GMM. 95% confidence intervals in brackets based on 100 bootstrap samples of admissions within insurer-hospital-markets.

In table (7) I present the resulting average marginal costs and the average Lerner indexes. Their 95% confidence intervals are also reported in brackets. Columns 5 and 6 report the percentage of hospitals and insurers that satisfy their incentive compatibility and rationality constraints, respectively, given $\hat{\tau}$ and $\hat{\beta}_j$. Although all insurers have non-binding constraints, I find that marginal costs are greater than prices for 11% of hospitals in department 05, 4% in department 08, 3% in department 11, and 6% in department 76. I calculate the (quantity-weighted) average price, marginal cost, and Lerner index in each department conditional on the subsample of insurer-hospital pairs with positive surplus. Hospitals enjoy greater markups in markets with fewer competitors. In department 08, the average Lerner index equals 0.52, followed by department 11 with an average of 0.44.

Table 7: Average marginal costs and Lerner indexes

Department	Price	Marginal cost	Lerner index	Total pairs	Hospitals with positive surplus	Insurers with positive surplus
	(1)	(2)	(3)	(4)	(5)	(6)
05	166.95	143.41	0.141	185	88%	100%
		[122.72, 177.42]	[-0.050, 0.239]			
08	156.33	75.46	0.517	55	96%	100%
		[46.79, 89.95]	[0.426, 0.700]			
11	180.72	101.80	0.437	75	97%	100%
		[53.10, 103.88]	[0.410, 0.661]			
76	187.48	107.61	0.426	64	94%	100%
		[70.12, 129.23]	[0.298, 0.607]			

Note: This table reports the (quantity-weighted) average price, average marginal cost, and average Lerner index per department in columns 1, 2, and 3, respectively. Column 4 presents the total number of insurer-hospital pairs in the market, and columns 5 and 6 report the percentage of hospitals and insurers with predicted positive surplus during Nash-in-Nash bargains. These last two columns serve as statistics of model fit. 95% confidence intervals reported in brackets based on 110 bootstrap resamples of admissions at the insurer-hospital-market level.

5.3 Elasticity decomposition

Using the estimated profit parameters, in table (8) I decompose the effective price elasticity into its three components Ω^i, Ω^j and Υ, and report their averages and 95% confidence intervals (in brackets) across admissions in the full sample, in the sub-sample before reaching the OOP maximum, and in the sub-sample after reaching the OOP maximum. The average effective own-price elasticity is statistically less than 1 in the three samples, and greater (in absolute value) in the sub-sample of admissions that happen before the expenditure limit is reached compared to the full sample. My estimated effective own-price elasticity is smaller than in Gowrisankaran et al. (2015) who find an average ranging from 1.7 to 4.6, so 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 that is explained by consumer sensitivity to prices (42.5%) and the bargaining protocol (11.0%) are both lower than the percentage explained by insurer steering (46.6%). Across admissions before the OOP maximum, consumers are the most elastic side of the market, representing 56.6% of the total effective own-price elasticity, followed by insurers with 34.4%, and by the bargaining protocol with the remaining 9.0%. In the sub-sample of admissions that happen after the OOP limit is reached, insurers explain 92.3% of the effective elasticity and the bargaining game explains only 7.7%. Recognizing the effect of coinsurance rate discontinuities due to OOP maximums on demand elasticity is important to determine if counterfactual changes to cost sharing are welfare improving. For example, increasing the coinsurance rate while having the OOP thresholds fixed would have two opposing effects: it would decrease welfare for consumers who fail to reach their spending thresholds and end up facing a higher proportion of their healthcare costs, but it would increase welfare for consumers who do reach their spending limits and face zero prices. I study the distributional welfare effects of counterfactual cost sharing rules in

the next section.

Table 8: Own-price elasticity decomposition per market

Sample	Effective Own-Price Elasticity	Consumer Sensitivity	Insurer Sensitivity	Bargaining Protocol
	(1)	(2)	(3)	(4)
Full	-0.151	42.49%	46.55%	10.96%
	[-0.154, -0.044]	[42.29, 46.02]	[46.07, 50.04]	[3.95, 11.57]
Before OOP max	-0.195	56.59%	34.41%	9.01%
	[-0.196, -0.082]	[56.03, 61.42]	[34.15, 37.40]	[1.18, 9.81]
After OOP max	-0.100	0.00%	92.33%	7.67%
	[-0.102, -0.011]	—	[91.46, 100.0]	[0.00, 8.54]

Note: Column 1 reports the (quantity-weighted) average effective own-price elasticity per sample. Column 2 shows the percentage of the effective own-price elasticity explained by consumer sensitivity to prices Ω_i, column 3 is the percentage explained by insurer sensitivity to prices Ω_j, and column 4 is the percentage explained by the bargaining protocol Υ. 95% confidence intervals in brackets based on 110 bootstrap resamples of admissions at 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 the bargaining protocol. 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 departments in the country as for the bargaining game, 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 overdemand 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:

$$\tilde{\mathbf{p}} = \mathbf{mc} - \left(\tilde{\Omega}^j(\tilde{\mathbf{p}}) + \tilde{\Upsilon}(\tilde{\mathbf{p}})\right)^{-1} \tilde{\mathbf{q}}(\tilde{\mathbf{p}}) \tag{11}$$

Here $\tilde{\Omega}^j$ is the counterfactual matrix of demand elasticities due to insurer sensitivity to prices, $\tilde{\Upsilon}$ is the counterfactual matrix of derivatives of insurer profits (defined previously), and $\tilde{\mathbf{q}}$ is the counterfactual demand. Equation (11) 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 the bargaining protocol 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 patient 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 sensitivity to prices 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.

The solid black line in figure (3) presents the (quantity-weighted) average equilibrium prices during 2011 in the four main departments under *uniform* counterfactual coinsurance rates, and the dashed line corresponds to the observed weighted average price in that year. In this counterfactual exercise, I eliminate the observed three-tier system for coinsurance rates and assign the same percentage to all patients, while holding fixed their income-indexed OOP maximums. Results show that the average equilibrium price is U-shaped with respect to the coinsurance rate. The average price decreases 9.7% when going from 0 to 30% coinsurance and then increases 4.0% from 30 to 100% coinsurance. This pattern is consistent with the intuition that as the coinsurance rate increases, demand becomes more elastic because consumers face a higher proportion of

their healthcare costs. But as the 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.

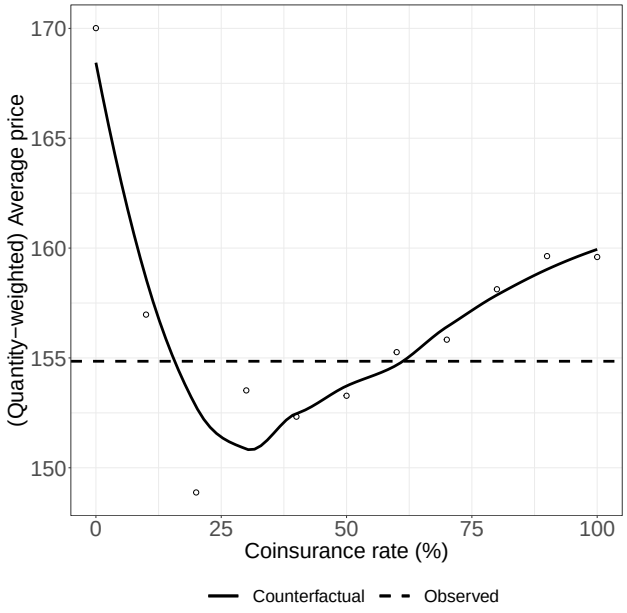
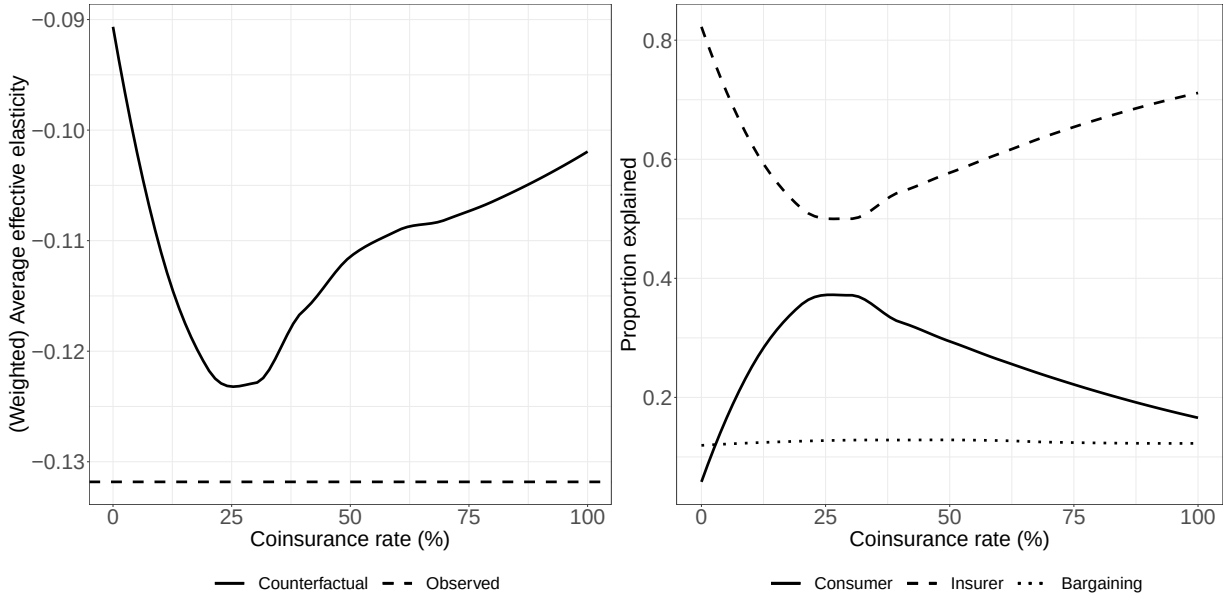


Figure 3: Average price under counterfactual coinsurance rates



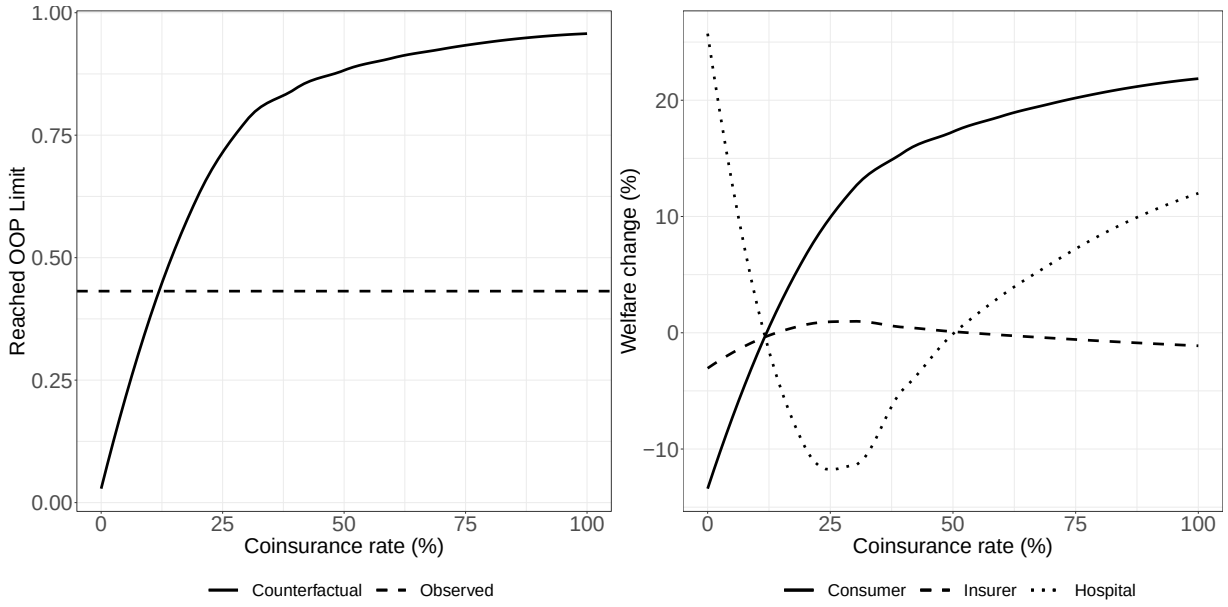
(a) Average effective elasticity

(b) Elasticity decomposition

Figure 4: Elasticities under counterfactual coinsurance rates

The elasticity patterns that explain the variations in counterfactual prices are depicted in panel (a) of

figure (4). The (quantity-weighted) average effective demand elasticity also follows a U-shaped pattern with respect to the coinsurance rate, going from -0.09 to -0.123 as the coinsurance rate increases from 0 to 30%, and then from -0.123 to -0.100 as the coinsurance rate increases from 30 to 100%. Panel (b) of this figure shows the decomposition of the effective elasticity into its three components due to consumer sensitivity to prices, insurer steering, and bargaining. While the percentage explained by the bargaining protocol remains mostly constant with respect to the coinsurance rate, the percentage explained by consumers is concave and achieves its maximum value at a coinsurance rate of 30%. At this value of the coinsurance rate, the change in proportion of patients that reach their OOP threshold is also maximal as seen in panel (a) of figure (5). Going from 0 to 30% coinsurance, the fraction of patients that reach their spending limit moves from 0 to 0.75, but as the coinsurance rate increases from 30 to 100%, this fraction only further increases by 0.25.



(a) Proportion that reach the OOP limit

(b) Welfare changes

Figure 5: Steering, and welfare under counterfactual coinsurance rates

Panel (b) of figure (5) presents the welfare change for consumers, insurers, and hospitals from counterfactual coinsurance rates relative to the observed scenario. Consumer welfare is calculated as $CS = \frac{1}{|\alpha_1|}\sum_{i,j} W_{ij}(\mathcal{H}_j, \mathbf{p}_{jh}, \kappa_i, \mathbf{g}_h)$, under the restriction that $(1-\kappa_i)p_{jh} = 0$ or assuming there is no insurer steering. Insurer surplus is calculated as $IS = \sum_j \pi_j(\mathcal{H}_j, \{\mathbf{p}_{jh}\}_{h\in\mathcal{H}_j}, \kappa_i)$ and hospital surplus as $HS = \sum_h \pi_h(\mathcal{J}_h, \{\mathbf{p}_{jh}\}_{j\in\mathcal{J}_h})$. As expected, the figure shows that hospital profits follow the same pattern as average equilibrium prices. When the coinsurance rate equals 0, demand is relatively inelastic and prices increase in a way that hospital profits grow more than 20% relative to the observed scenario. At 30% coinsurance, when the demand is most elastic, hospital profits decrease around 10%. Consumer surplus follows the same pattern with respect to the coinsurance rate as the proportion of patients that reach the spending limit.

Consumer welfare decreases relative to the observed scenario at a uniform coinsurance rate of 10% because future healthcare prices are now relatively higher or, in other words, because consumers will now take longer to reach full insurance coverage relative to the observed case. At a coinsurance rate of 75%, consumer welfare increases 20% relative to the observed scenario, because even though they face a higher proportion of their healthcare costs, patients reach their spending limits sooner than in the observed scenario. These findings show the importance of controlling for discontinuities in OOP prices when analyzing welfare changes in health insurance markets. Without accounting for these discontinuities, we would wrongly predict that an increase in the coinsurance rate would indistinctly lead to lower consumer surplus. However, the findings also suggest that a policymaker should take into account the distributional effects of these policies, for example how the financial default risk varies across income groups when the coinsurance rate is set above the average observed rate of 13.2% and OOP limits remain fixed. In figure (6), I present the mean (solid line), 25th and 75th percentiles (dashed lines) of the distribution of changes in consumer surplus at different values of the coinsurance rate and conditional on income level. Individuals earning less than 2 times the MMW experience net surplus gains at coinsurance rates above 12.5%, with a lower bound on the welfare change equal to 5% and an upper bound equal to 50%. While those earning more than 2 times the MMW, experience net welfare losses at uniform coinsurance rates between 12.5% and 60%.

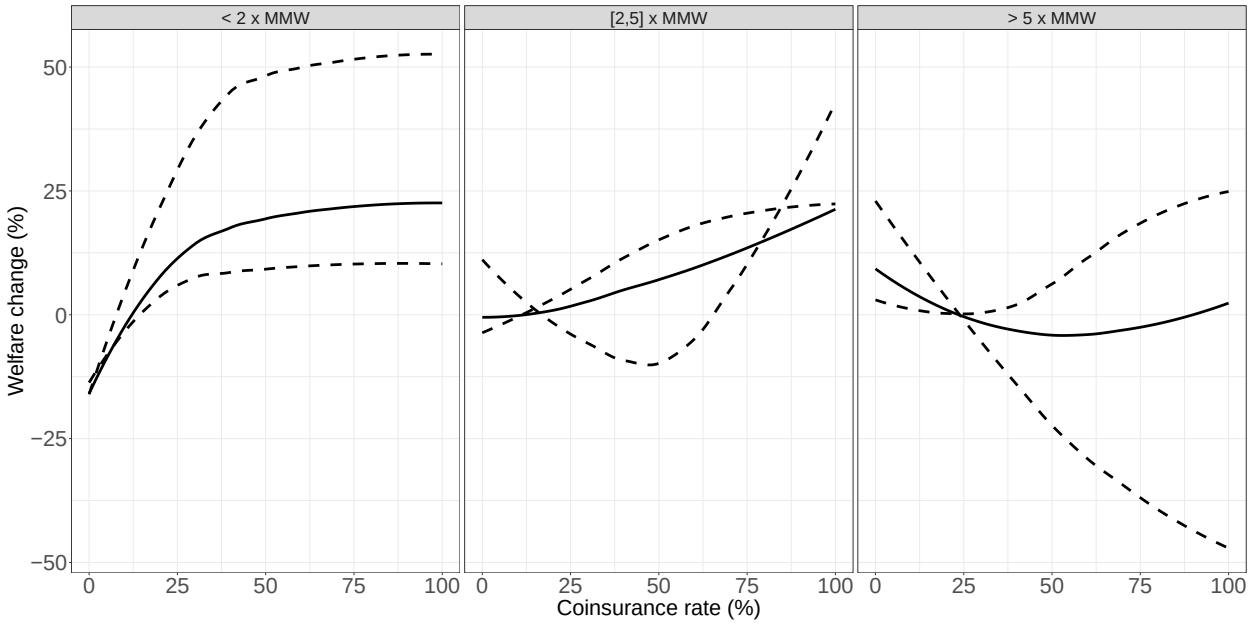


Figure 6: Consumer welfare change by income level under counterfactual coinsurance rates

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 three-tier coinsurance rate scheme 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. The observed average OOP limit equals 495.5 thousand COP. Figure (7) depicts the (quantity-weighted) average equilibrium prices under these counterfactual spending amounts in the main departments during 2011. Compared to the counterfactuals of the previous section, the average price does not respond significantly to changes in the OOP limits, even though the elasticity of demand varies highly as seen in panel (a) of figure (8). At the lowest value of the spending threshold, when the elasticity of demand decreases (in absolute value) by 22%, prices only increase 2.5% relative to the observed scenario. When the OOP limit is set to a MMW, demand elasticity increases (in absolute value) almost 46% relative to the observed case, but prices do not decrease as a result and are in fact statistically equal to observed ones. The explanation for the disconnection between average prices and demand elasticity is that supply of health insurance is highly elastic in these counterfactuals. Even as consumer demand decreases with the OOP limit, insurers have to cover a proportion of $1-\kappa_i$ of healthcare costs for longer as the expenditure limit is increased. As consumers eventually reach their spending thresholds, insurance supply remains sensitive to price given that carriers now have to cover the full cost of healthcare.

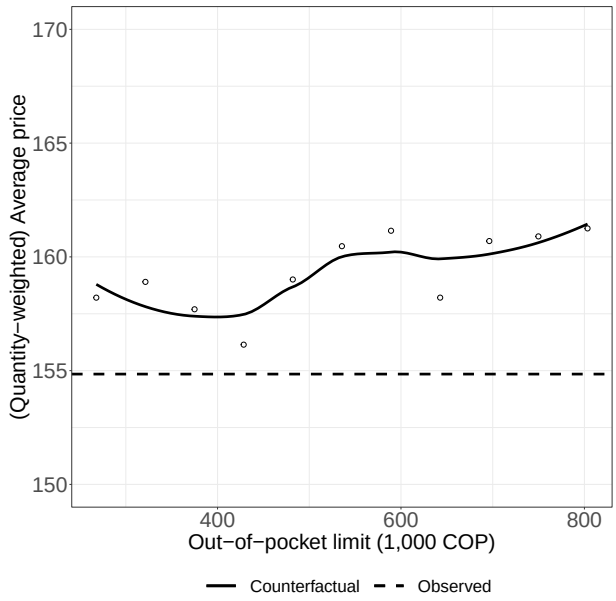
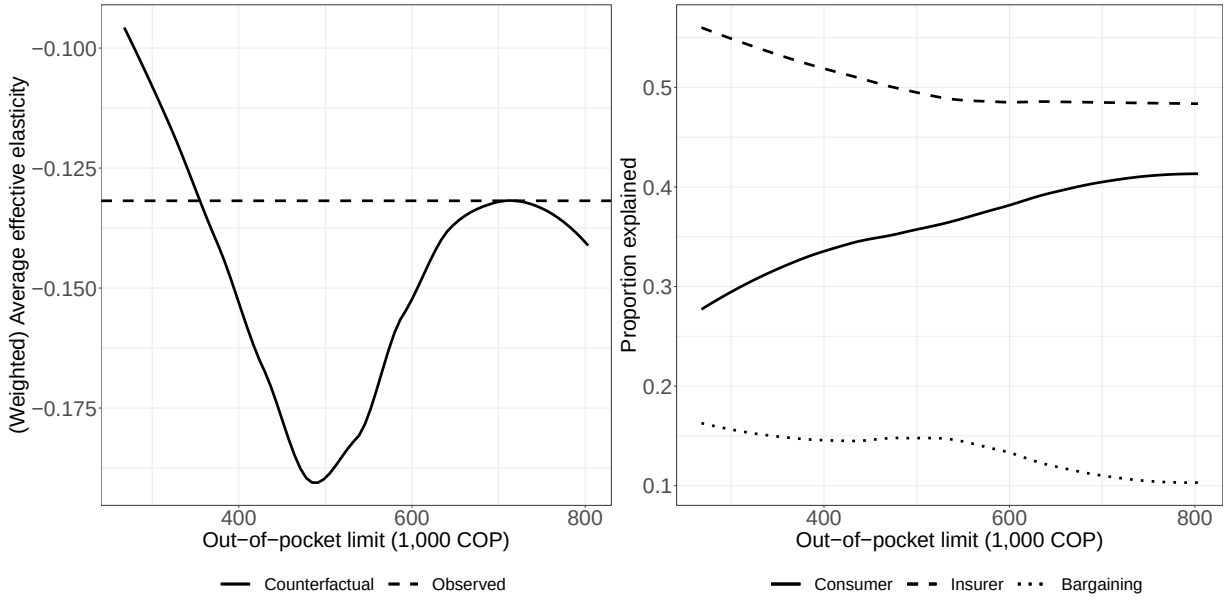


Figure 7: Average price under counterfactual OOP limits

Panel (b) of figure (8) presents the decomposition of demand elasticity into its portions explained by consumers, insurers, and the bargaining protocol. As the uniform OOP limit is increased from $MMW \times 0.5$ to $MMW \times 1.5$, the proportion of the elasticity that is explained by consumer sensitivity to prices increases from 0.3 to 0.4, while the proportion explained by insurer steering falls from 0.55 to 0.48 and the part due to bargaining falls from 0.17 to 0.10. Insurer and consumer sensitivity to prices converge as the OOP limit increases while the role of bargaining disappears, which means that market outcomes converge to Nash-Bertrand outcomes.

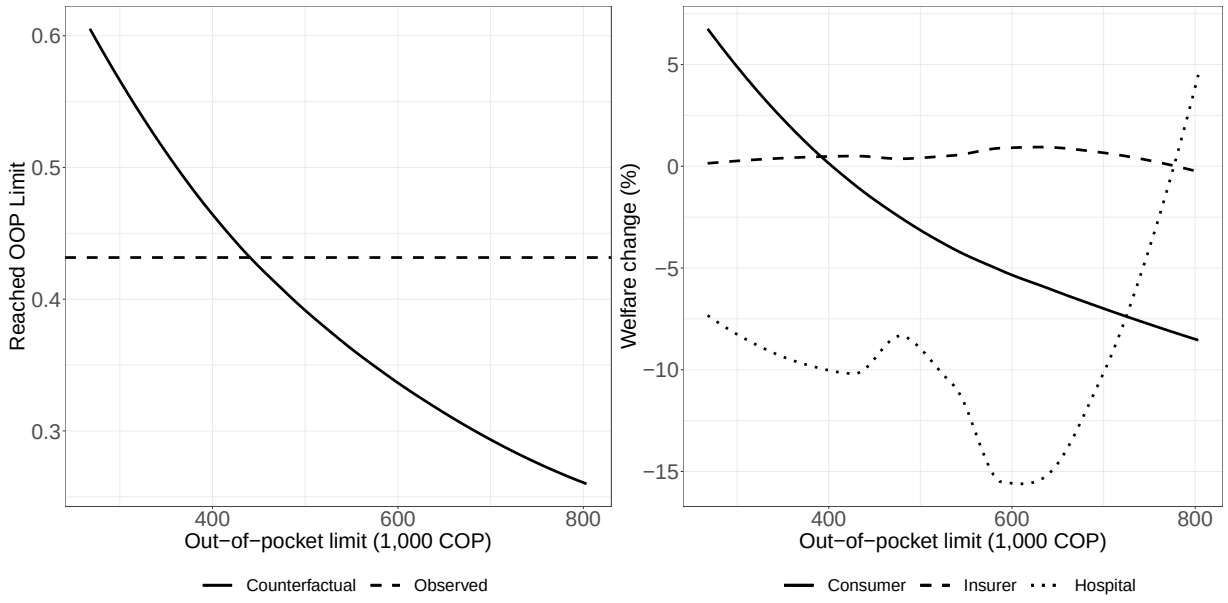
Panel (d) of the figure shows the welfare change for consumers, insurers, and hospitals under these

counterfactuals relative to the observed scenario. The consumer surplus function is decreasing and convex with respect to the OOP limit as is the proportion of patients that reach the spending thresholds shown in panel (c). Insurers experience no significant changes in surplus at different values of the OOP limit, while hospital surplus increases almost 5% relative to the observed scenario when the OOP limit is set to $MMW \times 1.5$.



(a) Average own-price elasticities

(b) Elasticity decomposition



(c) Proportion that reach the OOP limit

(d) Welfare changes

Figure 8: Elasticities, steering, and welfare under counterfactual OOP limits

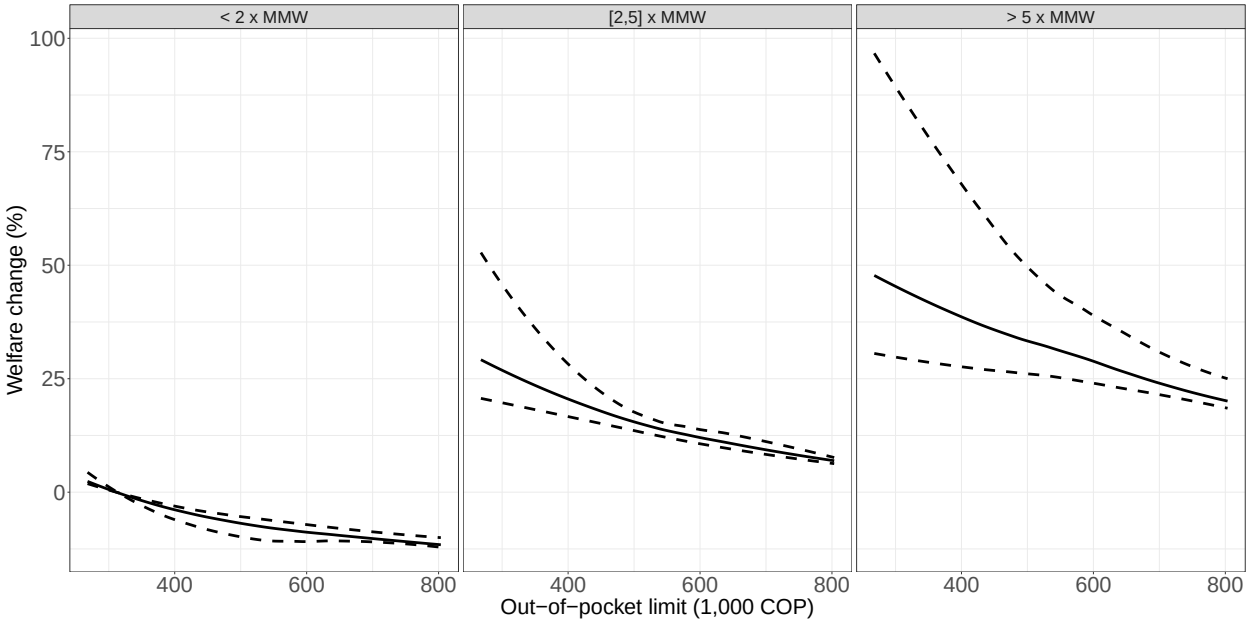


Figure 9: Consumer welfare change by income level under counterfactual OOP limits

In figure (9) I present the distributional welfare effects of the type of policy studied in this subsection, by conditioning consumer welfare changes to the enrollee's monthly income level. The solid line in each panel represents the average welfare change and the dashed lines are the associated 25th and 75th percentiles. Average welfare changes and their variance are, in general, increasing with income. When the OOP limit is uniform and set to 500 thousand COP, low income individuals –earning less than 2 times a MMW– experience net surplus losses of at most 10%, while individuals earning between 2 and 5 times the MMW, experience net welfare gains of 12.5%, relative to the observed scenario. Conditional on the income level, welfare changes follow a decreasing pattern with respect to the OOP limit. Finally, for consumers with incomes above 5 times the MMW, welfare gains can be as high as 50% and as low as 25% when the OOP limit equals 500 thousand COP. The implication of these results for the design of optimal health insurance plans is that in systems where premiums are fixed, cost sharing and maximum dollar amounts should be proportional to the elasticity of demand, which in Colombia is determined by the enrollee's income level. A non-linear cost sharing schedule like the one in Colombia can generate important welfare gains relative to a uniform cost sharing plan.

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.

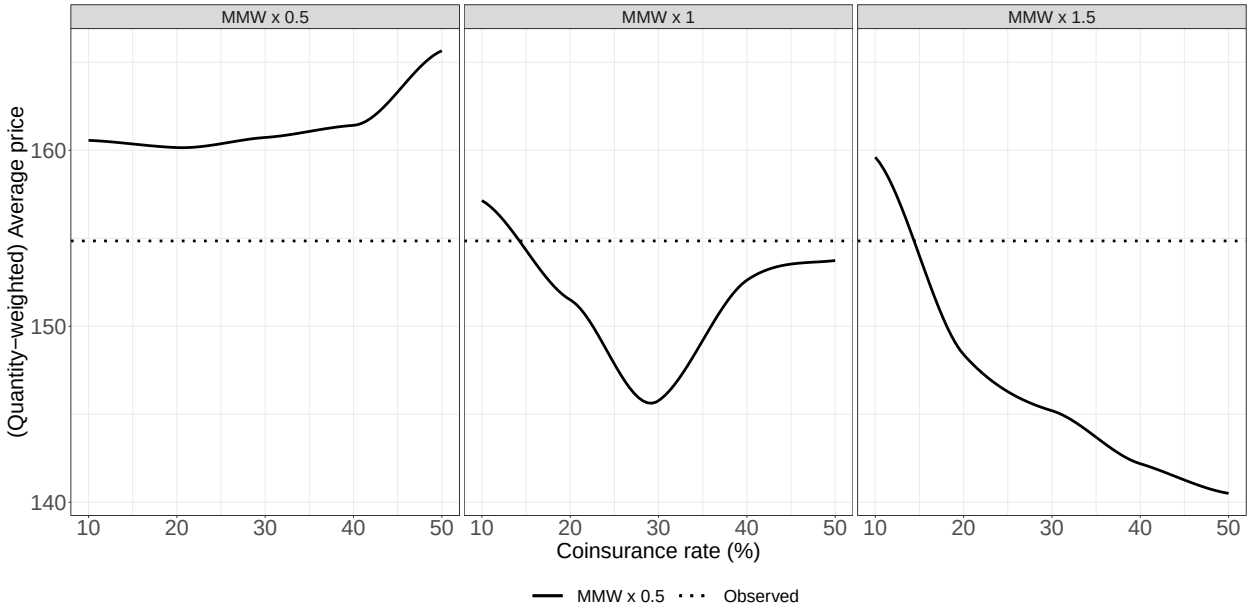


Figure 10: Average price under counterfactual coinsurance rates and OOP limits

Figure (10) presents the (quantity-weighted) average counterfactual prices from these exercises. The x-axis denotes the coinsurance rate, the y-axis is the average price, and each panel 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 $MMW \times 0.5$, a higher proportion of patients reach their expenditure thresholds (see figure A1 in the appendix) and demand is more inelastic, relative to the observed scenario (see panel (a) of figure A2 in the appendix). Equilibrium prices, thus, increase overall 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 is increased to $MMW \times 1.0$, which is above the observed average limit, the proportion of patients that reach full insurance coverage at every possible value of the coinsurance rate decreases. Demand becomes more elastic relative to the previous counterfactual because consumers now face higher future OOP expenditures. So, equilibrium prices decrease overall, following a U-shaped pattern with respect to the coinsurance rate. In the last panel of the figure where I set the OOP limit to $MMW \times 1.5$, even fewer individuals reach the spending threshold at every value of the coinsurance rate and prices exhibit even larger decreases compared to the counterfactual where I set the OOP limit to $MMX \times 1.0$. The price curve in this case is decreasing and convex with respect to the coinsurance rate.

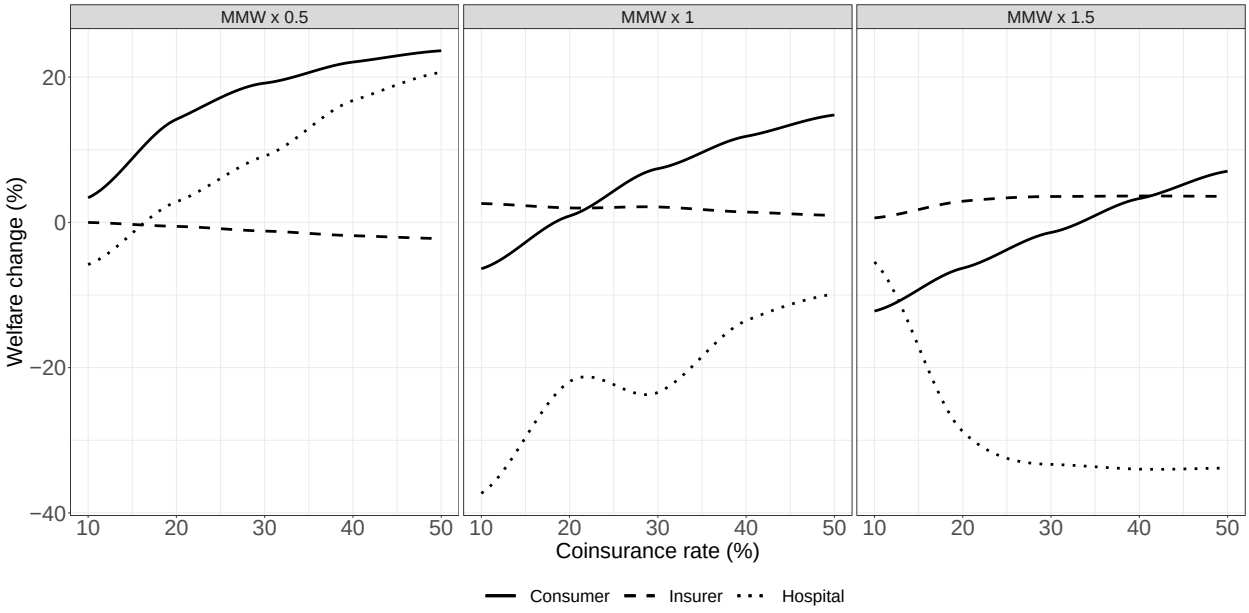


Figure 11: Welfare under counterfactual coinsurance rates and OOP limits

Figure (11) presents the welfare change for consumers, insurers, and hospitals in each counterfactual exercise relative to the observed scenario. While insurer surplus remains mostly constant across different coinsurance rates within an OOP limit and across different OOP limits, consumer surplus and hospital surplus exhibit large changes relative to the observed scenario. I find that the consumer surplus function decreases as the OOP limits increase and, conditional on an OOP threshold, the function is increasing with respect to the coinsurance rate and follows the same pattern as the proportion of patients that reach full insurance coverage. Hospital surplus also decreases with the OOP limits because prices are decreasing. But depending on the value of the OOP threshold, the coinsurance rate has different effects on the curvature of the hospital surplus function. At low values of the spending limit like those equal to $MMW \times 0.5$, the hospital surplus function is increasing and concave with respect to the coinsurance rate; while at high values of the threshold like those equal to $MMX \times 1.5$, the function is decreasing and convex. Although the second panel of figure (11) shows that a higher coinsurance rate has the potential to increase consumer surplus as much as 17% relative to the observed scenario, figure (12) where I present the distribution of consumer gains conditional on income level and on an OOP limit of $MMW \times 1$, indicates that the variance of such welfare gains not only increase with the income level but also with the coinsurance rate.

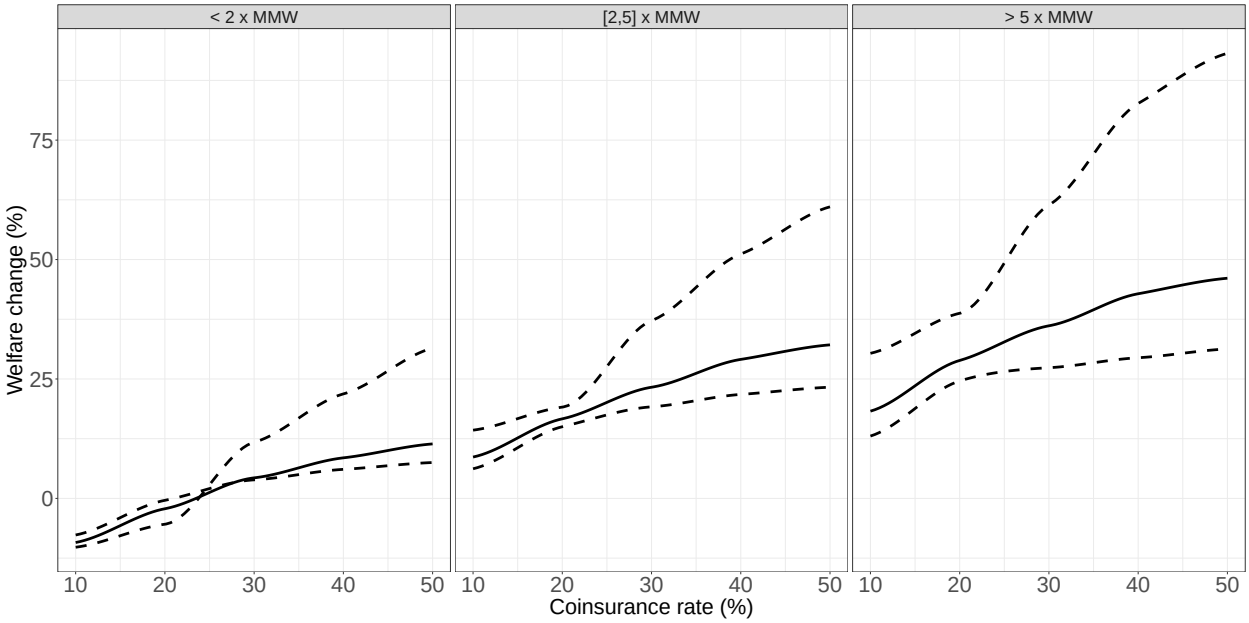


Figure 12: Consumer welfare change by income level under counterfactual coinsurance rates and an OOP limit of $MMW \times 1$

7 Conclusions

In this paper I quantify the effect of counterfactual cost sharing rules (coinsurance rates and maximum out-of-pocket expenditures) on negotiated prices between insurers and hospitals, social welfare, and distributional welfare effects. I also provide evidence to the question of whether a uniform cost sharing scheme is optimal in social insurance programs. My empirical application is the Colombian healthcare system where coinsurance rates and maximum out-of-pocket amounts are indexed to the enrollee's monthly income level in a three-tier system exogenously determined by the government. I model the Colombian healthcare market as a two stage game. First, insurers and hospitals engage in bilateral negotiations over the base 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 can affect consumer demand for hospitals because in a tightly regulated market such as in Colombia where carriers are unable to compete in premiums, they engage in steering consumer health claims towards certain preferred providers. These steering incentives are identified from the discontinuity in coinsurance rates due to the out-of-pocket maximums and get reflected in hospital demand as 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 the bargaining protocol. I find that consumer demand for hospitals in Colombia is inelastic relative to other elasticity estimates in the literature, even after accounting for bargaining and insurer steering motives. This suggests that the strong regulation

of coinsurance rates and premiums in Colombia can make demand relatively more inelastic than it would otherwise be if insurers had discretion on plan design.

I find that when holding out-of-pocket limits fixed, average equilibrium hospital prices are U-shaped with respect to the coinsurance rate, with the inflection point happening at 30% coinsurance; but they remain mostly invariant in a counterfactual scenario where the coinsurance rates are fixed and spending thresholds are allowed to vary. This shows that patients in Colombia are mostly myopic about their future healthcare costs and respond more heavily to changes in the spot price rather than the future price of healthcare. I find that consumer welfare is directly proportional to the number of patients that reach their spending thresholds, so that even in a counterfactual where out-of-pocket limits are held fixed and the coinsurance rate is increased, consumers on average experience welfare gains relative to the observed cost sharing system. These welfare gains are driven by the subsample of patients with monthly incomes below 2 times the minimum wage, but those who earn more than that experience net welfare losses. When I allow the spending thresholds to vary while holding the coinsurance rates fixed, the distributional effects are reversed: low income individuals tend to experience welfare losses while high income patients experience welfare gains relative to the observed scenario. Finally, when combining both sets of counterfactual analyses, my results show that coinsurance rates determine variations along the price curve while out-of-pocket limits determine overall price levels.

The results in this paper have two implications for the optimal 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. If consumers account for a higher proportion of demand elasticity, then a combination of higher coinsurance rates and lower out-of-pocket limits can help maximize consumer welfare. But a policymaker should also take into account the distributional welfare effects and understand how uniform cost sharing policies can 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.

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.

Comisión de Regulación en Salud (2010). Acuerdo 19 de 2010. http://legal.legis.com.co/document/Index?obra=legcoldocument=legcol$_9$90953$ffb8bca086e0430a010151a086.Lastchecked$

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.

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.

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.

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). Formulary Design and Targeted Rationing of Care: The Case of 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. 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.

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.

Appendices

A Additional Outcomes for Simultaneous Counterfactual Cost Sharing and OOP Limits

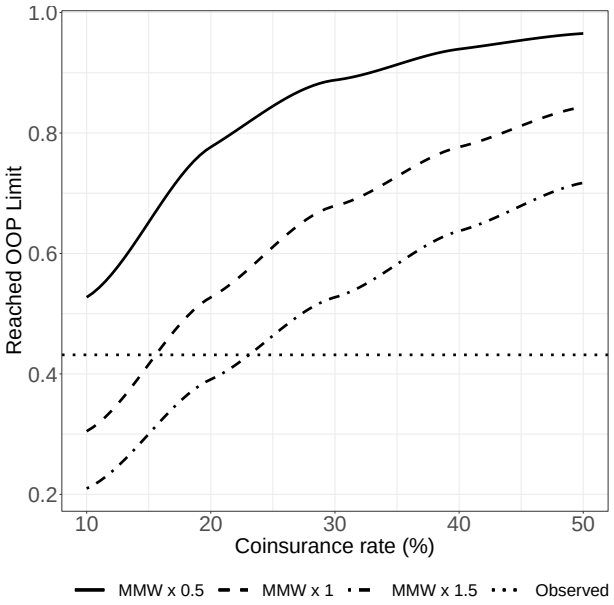
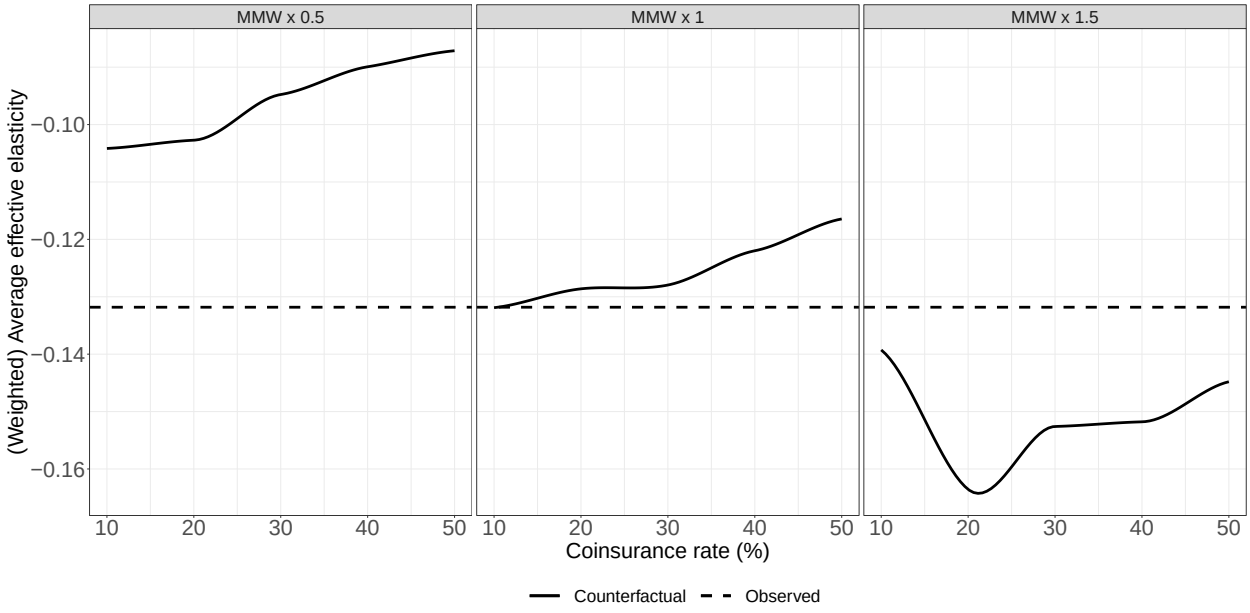
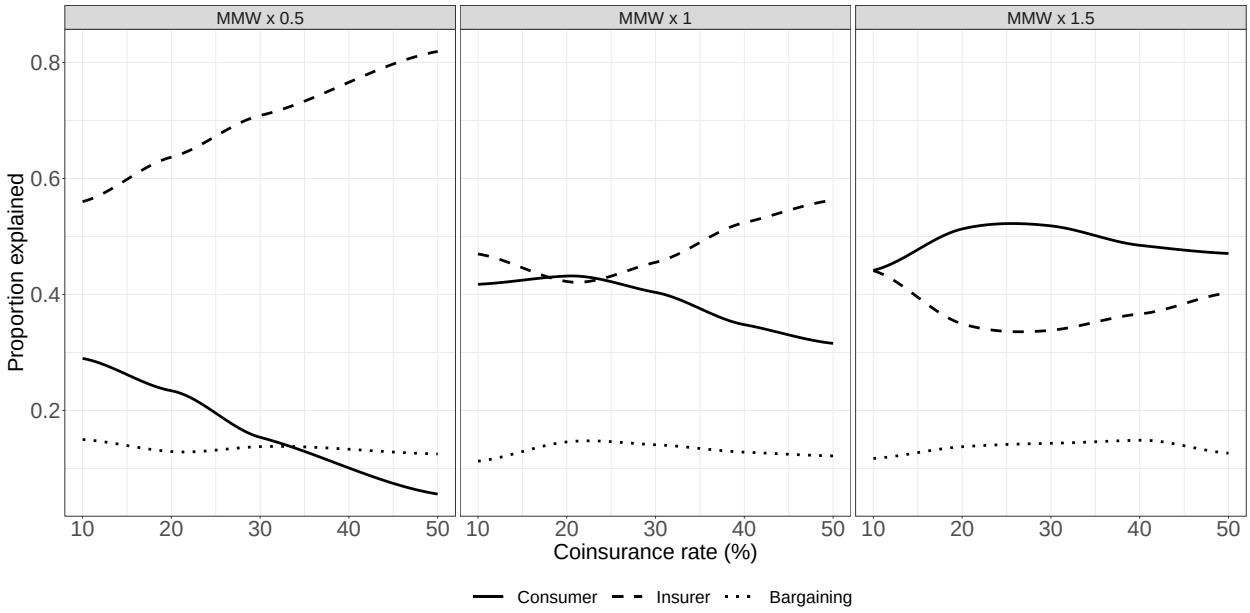


Figure A1: Proportion of patients that reach their OOP limit



(a) Average effective elasticity



(b) Elasticity decomposition

Figure A2: Elasticities under counterfactual coinsurance rates and OOP limits