Health Insurer Gatekeeping

Natalia Serna*

Stanford University

Abstract

Health care costs increase every year despite widespread use of cost-sharing. In this paper I study the role of insurer gatekeeping as an alternative, more effective cost containment mechanism that does not deteriorate patient health. I identify effects of gatekeeping using the discontinuity in cost-sharing introduced by the out-of-pocket maximum. I find that gatekeeping has much larger effects on health care utilization and spending than cost-sharing, which suggests that providing free insurance is not detrimental for the financial stability of health systems. I show that gatekeeping has no effects on individual mortality and that insurers are more likely to gatekeep discretionary services than necessary care. Estimates from a structural model of hospital demand show that gatekeeping induces patients to pay 2 percent more to reduce commuting distance to hospitals by 1 kilometer. Information frictions mute the effects of gatekeeping.

*e-mail: nserna@stanford.edu. I am deeply grateful to the Colombian Ministry of Health for providing the data for this research. I want to thank Marguerite Burns, Ken Hendricks, Grant Miller, Corina Mommaerts, Maria Polyakova, and Alan Sorensen for their advice. The findings of this paper do not represent the views of any institution involved. All errors are my own.

1 Introduction

Substantial research in health economics has found evidence that individuals respond to cost-sharing by reducing their consumption of health care (Serna, 2021; Aron-Dine, Einav, and Finkelstein, 2013; Eichner, 1998; Newhouse, 1993; Manning, Newhouse, Duan, Keeler, and Leibowitz, 1987). Empirical evidence has also shown that access to health insurance increases spending (Finkelstein, Taubman, Wright, Bernstein, Gruber, Newhouse, Allen, Baicker, and Group, 2012), but has small impacts on patient health outcomes for certain populations (Finkelstein and McKnight, 2008). This evidence has motivated the use of patient cost-sharing as a cost containment mechanism in many countries. However, health care costs continue to increase every year, which raises questions on whether cost-sharing is an effective tool to control health care spending or whether it exacerbates underuse of high-value care (Baicker, Mullainathan, and Schwartzstein, 2015).

In this paper I explore the role of insurer gatekeeping as an alternative, more effective mechanism for controlling rising health care costs that, at the same time, does not deteriorate patient health. I show that effects of insurer gatekeeping on health care utilization and spending are much larger than cost-sharing, which suggests that the provision of free health insurance is not detrimental for the financial stability of health care markets.

I study the effects of gatekeeping in the context of Colombia's contributory health care system. This health system covers individuals who can pay their taxes and provides access to the national health insurance plan. The government strictly regulates several aspects of this plan including premiums and cost-sharing. Cost-sharing rules (copays, coinsurance rates, and maximum out-of-pocket (OOP) amounts in the year) are a function of the enrollee's monthly income level but are standardized across

insurers and hospitals.

To identify gatekeeping I leverage the discontinuity in coinsurance rates introduced by the OOP maximum. Coinsurance rates drop to zero after patients reach this limit and the insurer has to cover the full cost of health care. I show that reaching the OOP maximum in my setting is a random and sudden event, typically a hospitalization, and therefore individuals cannot preempt it. This provides a unique framework to estimate the causal effect of insurer gatekeeping on health care utilization, spending, and mortality. I use enrollment and claims data from a random panel of 8 million enrollees from 2009 to 2011, who did not switch their insurer over the sample period.

I start by comparing the mean claim cost and the likelihood of making claims between individuals who reach their OOP maximum and those who don't, in a difference-in-differences event study design. I show that treated and control individuals have parallel spending and utilization patterns before reaching the OOP maximum, which confirms that treatment in my setting is as good as random. After reaching this maximum, treated individuals consume significantly cheaper services and have a substantially lower likelihood of making claims than controls. These findings are at odds with behavioral assumptions about consumers when they face zero prices and are also inconsistent with individuals' health status worsening due to sudden health shocks or hospitalizations. Instead, results are in line with the expected effects of insurer gatekeeping and information frictions that make consumers unaware of zero prices.

I find that treatment effects on the mean claim cost are entirely driven by individuals with poor health, but effects on the likelihood of making claims are homogeneous across healthy and sick patients. Individuals who are diagnosed with a chronic disease consume services that are nearly 100 thousand pesos cheaper than controls 5 months after reaching the OOP maximum; but I find no change in the mean claim cost among individuals without diagnoses after the event relative to controls.

Although healthy and sick patients likely face information frictions, there are no reasons to believe that these frictions differ across the two groups. In fact, intuition would potentially suggest the opposite: if sick individuals have more interaction with the health care system, they may be more likely to know when they reach their OOP maximum. To separate the effect of information frictions from the effect of gatekeeping, I leverage the fact that information frictions should disappear over time the more patients claim services after reaching the OOP maximum.

To test this hypothesis, I estimate my event study specification separately for cohorts of patients that reach their OOP maximum in different months of the year. If information frictions disappear over time, then cohorts who reach their maximum early on should consume more expensive services the closer they are to the end of the calendar year. My findings show no evidence that negative treatment effects vanish for any cohort. Reductions in the mean claim cost and the likelihood of making claims are both persistent, suggesting that information frictions are not the main driver of reductions in spending.

Event study results conform to the idea that insurer gatekeeping is an important source of price elasticity in health care demand. I show that insurers are less likely to steer or deter claims made in an inpatient setting, but are more likely to gatekeep discretionary care made in an outpatient setting, such as imaging and laboratory tests. Because gatekeeping may involve steering patients towards cheaper providers or denying claims altogether, its use raises questions about the impact on patient health. Conditional on patients who reach their OOP maximum in a regression discontinuity framework, I find no change in individual mortality after the event, hence gatekeeping does not deteriorate patient health.

In the last part of the paper I turn to examining the impacts of gatekeeping and information frictions on the types of providers that consumers choose to receive care.

I develop and estimate a structural model of hospital demand that incorporates information frictions, and consumers' and insurers' responsiveness to prices in two states of the world: before and after patients reach their OOP maximum. The structural model allows me to derive changes in the marginal disutility of distance that are due to insurer gatekeeping and information frictions. The model also allows me to quantify the bias in demand elasticity estimates if researchers fail to consider the, possibly unobserved, effects of gatekeeping.

My model estimates show significant responsiveness to prices before and after patients reach their OOP maximum, in line with the reduced-form evidence. Estimates show that consumers dislike commuting to visit health care providers. In a partial equilibrium exercise where I prohibit insurer gatekeeping, I find that patients on average would be willing to pay 2 percent more than in the observed scenario to reduce commuting distance by 1 kilometer. Put differently, gatekeeping induces individuals to travel on average 2.5 additional blocks to receive care. Information frictions have similar impacts on the marginal effect of distance.

This paper contributes to the literature on the use of non-price mechanisms to contain health care costs and unnecessary spending. For example, Brot-Goldberg, Burn, Layton, and Vabson (2023) study insurers' use of prior authorization in Medicare Part D, Shi (2023) analyzes the impacts of Medicare spending audits, and League (2022) examines claim denials in Traditional Medicare. I quantify the overall effects of insurer gatekeeping on utilization, spending, mortality, and provider choice. My paper also contributes to the literature on cost-sharing discontinuities and, in particular, what happens when health care prices are zero. Iizuka and Shigeoka (2022) show that health care demand increases discontinuously when prices are zero in line with behavioral moral hazard. My findings in the Colombian setting show reductions in demand after prices are zero, which is incompatible with consumers' behavioral

responses.

The remainder of this paper is structured as follows: section 2 describes the empirical setting, section 3 describes my data, section 4 provides the designed-based empirical analysis to identify gatekeeping, section 5 presents the structural model of hospital demand, section 6 presents results from the partial equilibrium analyses, and section 7 concludes.

2 Cost-Sharing in Colombia

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

The government regulates several aspects of the national plan: premiums are set to zero in both regimes, individuals in the contributory system have to pay a fraction of their health care expenditures through cost-sharing, and health care is free in the subsidized system. Insurers have no discretion on how to design these elements of the insurance plan, but they can decide on their network of preferred providers and negotiate health service prices with them.

Cost-sharing rules in the contributory system are a function of the enrollee's monthly income level but are standardized across insurers and hospitals. These rules involve a three-tier system of copayments, coinsurance rates, and maximum out-of-pocket (OOP) amounts in the year as seen in table 1. Individuals are assigned specific cost-sharing rules depending on whether they make less than 2, between 2 and 5, or

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

Income level y	Copay	Coinsurance rate	OOP Maximum	
		Per claim	Per claim	Per year
$y < 2$ MMW	1,900	11.5%	28.7%	57.5%
$y \in [2, 5]$ MMW	7,600	17.3%	115%	230%
$y > 5$ MMW	20,100	23.0%	230%	460%

Note: Table shows the copays, coinsurance rates, and OOP maximum per income level that apply to individuals enrolled in Colombia's contributory health care system. The monthly minimum wage (MMW) in 2009 equals 496,900 COP or roughly 231 USD. The coinsurance rates are percentages of claims cost, whereas the OOP maximums are percentages of the MMW.

more than 5 times the monthly minimum wage (MMW). For example, for individuals who make less than 2 times the MMW, the copay equals 1,900 pesos (nearly $1), the coinsurance rate is 11.5 percent of the price per health claim, and the maximum OOP amount is 28.7 percent of the MMW per health claim and 57.5 percent of the MMW per year. Enrollees make copayments every time they go to a primary care or specialist appointment and they pay coinsurance rates for every health service that they claim. After individuals reach their OOP maximum in the year, copays and coinsurance rates drop to zero and the insurer covers the full cost of their health care.

These cost-sharing rules have not changed since the establishment of the health care system and vary only to the extent that the monthly minimum wage changes every year with inflation. Previous studies have analyzed the impacts of coinsurance rate discontinuities in the Colombian health care system on utilization, spending, and health outcomes (Serna, 2021; Buitrago, Miller, and Vera-Hernández, 2021). These studies leverage comparisons across consumers in the different income tiers. In this paper instead I study within-patient changes in outcomes before and after they reach their OOP maximum.

3 Data and Descriptives

The data to study demand responses to the OOP maximum and to identify insurer gatekeeping come from Colombia's contributory healthcare system. The data consist of all the health claims of a random sample of nearly 8 million enrollees from 2009 to 2011 who made at least one claim and who did not switch their insurer during the sample period.

For every individual I observe basic socio-demographic characteristics including sex, age, income, and municipality of residence. The data reports insurer, provider, service, ICD-10 code, type of contract, and negotiated price associated with each health claim. Using the enrollee's income I recover their level of cost-sharing and the OOP maximum that applies to each of them. With the health claims data I construct different measures of monthly utilization and spending and determine whether and when they reach their OOP maximum.

I consider observations from one individual in different years as different individuals. This is because cost-sharing resets at the beginning of each calendar year, hence we can expect consumer and insurer behavior relative to the cost-sharing rules to be similar across years. This assumption implies that I consider my data as repeated cross-sections, and exploit the variation within years. For tractability, I choose a random sample of 200,000 individuals per year.

Table 2 presents some summary statistics of my sample. An observation is an individual. Column (1) shows descriptives for the full sample, column (2) for the sample of people who reach their OOP maximum in the year, and column (3) for the sample of people who do not reach their OOP maximum. 3 percent of individuals in my sample reach their OOP limit. These individuals are on average older and of lower income than those who do not reach the limit. 65 percent of those in column (2)

TABLE 2: Summary Statistics

	Full sample (1)	Above OOP max (2)	Below OOP max (3)
Socio-demographic			
Male	0.48 (0.50)	0.47 (0.50)	0.48 (0.50)
Age	47.0 (17.0)	58.5 (17.8)	46.7 (16.9)
Low income	0.75 (0.43)	0.93 (0.26)	0.75 (0.43)
Medium income	0.19 (0.39)	0.06 (0.24)	0.19 (0.39)
High income	0.06 (0.23)	0.01 (0.10)	0.06 (0.24)
Health			
Cancer	0.17 (0.38)	0.29 (0.45)	0.17 (0.37)
Cardiovascular	0.32 (0.47)	0.65 (0.48)	0.31 (0.46)
Pulmonary	0.05 (0.21)	0.19 (0.39)	0.04 (0.20)
Renal	0.03 (0.18)	0.16 (0.37)	0.03 (0.17)
Monthly health care use			
Mean monthly cost	39.9 (195)	335.4 (956)	30.0 (76.9)
Total monthly cost	55.0 (254)	787 (1,096)	30.7 (91.3)
Prescription claims	0.28 (0.99)	1.30 (2.96)	0.24 (0.83)
Outpatient claims	1.16 (1.61)	3.72 (3.87)	1.08 (1.40)
Hospitalization	0.01 (0.03)	0.08 (0.09)	0.00 (0.02)
Observations	600,000	19,262	580,738

Note: Mean and standard deviation in parenthesis of consumer characteristics in the full sample in column (1), in the sample of those who reach their OOP maximum in column (2), and in the sample of those who do not reach their OOP maximum in column (3). An observation is an individual. Cost variables are measured in millions of pesos.

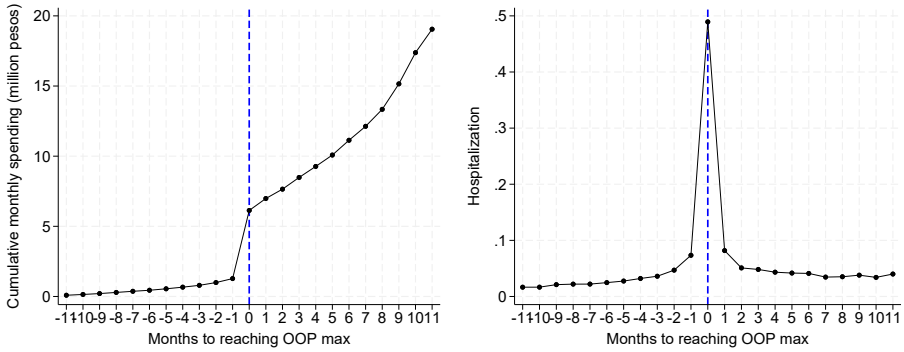
have a cardiovascular disease (such as hypertension), while only 31 percent of those in column (3) have this type of diagnosis. Consumers who reach their maximum have higher health care utilization than their counterparts, a difference that is driven only by the health claim made when they reach this maximum. For example, the mean claim cost is over 10 times higher and the likelihood of being hospitalized in a month is 8 percentage points higher for those in column (2) relative to column (3).

Reaching the OOP maximum is a sudden event. Panel A of figure 1 shows that cumulative monthly spending increases smoothly until the month before reaching the OOP maximum and has a sharp discontinuity when the limit is reached. This sudden event is typically a hospitalization as seen in panel B of the figure. A little under 50

percent of individuals who reach the OOP maximum have a hospitalization and the remaining half either claim an expensive imaging service or an expensive visit to the specialist as seen in appendix figure 1.

FIGURE 1: Cumulative spending and hospitalizations by month

(A) Cumulative spending (B) Hospitalizations



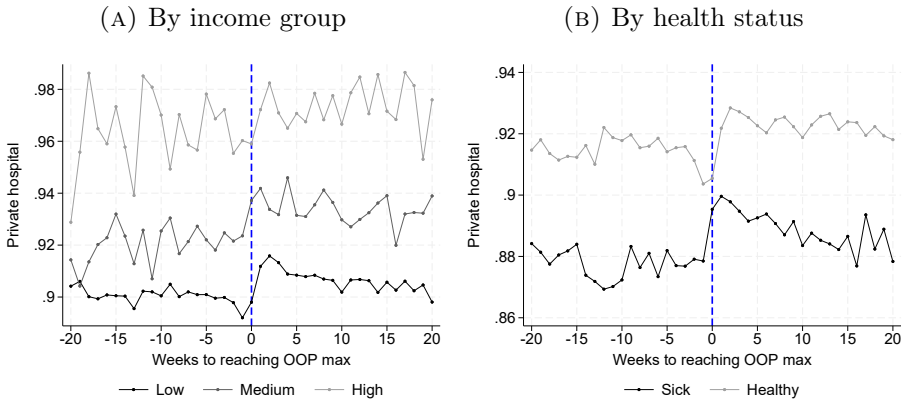
Note: Figure shows average cumulative monthly spending in panel A and average number of hospitalizations in panel B by month relative to the month in which the individuals reaches her OOP maximum.

Individuals who reach their OOP maximum are more likely to be treated at a private hospital after the event. Panel A of figure 2 shows that while the baseline probability of visiting a private hospital is increasing with income, this probability is higher after the reaching the OOP maximum conditional on the income group. Panel B presents qualitatively similar changes in the probability of visiting a private hospital among individuals with a chronic disease diagnosis and those without diagnoses. The baseline probability of visiting a private hospital is higher among healthy individuals, potentially reflecting an underlying correlation with income.

4 Identifying Gatekeeping

To identify insurer gatekeeping, I leverage exogenous variation in health care demand introduced by discontinuities in patient cost-sharing rules. In particular, I focus on the sample of individuals who reach their OOP maximum and face zero prices.

FIGURE 2: Probability of visiting a private hospital



Relative to individuals who pay a fraction of their health care cost through cost-sharing, gatekeeping incentives should be stronger among the group of patients who have their insurer completely cover the cost of care.

The strategy of using individuals who face zero prices to identify the magnitude of gatekeeping has several identification threats. The first is a selection bias problem: people who reach their OOP limit may be unobservably sicker and less responsive to prices compared to those who don't reach their limit. This type of unobserved heterogeneity might lead a researcher to underestimate the effects of insurer gatekeeping. The second is a confounding bias problem: changes in demand when individuals face zero prices may come from patients facing choice frictions, patients' health status worsening over time, or insurers steering patients towards cheaper care. These types of unobserved confounders might lead a researcher to overestimate the effects of gatekeeping.

I start by exploring the first source of bias to determine whether selection into reaching the OOP maximum is a concern in my setting. The descriptive evidence suggested that reaching the OOP limit is a sudden event and therefore there are no reasons to believe that individuals anticipate the event. If this is true, then peo-

ple who reach their OOP limit and those who don't should have similar utilization and spending patterns before the event. To characterize these spending patterns more systematically, I compare people who reach the OOP spending limit (treated group) and those who don't (control group), before and after they reach the limit in a differences-in-differences (*did*) event study framework.

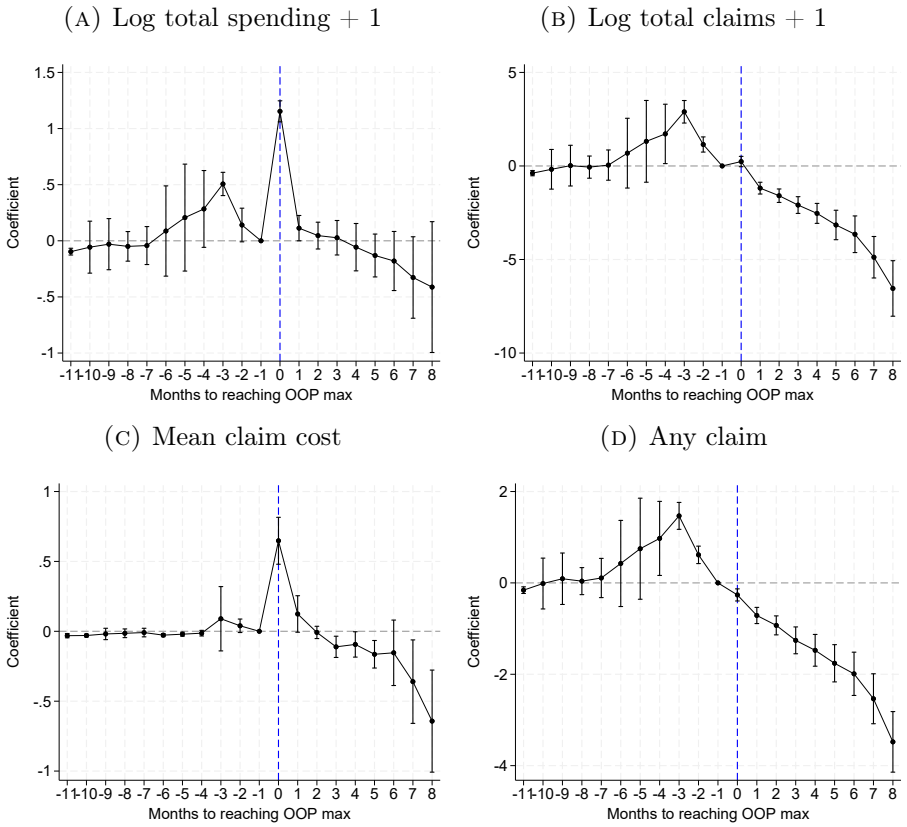
I exclude from my treated sample individuals who reach their OOP maximum during the first quarter of the year. For these individuals I do not observe pre-event periods and I cannot distinguish whether the health shock is random or whether reaching the OOP maximum was determined by their utilization or spending prior to the start of my sample period. The regression specification is as follows:

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

where y_{it} is the outcome of individual i in month t, t^* is the calendar month in which the treated individual reaches the OOP maximum; S_{it} is consumer i's cumulative out-of-pocket spending up to month t, which has differential effects by year y through θ_y; α_i is an individual fixed effect and γ_t is a calendar month fixed effect. The coefficients β_k measure the average treatment effect on the treated in month k relative to the month when individuals reach their OOP limit. For those in the control group, I normalize $k = -1$. Standard errors are clustered at the individual level, which defines the level of treatment. I use Callaway and Sant'Anna (2021)'s estimator to deal with staggered treatment, as different treated individuals may reach the OOP limit in different months. I use the never-treated units as controls.

Figure 3 shows event study coefficients and 95 percent confidence intervals using as outcome the log of total spending in panel A, the log of total claims in panel

FIGURE 3: Utilization and spending after reaching the OOP limit



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (1) for the log of total spending in panel A, log of total claims in panel B, mean claim cost in panel C, and an indicator for making claims in panel D. Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to −1 for the control group.

B, the mean claim cost in panel C, and an indicator for making claims in panel D. I find that trend differences in total spending and mean claim cost before the event between individuals who reach the OOP maximum and those who don't are negligible, suggestive of parallel pre-trends. More importantly, to the extent that these outcomes capture changes in health status, evidence of parallel trends before the event indicates limited selection into reaching the OOP maximum.

Consistent with the descriptive evidence, at the time of the event there is a sharp discontinuity in total spending and mean claim cost among people who reach the OOP limit. These individuals spend twice as much and claim services that are around 500

thousand pesos more expensive than individuals who do not reach their limit. The difference in total spending and mean claim cost becomes negative over time. Treated individuals spend half as much and make claims that are roughly 250 thousand pesos *cheaper* than controls 7 months after the event. The negative causal effect on spending may be the result of insurer gatekeeping or of information frictions regarding the shadow price of health care. I expand on the role of these factors later on.

While individuals who reach the OOP limit claim relatively cheaper services, their number of claims and likelihood of making claims does not increase and, in fact, decreases substantially after the event as seen in panels B and D of figure 3, respectively. At the time of the event, individuals who reach the OOP maximum are 100 percentage points less likely to make a claim. This difference increases over time, with treated individuals being 300 percentage points less likely to make claims and making 5 fewer claims between 7 to 8 months after reaching the OOP maximum. Appendix table 1 presents the associated event study coefficients and standard errors. These results are robust to excluding individuals who die during the sample period as seen in appendix table 5.

4.1 Heterogeneity Analysis

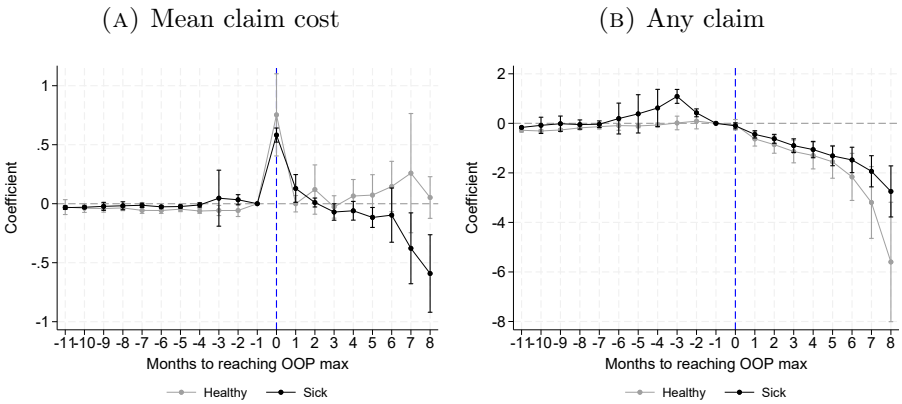
Differences in utilization and spending after reaching the OOP limit are driven by differences health status. Figure 4 reports *did* event study coefficients and 95 percent confidence intervals conditional on individuals who received a chronic disease diagnosis at any point during the year in black, and conditional on those who never received a diagnosis in gray.

Results in panel A show that healthy individuals who reach their OOP limit claim services that are as expensive as those claimed by consumers who don't reach their

spending limit, except during the month when the event occurs. Instead, consumers with chronic diseases claim services that are on average 450 thousand pesos cheaper than those claimed by the control group 7 months after reaching the OOP limit. The opposite patterns in mean claim cost between healthy and sick patients may reflect an underlying correlation with income.

Panel B shows that reductions in the probability of making claims are substantial among both healthy and sick consumers after reaching the OOP maximum. However, reductions are almost twice as large for the former than for the latter 8 months after the event. The fact that health care consumption falls by a greater magnitude among healthy individuals suggests that it may be easier for insurers to gatekeep this type of patient and, more generally, that steering incentives depend on the consumer's health status.

FIGURE 4: Utilization and spending by health status



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (1) for the mean claim cost in panel A, and an indicator for making claims in panel B. Estimates in black condition on individuals with a chronic disease diagnosis and those in gray condition on individuals without diagnosis. Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to −1 for the control group.

Treatment effects also vary across age and sex in ways consistent with a correlation with health status as seen in figure 5. Panels A and B present results conditional on the sample of individuals aged less than 65 in black, and aged 65 or older in gray.

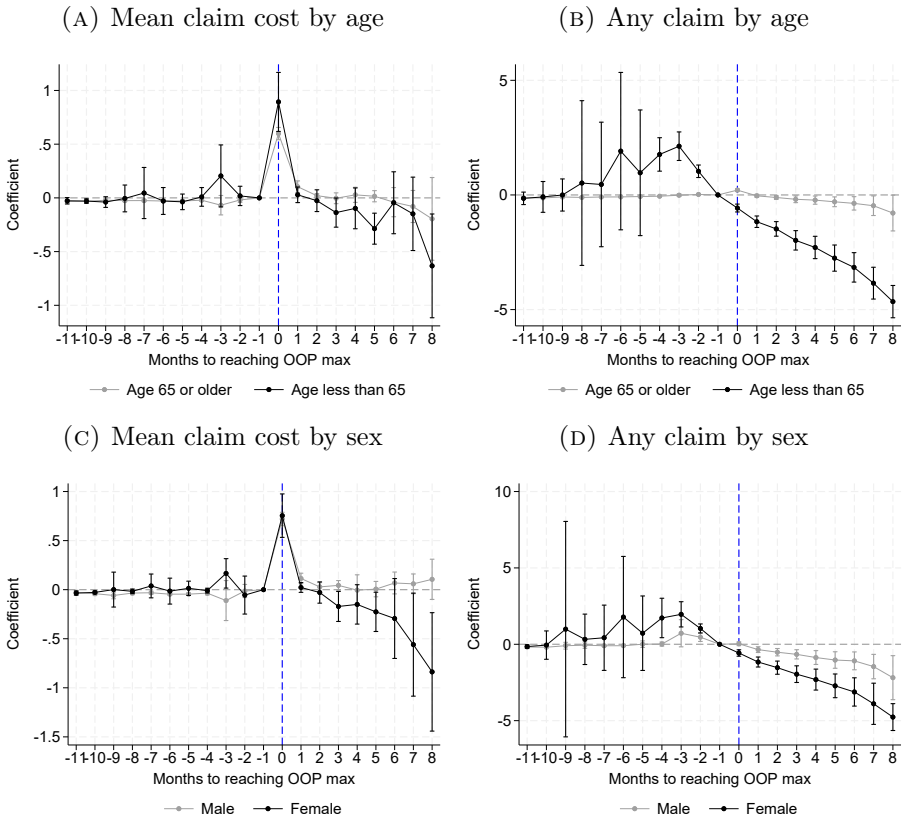
Panels C and D show results for females in black and for males in gray. I find that reductions in the mean claim cost and the probability of making claims after reaching the OOP maximum are larger for treated individuals aged less than 65 than for those aged 65 or older. In particular, the treatment effect on the probability of making claims for the former is more than three times larger than the effect for the latter. Panels C and D also show that gatekeeping or information frictions are stronger among females than males. Females in the treated group see reductions in the mean claim cost that are nearly twice as large as for treated males in the first 4 months after reaching the OOP maximum.

Effects on the mean claim cost and the probability of making claims are heterogeneous across insurers, suggestive of differences in gatekeeping efforts. Figure 6 reproduces the event study specification conditional on the set of three largest insurers (EPS010, EPS013, and EPS037) in panels A and B, and conditional on a set of smaller insurers (EPS005, EPS001, and EPS002) in panels C and D. Even when conditioning on each insurer, the figures provide evidence of parallel spending and utilization patterns before the event between treated and control individuals, which reinforces the idea that reaching the OOP maximum is a sudden event in my context.

The magnitude of the causal effect does differ across insurers. Panels A and C show that treated individuals enrolled with EPS013 claim substantially cheaper services 7 to 8 months after reaching the OOP maximum. For smaller insurers, such as EPS002 and EPS001, treated individuals consume services that are as expensive as those claimed by control units throughout the post-period. Panels B and D show that the probability of making claims falls substantially for EPS037 and EPS002 in line with findings in the full sample, but does not change for the rest of insurers after the event.

The results presented in this section provide strong evidence of dynamic incentives

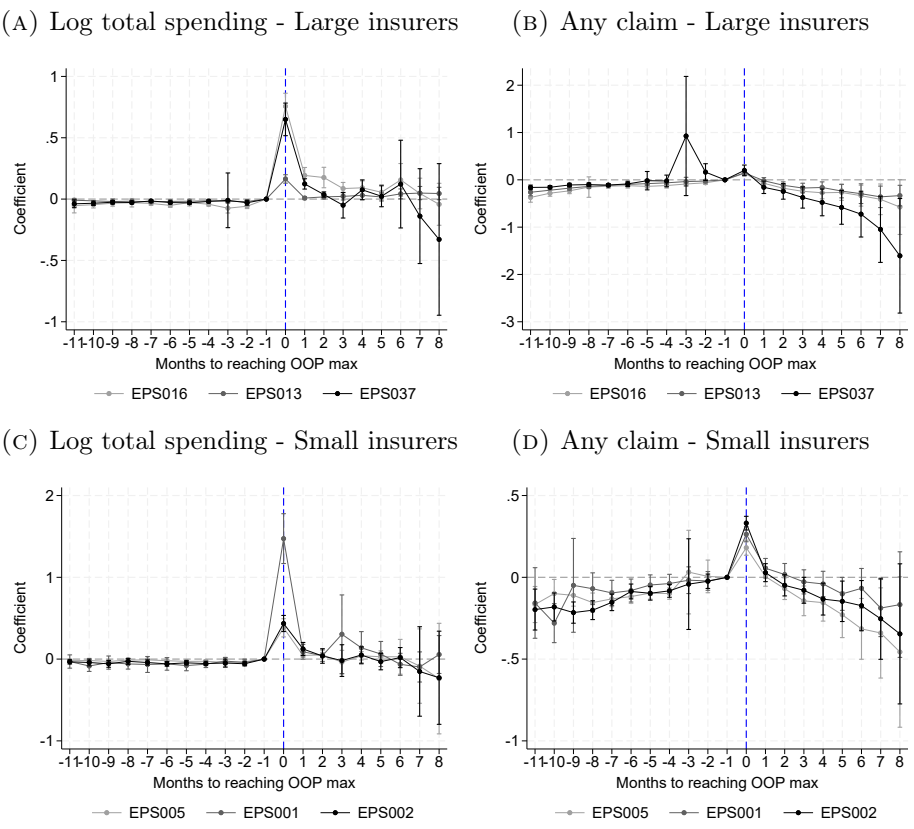
FIGURE 5: Utilization and spending by age and sex



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (1) for the mean claim cost in panels A and C, and an indicator for making claims in panels B and D. Estimates in black in panels A and B condition on individuals aged less than 65 and those in gray condition on individuals aged 65 or older. Estimates in black in panels C and D condition on females and those in gray condition on males. Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to −1 for the control group.

in the decision to consume and provide health care. Changes in the magnitude of these incentives are introduced by discontinuities in cost-sharing rules. For example, after reaching the OOP maximum, individuals face zero coinsurance rates, which may induce them to consume more expensive care. However, insurers also pay the full cost of care after enrollees hit their OOP limit, which may induce them gatekeep patients depending on their health status and demographic characteristics.

FIGURE 6: Utilization and spending by insurer



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (1) for the log of total spending in panels A and C, and an indicator for making claims in panels B and D. Each set of estimates restricts the treatment group to individuals enrolled with a particular insurer. Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to -1 for the control group.

4.2 Health Services

To see whether the reduction in spending stems from individuals choosing cheaper providers conditional on the service, I estimate event study designs for subsets of health services. To the extent that insurers' gatekeeping incentives vary with the type of health service that individuals claim, these exercises also provide evidence of the type of care that insurers are more likely to deter or steer. Figure 7 presents results of my event study specification using as outcomes the log of total spending and the likelihood of making claims for inpatient services in panels A and B, for

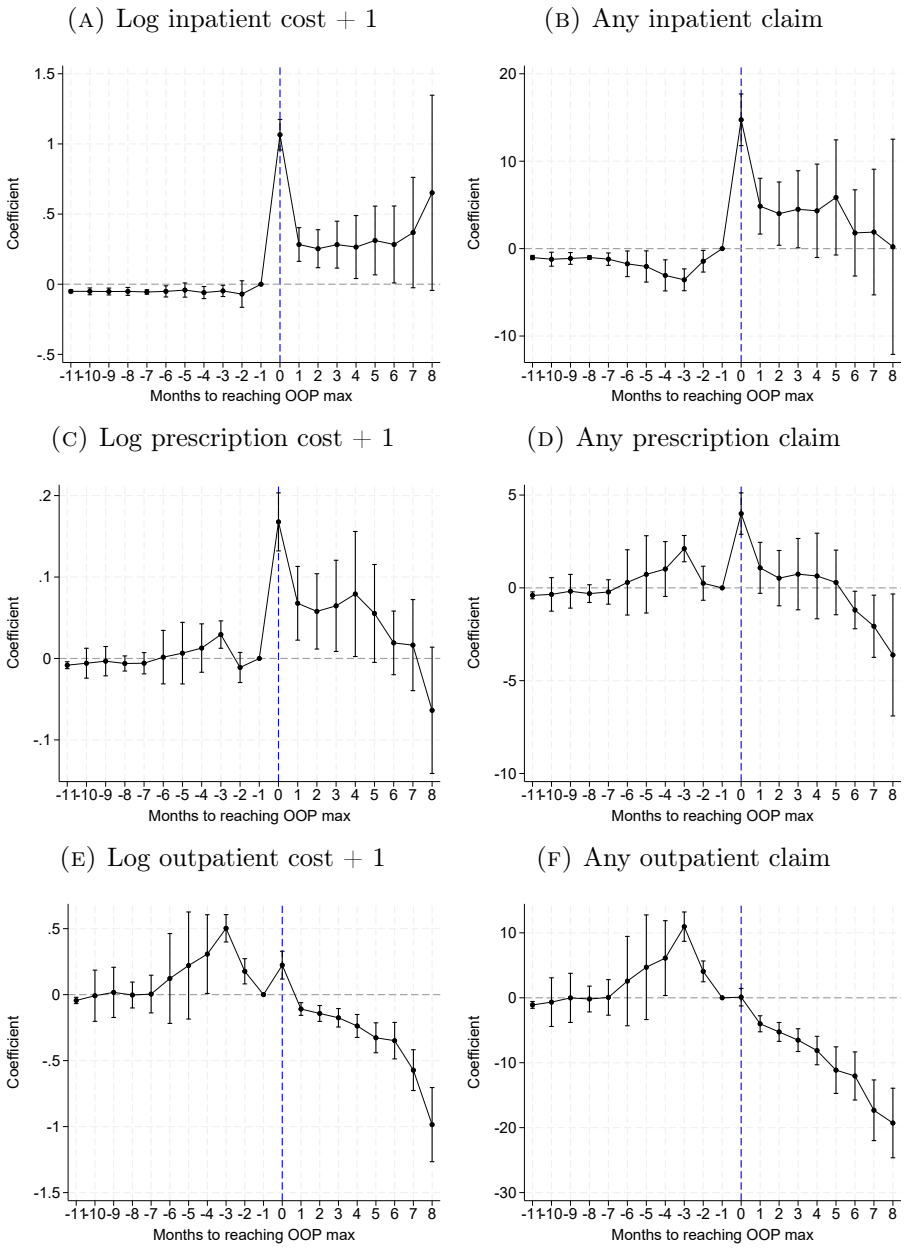
prescription medications in panels C and D, and for outpatient care in panels E and F.

Results are suggestive of gatekeeping efforts being smaller for more acute or necessary care. Panels A and B show that treated individuals spend twice as much as controls when reaching the OOP maximum. The treatment effect remains positive and grows in magnitude the closer individuals get to the end of the calendar year, which is opposite to the pattern in the full sample in figure 3. The likelihood of making inpatient claims is also 500 percentage points higher for treated individuals than for controls 1 month after the event. This treatment effect remains positive 5 months after the event, but becomes indistinct from zero after 6 months.

Panels C and D show qualitatively similar results for prescription medications. Individuals who reach their OOP maximum due to sudden health shocks are usually treated by their physician with medications together with other types of expensive care. This explains why treated individuals spend nearly 50 percent more than controls 1 month after the event, and why their likelihood of making prescription claims does not change relative to controls in the first 4 months after reaching the OOP maximum.

Because the health status of patients who reach their OOP maximum likely worsens after the event, we would expect them to consume more expensive services across all types of care. The evidence in panels A to D is consistent with this intuition, but consumption patterns for more discretionary care such as outpatient services in panels E and F present a different story. The panels show that treated individuals spend 50 percent less than controls in outpatient care between 5 to 7 months after reaching their OOP maximum. This difference increases over time, with outpatient spending being around a hundredth of that of controls.

FIGURE 7: Utilization and spending by type of care



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (1) for the log of inpatient cost in panel A, indicator for making inpatient claims in panel B, log of prescription cost in panel C, indicator for making prescription claims in panel D, log of outpatient cost in panel E, and indicator for making outpatient claims in panel F. Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to -1 for the control group.

Panel F also shows that the likelihood of making outpatient claims falls 500 percentage points for treated individuals relative to controls 2 months after reaching the OOP maximum. This reduction in the consumption of care that is less acute suggests that, despite their worsened health and their zero cost-sharing, treated individuals behave as if they face non-zero prices for outpatient care. Appendix figure 2 reports event study results for other types of potentially discretionary services such as imaging and laboratory tests.

4.3 Zero-Price and Health Shock Effects

The event studies in figure 3 generally suggest that after reaching the OOP spending limit, treated individuals differ systematically from controls. This difference conflates two effects at play: the effect on utilization and spending due to sudden hospitalizations ("health shock effect") and the effect due to facing zero prices after reaching the OOP maximum ("zero-price effect"). These effects are highly correlated in the sense that individuals face sudden hospitalizations and zero prices almost at the same time.

Since gatekeeping incentives may be different for people who face sudden hospitalizations versus those who face zero prices without having a hospitalization, it is important to disentangle the zero-price effect from the health shock effect. For instance, we might expect an insurer to be more willing to send patients to expensive providers when it has to cover the full cost of care if the patient is relatively sick compared to when the patient is relatively healthy. In this subsection I measure the relative magnitude of the zero price and the health shock effects, which will bound the potential impacts of insurer gatekeeping.

Let $t = 0$ denote the period before reaching the OOP maximum and $t = 1$ the period after reaching the maximum. Let $H(i, t)$ be an indicator for individual i having

a hospitalization in period t, $D(i,t)$ be the price of healthcare that individual i faces in period t, and $Y(i,t)$ be the mean claim cost of individual i in period t. The event study of mean claim cost in figure 3 identifies the average treatment effect on the treated (ATT) as

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

Underlying this treatment effect is the effect of sudden changes in health status caused primarily by hospitalizations. Both treated and control units may face hospitalizations in the pre- or post- periods. This implies that each element of the previous equation is a weighted average across hospitalization status, for example,

$$\begin{aligned}&E[Y(i,1)|D(i,1)=0] = \\ &\quad P(H(i,1)=1|D(i,1)=0)E[Y(i,1)|D(i,1)=0,H(i,1)=1] \\ &+P(H(i,1)=0|D(i,1)=0)E[Y(i,1)|D(i,1)=0,H(i,1)=0]\end{aligned}$$

Using this fact, we can rewrite the ATT as:

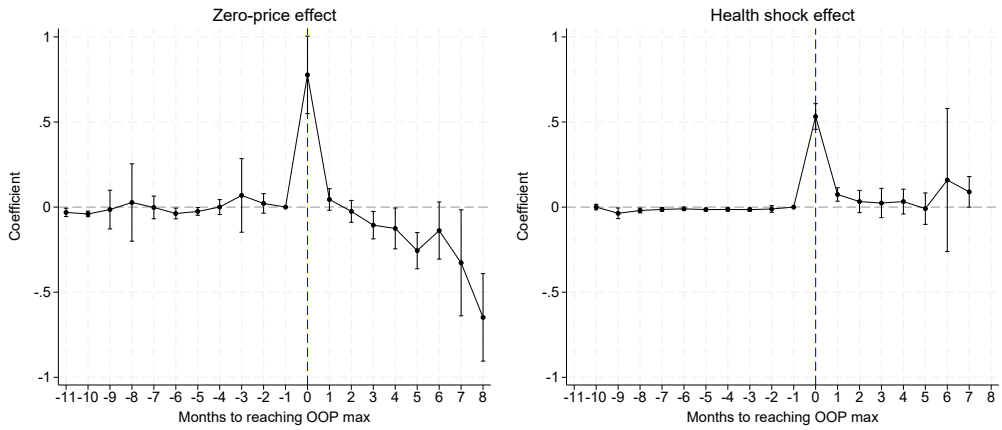
$$
\begin{aligned}
\beta = &\Big(E[Y(i,1)|D(i,1)=0,H(i,1)=0] - E[Y(i,1)|D(i,1)>0,H(i,1)=0]\Big) \\
&\underbrace{-\Big(E[Y(i,0)|D(i,1)=0,H(i,1)=0] - E[Y(i,0)|D(i,1)>0,H(i,1)=0]\Big)}_{\text{Zero-price effect}} \\
&+ x\Big(E[Y(i,1)|D(i,1)=0,H(i,1)=1] - E[Y(i,1)|D(i,1)=0,H(i,1)=0]\Big) \\
&- x\underbrace{\Big(E[Y(i,0)|D(i,1)=0,H(i,1)=1]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

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

The first part of the equation compares the mean claim cost of individuals who reach the OOP maximum and those who don't, conditional on not having a hospitalization, before and after reaching the OOP maximum. The second part of the equation compares individuals who have a hospitalization and those who don't, conditional on reaching the OOP maximum, before and after this event. The third part of the equation compares the mean claim cost of consumers who have a hospitalization and those who don't, conditional on not reaching the OOP maximum.

Results of this decomposition exercise for the mean claim cost are presented figure 8. The left panel depicts event study coefficients for the zero-price effect, that is, conditional on people who never have a hospitalization in the study period. The right panel depicts event study coefficients for the treated health shock effect, that

FIGURE 8: Effect Decomposition on the Treated



Note: Coefficients and 95 percent confidence intervals of the event study specifications for the mean claim cost due to the zero-price effect in the left panel and due to the health shock effect in the right panel. The zero-price effect uses the sample of individuals who are not hospitalized during the sample period. The health shock effect uses the sample of treated individuals. Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to -1 for the control group.

is, conditional on individuals who reach the OOP maximum. For the treated health shock effect, relative time indicators are the same as those for reaching the OOP maximum as the hospitalization happens when the individual reaches this limit.

Findings show that when the event occurs, the zero-price effect is 50 percent larger than the treated health shock effect. The effect on the mean claim cost from facing zero prices quickly disappears one month after the event and becomes negative thereafter. However, the health shock effect is positive up to 7 months after the event although estimates are noisy. The health shock effect goes in line with the intuition that individuals with poor health tend to be less price sensitive. But, the reduction in the zero-price effect and, in particular, the fact that it becomes negative, is irreconcilable with health care demand being perfectly inelastic after consumers reach their OOP limit. This suggests that factors other than consumers' OOP price may explain why health care demand responds to cost-sharing. I delve into these factors in the next subsection.

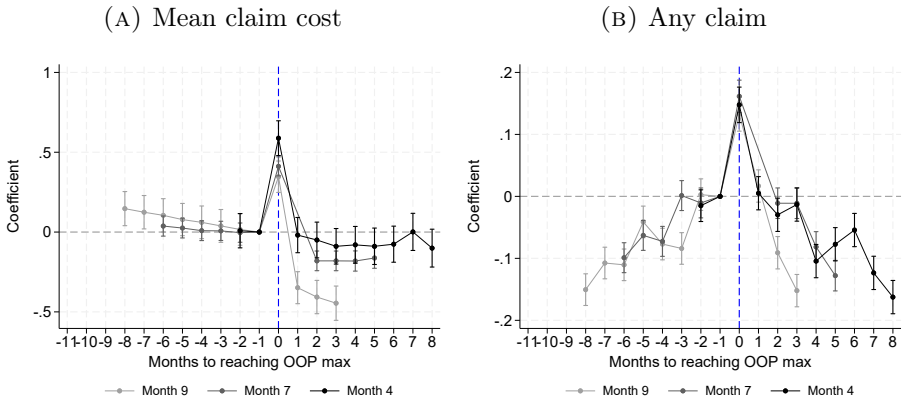
4.4 Dynamic incentives

After reaching the OOP maximum, consumers face zero prices because their coinsurance rate drops to zero. If consumers were forward-looking and faced no information frictions (such that they know exactly what their OOP prices are at every point in time), they would either consume more health care services or more expensive services after the event than before the event. Cumulative spending can also increase over time after reaching the OOP maximum if consumers foresee that prices will be non-zero at the start of the next calendar year when cost-sharing resets. This is apparent from panel A of figure 1 which shows that cumulative spending ramps up after reaching the OOP limit, in the last three months of the year.

However, when compared to individuals who do not reach the OOP maximum in an event study specification, the zero-price effect in the left panel of figure 8 is at odds with these behavioral assumptions about consumers when they face zero prices. The reduction in the zero-price effect is consistent with insurers steering patients towards cheaper providers and with gatekeeping incentives being stronger among the group of patients who reach the OOP maximum. Nonetheless, the finding is also consistent with patients facing information frictions. If consumers are uncertain about whether they have reached their OOP limit, they might behave as if they face non-zero prices after the event. If this type of information friction disappears over time, then we should see consumers either making more expensive claims or claiming more services the further they are from having reached their OOP maximum and the closer they are to the end of the calendar year.

To get at the role of information frictions in explaining the spending and utilization patterns in figure 3, I estimate separate event study specifications *conditional* on individuals who reach the OOP maximum in different months of the year (i.e., without

FIGURE 9: Utilization and spending by cohort



Note: Coefficients and 95 percent confidence intervals of the event study specifications following equation (1) for mean claim cost in panel A and an indicator for making claims in panel B conditional on treated individuals who reach their OOP maximum in September, July, and April.

a control group). Appendix table 7 presents the set of event study coefficients.

Findings in figure 9 show that treated individuals consume significantly cheaper services and are significantly less likely to make claims after reaching the OOP limit regardless of when the event happens. Comparing across months, we see that individuals who reach the OOP maximum in April consume relatively more expensive services by the end of the calendar year compared to those who reach their maximum in July, and this latter group in turn consumes relatively more expensive services than those who are treated in September. While this evidence is consistent with the presence of information frictions, the fact that estimates of the difference in mean claim cost are negative after the event for consumers who are treated in April suggests a role for insurer gatekeeping in my setting.

Results in figures 3 and 9 are in contrast to Brot-Goldberg, Chandra, Handel, and Kolstad (2017), who find no evidence of consumers price-shopping or of consumers responding to the true shadow price of care after they hit their deductible. Instead, figures 3 and 9 indicate that health care demand responds substantially to the shadow price of care after patients reach their OOP maximum and that this response increases

over time after the event. Importantly, the figures also suggest that insurer gatekeeping is an important –usually unobserved– source of variation in the responsiveness of health care demand to prices.

Although the event study analyses identify insurer gatekeeping and consumer information frictions as sources of price sensitivity, they do not speak to the mechanisms by which insurers gatekeep their enrollees. Given that insurers in Colombia cannot design their cost-sharing rules nor premiums, they can engage in steering through non-price mechanisms such as denying claims or requiring prior authorization for certain services or to visit certain providers.[1] The event study analyses do shed light on what the effects of gatekeeping are on the consumption of health care. Individuals make fewer claims and visit different providers after discontinuous changes in cost-sharing.

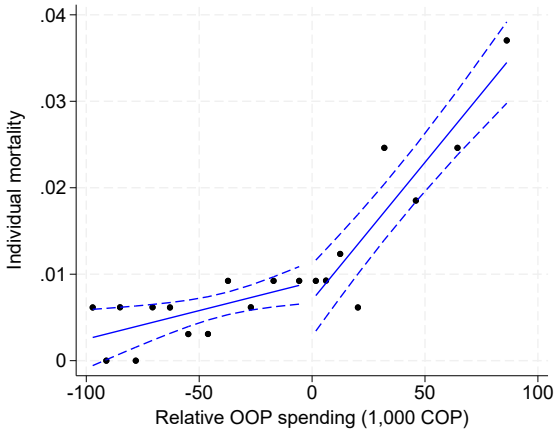
4.5 Health Outcomes

A few papers in the health economics literature have found evidence of a causal effect of cost-sharing on individual mortality. Chandra, Flack, and Obermeyer (2021) show for example that $100 reductions in patient drug budgets among the elderly increases mortality by 13.5 percent. Along a similar line, Buitrago et al. (2021) find that mortality increases by 0.8 per 1,000 after a three-fold increase in copayments using data from the Colombian health care system. Because insurer gatekeeping may potentially involve the provision of inadequate care and has similar effects on health care demand as cost-sharing, in this subsection I study the impact of gatekeeping on individual mortality.

My empirical approach is a regression discontinuity design (RD). I do not use

[1]The impact of prior authorizations has been the focus of other papers such as Brot-Goldberg et al. (2023).

FIGURE 10: Individual mortality



Note: Regression discontinuity plot for individual mortality. Linear regressions are estimated on vigintiles of OOP spending relative to the OOP maximum. Black dots correspond to average mortality in the bin, solid blue lines represent a linear fit, and dashed blue lines represent 95 percent confidence intervals.

a difference-in-differences specification because, by definition, the outcome does not vary before the event for treated individuals. Let y_i be an indicator for whether individual i dies in year t; S_i be the individual's relative OOP spending (cumulative OOP spending minus OOP maximum) by the end of the year or when she dies, whichever happens first; and $T_i = \mathbf{1}\{S_i - oop_i \geq 0\}$ be an indicator for whether the individual reaches her OOP maximum denoted by oop_i. I estimate the following regression in binned data based on vigintiles of S_i and for $S_i \in [-100, 100]$ thousand pesos:

$$y_i = \alpha T_i + \beta T_i \times S_i + \gamma S_i + \varepsilon_i$$

Results are presented in figure 10. The main takeaway is that individual mortality does not change around the cutoff. That is, patients are not more likely to die after reaching their OOP maximum as they are before reaching this limit. Appendix figure 3 presents RD plots to test for covariate smoothness around the cutoff. Although individuals who reach their OOP maximum are disproportionately older and are more likely to have a chronic disease, these discontinuities should bias the mortality effect

upwards.

5 The Cost of Gatekeeping and Information Frictions

The reduced-form findings of the previous sections provide evidence that insurer gatekeeping affects the price sensitivity of health care demand and impacts consumers' decisions of which providers to visit. An important caveat of such analysis is that it identifies changes in health care consumption due to changes in gatekeeping incentives, conditional on these incentives being at play along the distribution of cost-sharing. For example, if the magnitude of gatekeeping is determined by cost-sharing levels, then the event study analyses quantify the effect of an 11.5 percent increase in insurers' costs: for a low-income consumer, insurers cover 88.5 percent of health care costs before the individual reaches the OOP maximum and cover 100 percent after the consumer reaches this maximum. The exercises however cannot speak to what health care consumption or access to care would be in absence of gatekeeping nor can disentangle the relative magnitude of gatekeeping and information frictions.

In this section I develop a structural model of demand for hospitals that allows me to quantify the impact of gatekeeping and information frictions on access to health care. In the event studies, health care access was approximated by the likelihood of making claims. Here instead I measure access as changes in the value of distance to the hospitals that consumers choose. I am therefore interested in measuring how much farther do consumers have to travel to receive care because of insurer gatekeeping and because of information frictions.

Suppose consumer i of type θ lives in municipality m and is enrolled with insurer j.

The consumer chooses a hospital h in the network of her insurer based on her indirect utility in two states of the world, before and after reaching the OOP maximum:

$$u_{ijhm} = \begin{cases} \alpha_i r_i p_{jh} + \sum_{l \in m} q_{\theta l}(\tau + \sigma_\omega \omega_i) d_{lh} + \xi_h + \varepsilon_{ijhm} & \text{if } c_i + r_i p_{jh} + \nu_i \leq oop_i \\ \beta_i p_{jh} + \sum_{l \in m} q_{\theta l}(\tau + \sigma_\omega \omega_i) d_{lh} + \xi_h + e_{ijhm} & \text{o.w} \end{cases} \tag{2}$$

In this utility function, p_{jh} is the price that insurer j pays at hospital h for an admission and r_i is the coinsurance rate of a consumer type θ. $q_{\theta l}$ is the probability that the consumer lives in locality l within municipality m. This location probability is given by the population density at each locality. d_{lh} is the distance from locality l's centroid to hospital h, and $\omega_i \sim N(0,1)$ is a random normal shock to the bilateral distance between locality l and hospital h. The coefficients are given by $\alpha_i = x_i'\alpha$, $\beta_i = x_i'\beta$, where x_i is a vector of consumer demographics, diagnoses, and insurer. Moreover, ξ_h is a hospital fixed effect that captures shared preferences for hospital h across consumers.

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

$$\gamma_{ijh} = E[\mathbf{1}\{c_i + r_i p_{jh} + \nu_i \leq oop_i\}]$$

where, c_i is the OOP cost of consumer i up to but not including the hospital admission and oop_i is the OOP maximum. Price sensitivity in both states of the world depend also on the magnitude of information frictions. These frictions may arise either because the consumer does not know the true shadow price of health care or

because insurer j is uncertain about the patient's total OOP costs. I capture the impact of these information frictions on the probability of each state of world through $\nu_i \sim N(0, \sigma_\nu^2)$. This parameterization implies that

$$\gamma_{ijh} = \Phi\Big(\frac{oop_i - c_i - r_i p_{jh}}{\sigma_\nu}\Big)$$

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

Because I do not observe the patients' residence address but only their municipality of residence, I complement my enrollment and claims data with information on the distribution of population density by age across localities within a municipality. This information comes from the 2018 census. I limit my analysis to the 13 main municipalities in the country, for which locality level information exists. These municipalities have on average 14 localities, each with an average area of 128 squared kilometers. The third term on the right side of equation (2) therefore captures the expected distance to each hospital from each census tract conditional on the patient's age and municipality of residence.

Appendix 3 describes this census data by reporting maps of the 4 largest municipalities in my sample with their localities and the location of hospitals. In this appendix I also explain my methodology for obtaining admission prices given that prices observed in the claims data may sometimes vary with admission characteristics that are unobserved when insurers and hospitals negotiate.

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

$$L = \sum_i \Big(\gamma_{ijh} \log\Big(\prod_{h \in H_j} P_{ijhm}^{1^{y_{ijhm}}}\Big) + (1 - \gamma_{ijh}) \log\Big(\prod_{h \in H_j} P_{ijhm}^{2^{y_{ijhm}}}\Big)\Big)$$

where

$$P^s_{ijhm} = \int \frac{\exp(\delta^s_{ijhm})}{\sum_k \exp(\delta^s_{ijkm})} \phi(\omega) \quad \text{for } s = \{1, 2\} \tag{3}$$

and

$$\delta^1_{ijkm} = \alpha_i r_i p_{jh} + \sum_{l \in m} q_{\theta l}(\tau + \sigma_\omega \omega_i) d_{lh} + \xi_h$$

$$\delta^2_{ijkm} = \beta_i p_{jh} + \sum_{l \in m} q_{\theta l}(\tau + \sigma_\omega \omega_i) d_{lh} + \xi_h$$

Identification. To separately identify the coefficients associated with admission prices in the two states of world, α_i and β_i, I use the discontinuity in coinsurance rates introduced by the OOP maximum. Before reaching this maximum, consumers face prices up to the coinsurance rate, but afterwards out-of-pocket prices are zero and demand responds to prices only to the extent that insurers cover the full price of the admission. Price variation within hospital and coinsurance rate variation across patients are therefore needed to identify the price coefficients. While it is reasonable to think that consumers who reach their OOP maximum may differ in unobserved ways from those who don't, event study results from the previous sections provide evidence of limited selection bias of this style.

I identify the impact of information frictions captured by σ_ν from variation in cumulative OOP spending up to but not including the hospital admission. Identification essentially requires that, conditional on reaching the OOP maximum, choices differ across individuals beyond what can be explained by admission prices. Given that some individuals reach their OOP maximum before they make their choice of hospital, comparing their choices against those made by individuals who will reach the maximum right after the admission, identifies this parameter.

Finally, identification of τ and σ_ω requires observing two individuals who are

identical in terms of age, income group, municipality of residence, and insurer, who choose different hospitals.

Estimates. I estimate the hospital demand model using simulated maximum likelihood to approximate the integrals in equation (3). Results are presented in table 3. Consistent with the reduced form evidence I find that hospital demand responds to prices before and after consumers reach their OOP maximum. Before reaching this maximum, findings show that a 10,000 pesos increase in OOP prices reduces the probability of choosing a hospital by 6.78 percent. After reaching the maximum, a 10,000 pesos increase in admission prices reduces the choice probability by 1.57 percent.

Because OOP prices are zero after patients reach their OOP maximum, price sensitivity in this state of the world can be explained by insurer gatekeeping. Reassuringly, the price coefficient is smaller in magnitude than the OOP price coefficient, because the former captures only a change in gatekeeping incentives from covering 88.5 to 100 percent of an (low-income) individual's health care cost, but does not capture the extensive margin effect of gatekeeping.

Interactions of prices with consumer demographics and diagnoses are in line with intuition. Younger patients are much more responsive to OOP prices than older patients, while those with chronic disease are significantly less price sensitive than patients without diagnoses. Findings also show that gatekeeping incentives are stronger among older individuals who are potentially more expensive to the insurer. But insurers are less likely to gatekeep claims from individuals with chronic diseases as they are to gatekeep healthy individuals consistent with evidence from panel B of figure 4.

In line with previous literature (e.g., Ho, 2006), I find that patients dislike commuting. If a patient has to travel one additional kilometer to visit a hospital, her probability of choosing this hospital decreases by over 40 percent. This marginal

disutility from commuting is relatively homogeneous across consumers as seen by my estimate for σ_ω which is indistinct from zero. The model instead rationalizes variation in choice probabilities in the two states of world, conditional on prices, with findings of significant information frictions. My estimate for the standard deviation of the distribution of relative OOP spending is large and significant.

TABLE 3: Hospital demand estimates

		coef	se
OOP price		-6.782	(0.401)
Price		-1.567	(0.110)
Distance (mean)		-0.469	(0.048)
Distance (sd)		0.290	(0.109)
State prob sd		1.721	(0.040)
Interactions			
OOP price	Male	0.681	(0.040)
	Age 10-19	-26.03	(1.632)
	Age 20-29	-14.29	(0.933)
	Age 30-39	-13.13	(0.981)
	Age 40-49	-2.007	(0.155)
	Age 50-59	-8.796	(0.695)
	Age 60-69	-9.602	(0.826)
	Age 70 or older	(ref)	
	Sick	7.138	(0.654)
Price	Male	0.670	(0.048)
	Age 10-19	1.264	(0.090)
	Age 20-29	0.275	(0.019)
	Age 30-39	0.244	(0.017)
	Age 40-49	0.033	(0.003)
	Age 50-59	0.604	(0.048)
	Age 60-69	0.585	(0.051)
	Age 70 or older	(ref)	
	Sick	1.089	(0.100)
Observations		548,481	

Note: Table shows simulated maximum likelihood estimates of the hospital demand model. Prices are measured in millions of COP. Distance is measured in kilometers. Bootstrap standard errors in parenthesis based on 80 resamples.

6 Partial Equilibrium Analysis

I use my model estimates to conduct two partial equilibrium analyses that reveal the relative importance of gatekeeping and information frictions on access to care. First, to quantify the extensive margin effect of gatekeeping on commuting I set $\beta_i = 0$. Second, to quantify the impact of information frictions, I set $\sigma_\nu = 0$. In each scenario, I recompute individual's choice probabilities and present summary statistics of monetized marginal effects of distance and price elasticities. The monetized marginal effect is:

$$\gamma_{ijhm}\Big(\frac{1}{\alpha_i}\frac{\partial P^1_{ijhm}}{\partial \sum_{l\in m} q_{\theta l} d_{lh}}\Big) + (1-\gamma_{ijhm})\Big(\frac{1}{\alpha_i}\frac{\partial P^2_{ijhm}}{\partial \sum_{l\in m} q_{\theta l} d_{lh}}\Big)$$

which can be interpreted as patient willingness-to-pay to reduce commuting distance. Let $\tilde{P}_{ijhm} = \gamma_{ijhm}P^1_{ijhm} + (1-\gamma_{ijhm})P^2_{ijhm}$, price elasticities are calculated as:

$$\Big(\gamma_{ijhm}\frac{\partial P^1_{ijhm}}{\partial p_{jh}} + (1-\gamma_{ijhm})\frac{\partial P^2_{ijhm}}{\partial p_{jh}} + P^1_{ijhm}\frac{\partial \gamma_{ijhm}}{\partial p_{jh}} - P^2_{ijhm}\frac{\partial \gamma_{ijhm}}{\partial p_{jh}}\Big)\frac{p_{jh}}{\tilde{P}_{ijhm}}$$

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

Table 4 presents the mean, 25th, and 75th percentiles of the distribution of monetized marginal effect of distance under the observed scenario and the exercises without gatekeeping and without information frictions. Note that if it weren't for gatekeeping, consumers should choose hospitals only based on distance and quality in the event they reach their OOP maximum. Consistent with this intuition, I find that without gatekeeping consumers on average would be willing to pay 2 percent more than in the observed scenario to reduce commuting distance by 1 kilometer. For the consumer in the 75th percentile of the distribution this effect equals nearly 5 percent.

To put these estimates in perspective, the price of a bus ticket in Bogotá during 2011 was 1,700 pesos (for a Transmilenio bus ride), hence gatekeeping induces individuals to pay 34 additional pesos to visit a hospital. If patients pay 1,700 pesos to commute the average distance in my data (8.96 kilometers), then my partial equilibrium findings also suggest that gatekeeping has consumers travel on average 2.5 additional blocks to visit a hospital.

TABLE 4: Monetized marginal effect of distance

	p25	mean	p75
Observed	-2,239.1	-1,014.8	-221.5
No gatekeeping	-2,285.9	-1,033.9	-231.9
No information frictions	-2,275.8	-1,027.7	-208.8

Note: Table shows the mean, 25th, and 75th percentiles of the distribution of the monetized marginal effect of distance in the observed scenario, the exercise without gatekeeping setting $\beta_i = 0$, and the exercise without information frictions setting $\sigma_\nu = 0$. Measured in 2011 COP.

When I eliminate information frictions, results in table 4 show that patients on average would be willing to pay 1.3 percent more than the observed scenario to reduce commuting distance by 1 kilometer. This effect is quite heterogeneous across patients as the consumer in the 75th percentile of the distribution would be willing to pay 6 percent less than in the observed scenario. Without information frictions, 56 percent of potential choices would have consumers reach their OOP maximum with certainty and their insurer cover the full cost of health care. For this set of potential choices, the monetized marginal disutility of distance increases because insurers are less price sensitive than consumers (α_i is smaller in the state of the world where patients reach their OOP maximum).

In table 5 I present summary statistics of the distribution of own-price elasticities in each scenario. In line with the intuition and reduced-form findings, gatekeeping induces additional price sensitivity in demand. When I eliminate gatekeeping the average own-price elasticity falls 37 percent (in absolute value) relative to the observed

scenario. These estimates highlight the importance of accounting for gatekeeping when making welfare calculations or policy predictions in health insurance markets. Gatekeeping is an important, usually unobserved variable that is correlated with out-of-pocket prices, hence estimation of hospital demand functions requires instruments otherwise estimates may reflect relatively inelastic hospital demand curves.

In the scenario without information frictions, I find instead that hospital demand becomes more price elastic. The average own-price elasticity increases 12 percent (in absolute value) relative to the observed scenario. This finding stems from the fact that eliminating information frictions exacerbates the effects of gatekeeping. Insurers steer demand away from the 56 percent of potential choices that would have consumers reach their OOP maximum with certainty.

TABLE 5: Own-Price elasticity

	p25	mean	p75
Observed	-0.839	-0.555	-0.141
No gatekeeping	-0.514	-0.351	-0.086
No information frictions	-0.930	-0.620	-0.098

Note: Table shows the mean, 25th, and 75th percentiles of the distribution of own-price elasticities in the observed scenario, the exercise without gatekeeping setting $\beta_i = 0$, and the exercise without information frictions setting $\sigma_\nu = 0$.

7 Conclusions

In this paper I show that health insurers play an important role in containing rising health care spending by engaging in gatekeeping practices. Gatekeeping has stronger effects on health care utilization and spending compared to cost-sharing, which provides an argument in favor of health systems that provide free health insurance through private insurers.

To identify the impact of gatekeeping on utilization and spending I leverage the

discontinuity in coinsurance rates introduced by the out-of-pocket (OOP) maximum. I use data from the Colombian health care system where cost-sharing rules are determined by the government and standardized across insurers and hospitals. I show that patients in my sample reach their OOP maximum as-if-randomly. I find that those who reach their maximum consume significantly cheaper services and are substantially less likely to make claims afterwards. These results are at odds with behavioral assumptions about consumers when they face zero prices and are in favor of gatekeeping and information frictions driving consumer choices. Estimates from a structural model of hospital demand show that without gatekeeping consumers would be willing to pay 2 percent more to reduce commuting distance to hospitals by 1 kilometer.

Importantly, I find no evidence of gatekeeping significantly impacting patient health outcomes such as mortality. In the discussion of how best to deliver universal health insurance coverage, these findings suggest a role for private insurers as buffers of patient moral hazard beyond cost-sharing. However, recent media attention to cases where gatekeeping has prevented patients from receiving adequate care for their chronic health conditions in the United States, indicate that government regulation of gatekeeping is needed. Future research may study counterfactual policies that limit gatekeeping for certain groups of patients.

References

Aron-Dine, A., L. Einav, and A. Finkelstein (2013): "The RAND Health Insurance Experiment, Three Decades Later," *Journal of Economic Perspectives*, 27, 197–222.

Baicker, K., S. Mullainathan, and J. Schwartzstein (2015): "Behavioral

Hazard in Health Insurance," *The Quarterly Journal of Economics*, 130, 1623–1667.

Brot-Goldberg, Z. C., S. Burn, T. Layton, and B. Vabson (2023): "Rationing Medicine Through Bureaucracy: Authorization Restrictions in Medicare," Tech. rep., National Bureau of Economic Research.

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

Buitrago, G., G. Miller, and M. Vera-Hernández (2021): "Cost-Sharing in Medical Care Can Increase Adult Mortality Risk in Lower-Income Countries," *medRxiv*, 2021–03.

Callaway, B. and P. H. Sant'Anna (2021): "Difference-in-Differences With Multiple Time Periods," *Journal of econometrics*, 225, 200–230.

Chandra, A., E. Flack, and Z. Obermeyer (2021): "The Health Costs of Cost-Sharing," .

Eichner, M. J. (1998): "The Demand for Medical Care: What People Pay Does Matter," *The American Economic Review*, 88, 117–121.

Finkelstein, A. and R. McKnight (2008): "What did Medicare do? The Initial Impact of Medicare on Mortality and Out of Pocket Medical Spending," *Journal of Public Economics*, 92, 1644–1668.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group (2012): "The Ore-

gon Health Insurance Experiment: Evidence from the First Year," *The Quarterly Journal of Economics*, 127, 1057–1106.

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

Iizuka, T. and H. Shigeoka (2022): "Is Zero a Special Price? Evidence from Child Health Care," *American Economic Journal: Applied Economics*, 14, 381–410.

League, R. (2022): "Administrative Burden and Consolidation in Health Care: Evidence from Medicare Contractor Transitions," .

Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, and A. Leibowitz (1987): "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *The American Economic Review*, 251–277.

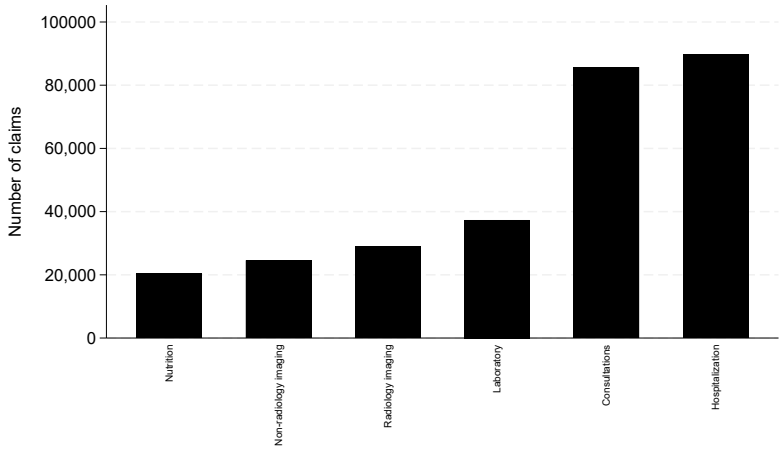
Newhouse, J. P. (1993): *Free for All?: Lessons from the RAND Health Insurance Experiment*, Harvard University Press.

Serna, N. (2021): "Cost Sharing and the Demand for Health Services in a Regulated Market," *Health Economics*, 30, 1259–1275.

Shi, M. (2023): "Monitoring for Waste: Evidence from Medicare Audits," *The Quarterly Journal of Economics*.

Appendix 1 Additional descriptives

Appendix Figure 1: Most expensive types of claims



Note: Figure shows the frequency of the top 6 most expensive types of services claimed by individuals who reach their OOP maximum in the week prior to reaching this limit.

Appendix 2 Event Study Coefficients

In this appendix I present event study coefficients and standard errors used to construct each figure in the main text. I also report additional event study results for imaging and laboratory claims, as well as regression discontinuity graphs to test for covariate smoothness around the OOP maximum related to my mortality analysis.

Appendix Table 1: Main Event Study Coefficients

	Log total spending + 1		Log total claims + 1		Mean claim cost		Any claim	
	coef	se	coef	se	coef	se	coef	se
t-11	-0.097	(0.015)	-0.384	(0.072)	-0.031	(0.008)	-0.159	(0.037)
t-10	-0.057	(0.118)	-0.178	(0.541)	-0.030	(0.005)	-0.014	(0.284)
t-9	-0.030	(0.116)	0.018	(0.555)	-0.019	(0.021)	0.091	(0.287)
t-8	-0.050	(0.067)	-0.061	(0.302)	-0.014	(0.016)	0.039	(0.151)
t-7	-0.042	(0.086)	0.047	(0.412)	-0.009	(0.016)	0.107	(0.219)
t-6	0.087	(0.205)	0.685	(0.951)	-0.027	(0.006)	0.425	(0.481)
t-5	0.207	(0.243)	1.316	(1.115)	-0.021	(0.007)	0.748	(0.564)
t-4	0.284	(0.175)	1.713	(0.810)	-0.013	(0.010)	0.974	(0.414)
t-3	0.506	(0.053)	2.896	(0.307)	0.090	(0.117)	1.467	(0.150)
t-2	0.141	(0.076)	1.148	(0.207)	0.039	(0.025)	0.614	(0.097)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	1.153	(0.048)	0.236	0.141	0.648	(0.086)	-0.263	(0.068)
t+1	0.113	(0.058)	-1.185	(0.161)	0.124	(0.067)	-0.713	(0.090)
t+2	0.046	(0.061)	-1.590	(0.184)	-0.008	(0.022)	-0.930	(0.106)
t+3	0.027	(0.078)	-2.089	(0.228)	-0.111	(0.039)	-1.258	(0.150)
t+4	-0.057	(0.108)	-2.535	(0.272)	-0.094	(0.046)	-1.476	(0.177)
t+5	-0.131	(0.097)	-3.155	(0.402)	-0.164	(0.050)	-1.759	(0.208)
t+6	-0.180	(0.134)	-3.651	(0.500)	-0.154	(0.119)	-1.990	(0.242)
t+7	-0.327	(0.185)	-4.879	(0.566)	-0.360	(0.152)	-2.536	(0.279)
t+8	-0.412	(0.297)	-6.548	(0.757)	-0.643	(0.186)	-3.480	(0.338)
Observations	7,154,832		7,154,832		7,154,832		7,154,832	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (1) for the log of total spending, log of total claims, mean claim cost, and an indicator for making claims, using Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to -1 for the control group.

Appendix Table 2: Event Study Coefficients by Health Status

	Mean claim cost				Any claim			
	(1) Healthy		(2) Sick		(3) Healthy		(4) Sick	
	coef	se	coef	se	coef	se	coef	se
t-11	-0.029	(0.032)	-0.033	(0.008)	-0.280	(0.038)	-0.164	(0.020)
t-10	-0.038	(0.018)	-0.030	(0.006)	-0.309	(0.027)	-0.081	(0.167)
t-9	-0.041	(0.016)	-0.022	(0.017)	-0.267	(0.026)	-0.015	(0.158)
t-8	-0.036	(0.012)	-0.017	(0.017)	-0.183	(0.038)	-0.049	(0.095)
t-7	-0.057	(0.013)	-0.013	(0.010)	-0.133	(0.050)	-0.035	(0.068)
t-6	-0.058	(0.013)	-0.026	(0.006)	-0.090	(0.097)	0.194	(0.318)
t-5	-0.046	(0.011)	-0.024	(0.007)	-0.105	(0.037)	0.383	(0.394)
t-4	-0.062	(0.010)	-0.011	(0.010)	-0.068	(0.042)	0.620	(0.384)
t-3	-0.058	(0.022)	0.047	(0.121)	0.015	(0.140)	1.088	(0.143)
t-2	-0.057	(0.026)	0.033	(0.022)	0.090	(0.160)	0.424	(0.081)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	0.753	(0.178)	0.582	(0.030)	-0.057	(0.108)	-0.101	(0.062)
t+1	0.000	(0.035)	0.129	(0.060)	-0.628	(0.147)	-0.445	(0.075)
t+2	0.120	(0.107)	0.010	(0.019)	-0.860	(0.177)	-0.624	(0.092)
t+3	-0.027	(0.048)	-0.070	(0.036)	-1.137	(0.234)	-0.900	(0.140)
t+4	0.065	(0.071)	-0.060	(0.041)	-1.299	(0.276)	-1.059	(0.169)
t+5	0.073	(0.088)	-0.116	(0.043)	-1.555	(0.337)	-1.315	(0.203)
t+6	0.146	(0.109)	-0.097	(0.117)	-2.162	(0.485)	-1.478	(0.262)
t+7	0.259	(0.258)	-0.378	(0.153)	-3.196	(0.739)	-1.936	(0.321)
t+8	0.052	(0.090)	-0.592	(0.168)	-5.597	(1.230)	-2.749	(0.526)
Observations	3,542,436		3,612,396		3,542,436		3,612,396	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (1) for mean claim cost and an indicator for making claims conditional on individuals without diagnoses in columns (1) and (3) and conditional on individuals with chronic diseases in columns (2) and (4). Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to -1 for the control group.

Appendix Table 3: Event Study Coefficients by Age

	Mean claim cost				Any claim			
	(1) Age<65		(2) Age>=65		(3) Age<65		(4) Age>=65	
	coef	se	coef	se	coef	se	coef	se
t-11	-0.027	(0.015)	-0.032	(0.008)	-0.148	(0.139)	-0.125	(0.022)
t-10	-0.030	(0.011)	-0.031	(0.007)	-0.086	(0.343)	-0.130	(0.015)
t-9	-0.039	(0.025)	-0.022	(0.006)	-0.002	(0.357)	-0.088	(0.013)
t-8	-0.005	(0.064)	-0.025	(0.005)	0.520	(1.833)	-0.108	(0.011)
t-7	0.045	(0.121)	-0.025	(0.005)	0.458	(1.386)	-0.089	(0.011)
t-6	-0.029	(0.064)	-0.029	(0.005)	1.912	(1.751)	-0.085	(0.009)
t-5	-0.037	(0.037)	-0.028	(0.004)	0.967	(1.398)	-0.069	(0.010)
t-4	0.009	(0.044)	-0.022	(0.005)	1.767	(0.372)	-0.055	(0.014)
t-3	0.204	(0.147)	-0.068	(0.046)	2.126	(0.318)	-0.011	(0.030)
t-2	0.018	(0.046)	-0.019	(0.012)	1.033	(0.139)	0.022	(0.024)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	0.893	(0.140)	0.598	(0.029)	-0.570	(0.087)	0.211	(0.032)
t+1	0.029	(0.037)	0.104	(0.028)	-1.167	(0.128)	-0.038	(0.034)
t+2	-0.026	(0.052)	0.021	(0.010)	-1.481	(0.162)	-0.111	(0.043)
t+3	-0.137	(0.069)	-0.004	(0.026)	-1.980	(0.216)	-0.189	(0.075)
t+4	-0.098	(0.097)	0.028	(0.036)	-2.291	(0.249)	-0.225	(0.091)
t+5	-0.287	(0.074)	0.015	(0.027)	-2.755	(0.291)	-0.298	(0.109)
t+6	-0.045	(0.147)	-0.040	(0.069)	-3.159	(0.327)	-0.367	(0.149)
t+7	-0.149	(0.174)	-0.080	(0.076)	-3.847	(0.354)	-0.470	(0.218)
t+8	-0.633	(0.246)	-0.195	(0.196)	-4.653	(0.359)	-0.789	(0.399)
Observations	5,938,236		1,216,596		5,938,236		1,216,596	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (1) for mean claim cost and an indicator for making claims conditional on individuals aged less than 65 in columns (1) and (3) and conditional on individuals aged 65 or older in columns (2) and (4). Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to -1 for the control group.

Appendix Table 4: Event Study Coefficients by Sex

	Mean claim cost				Any claim			
	(1) Male		(2) Female		(3) Male		(4) Female	
	coef	se	coef	se	coef	se	coef	se
t-11	-0.028	(0.011)	-0.035	(0.011)	-0.176	(0.022)	-0.149	(0.058)
t-10	-0.041	(0.010)	-0.028	(0.008)	-0.191	(0.017)	-0.054	(0.473)
t-9	-0.059	(0.017)	0.001	(0.091)	-0.093	(0.113)	0.990	(3.598)
t-8	-0.035	(0.009)	-0.017	(0.012)	-0.058	(0.105)	0.325	(0.842)
t-7	-0.031	(0.008)	0.038	(0.062)	-0.096	(0.030)	0.428	(1.090)
t-6	-0.046	(0.007)	-0.015	(0.067)	-0.094	(0.021)	1.782	(2.026)
t-5	-0.046	(0.007)	0.014	(0.037)	0.006	(0.110)	0.725	(1.244)
t-4	-0.035	(0.009)	-0.010	(0.013)	0.016	(0.068)	1.723	(0.659)
t-3	-0.111	(0.104)	0.166	(0.076)	0.722	(0.454)	1.961	(0.421)
t-2	-0.014	(0.034)	-0.055	(0.099)	0.465	(0.141)	1.035	(0.151)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	0.720	(0.035)	0.754	(0.113)	0.034	(0.064)	-0.571	(0.107)
t+1	0.114	(0.027)	0.022	(0.025)	-0.350	(0.096)	-1.161	(0.167)
t+2	0.028	(0.022)	-0.029	(0.055)	-0.519	(0.123)	-1.531	(0.220)
t+3	0.044	(0.025)	-0.171	(0.078)	-0.659	(0.154)	-1.956	(0.278)
t+4	-0.005	(0.080)	-0.150	(0.102)	-0.868	(0.230)	-2.311	(0.350)
t+5	0.006	(0.040)	-0.226	(0.102)	-1.030	(0.276)	-2.721	(0.396)
t+6	0.068	(0.057)	-0.294	(0.208)	-1.086	(0.298)	-3.121	(0.472)
t+7	0.059	(0.051)	-0.560	(0.268)	-1.460	(0.408)	-3.896	(0.688)
t+8	0.106	(0.105)	-0.838	(0.308)	-2.187	(0.735)	-4.766	(0.448)
Observations	3,406,044		3,748,788		3,406,044		3,748,788	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (1) for mean claim cost and an indicator for making claims conditional on males in columns (1) and (3) and conditional on females in columns (2) and (4). Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to -1 for the control group.

Appendix Table 5: Event Study Coefficients Excluding Deaths

	Mean claim cost		Any claim	
	coef	se	coef	se
t-11	-0.026	(0.010)	-0.133	(0.064)
t-10	-0.029	(0.007)	0.152	(0.570)
t-9	-0.003	(0.040)	0.193	(0.416)
t-8	-0.009	(0.026)	0.106	(0.313)
t-7	-0.010	(0.024)	0.264	(0.478)
t-6	-0.022	(0.007)	0.700	(0.760)
t-5	-0.017	(0.006)	0.603	(0.553)
t-4	-0.015	(0.008)	0.965	(0.411)
t-3	0.236	(0.061)	1.488	(0.165)
t-2	0.007	(0.040)	0.607	(0.136)
t-1	(ref)	(ref)	(ref)	(ref)
t	0.786	(0.121)	-0.279	(0.073)
t+1	0.015	(0.018)	-0.689	(0.097)
t+2	-0.025	(0.019)	-0.925	(0.123)
t+3	-0.102	(0.044)	-1.228	(0.163)
t+4	-0.136	(0.054)	-1.419	(0.192)
t+5	-0.198	(0.063)	-1.701	(0.228)
t+6	-0.060	(0.044)	-1.906	(0.247)
t+7	-0.154	(0.144)	-2.545	(0.317)
t+8	-0.465	(0.188)	-3.445	(0.378)
Observations	7,154,832		7,154,832	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (1) for mean claim cost and an indicator for making claims excluding individuals who die during the sample period. Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to -1 for the control group.

Appendix Table 6: Zero-price and Health Shock Effects Event Study Coefficients

	Zero-price		Health Shock	
	coef	se	coef	se
t-11	-0.031	(0.013)	—	
t-10	-0.040	(0.008)	0.000	(0.008)
t-9	-0.014	(0.058)	-0.036	(0.016)
t-8	0.027	(0.116)	-0.020	(0.007)
t-7	-0.002	(0.034)	-0.013	(0.006)
t-6	-0.038	(0.016)	-0.010	(0.004)
t-5	-0.025	(0.012)	-0.015	(0.004)
t-4	0.000	(0.022)	-0.014	(0.005)
t-3	0.069	(0.110)	-0.015	(0.005)
t-2	0.022	(0.029)	-0.011	(0.010)
t-1	(ref)	(ref)	(ref)	(ref)
t	0.777	(0.116)	0.532	(0.038)
t+1	0.045	(0.032)	0.074	(0.020)
t+2	-0.025	(0.033)	0.033	(0.033)
t+3	-0.106	(0.041)	0.024	(0.044)
t+4	-0.125	(0.061)	0.033	(0.037)
t+5	-0.256	(0.054)	-0.009	(0.047)
t+6	-0.138	(0.086)	0.159	(0.215)
t+7	-0.327	(0.159)	0.089	(0.046)
t+8	-0.648	(0.131)	—	
Observations	7,057,632		89,868	

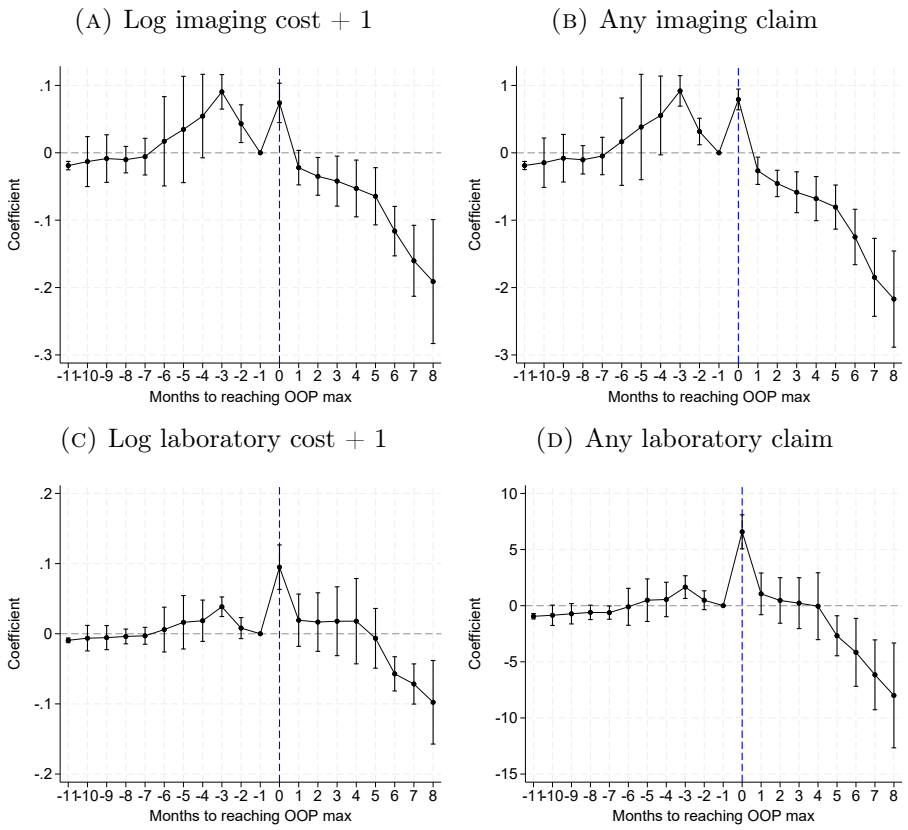
Note: Coefficients and standard errors in parenthesis of the event study specifications for the mean claim cost due to the zero-price effect and due to the health shock effect. The zero-price effect uses the sample of individuals who are not hospitalized during the sample period. The health shock effect uses the sample of treated individuals. Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to -1 for the control group.

Appendix Table 7: Event Study Coefficients by Cohort

	Mean claim cost			Any claim		
	(1) Month 9	(2) Month 7	(3) Month 4	(4) Month 9	(5) Month 7	(6) Month 4
t-8	0.147	—	—	-0.151	—	—
	(0.054)			(0.013)		
t-7	0.124	—	—	-0.108	—	—
	(0.053)			(0.013)		
t-6	0.105	0.038	—	-0.110	-0.099	—
	(0.053)	(0.032)		(0.013)	(0.012)	
t-5	0.078	0.025	—	-0.041	-0.063	—
	(0.052)	(0.032)		(0.013)	(0.012)	
t-4	0.061	0.009	—	-0.077	-0.073	—
	(0.052)	(0.032)		(0.013)	(0.012)	
t-3	0.038	0.007	0.003	-0.084	0.001	-0.048
	(0.052)	(0.031)	(0.056)	(0.013)	(0.012)	(0.013)
t-2	0.016	-0.004	0.008	0.003	-0.011	-0.015
	(0.050)	(0.031)	(0.055)	(0.013)	(0.012)	(0.013)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	0.347	0.412	0.588	0.133	0.161	0.148
	(0.050)	(0.030)	(0.056)	(0.014)	(0.013)	(0.015)
t+1	-0.349	-0.148	-0.019	0.017	0.013	0.005
	(0.051)	(0.031)	(0.056)	(0.013)	(0.013)	(0.014)
t+2	-0.407	-0.180	-0.050	-0.091	-0.011	-0.030
	(0.053)	(0.031)	(0.057)	(0.013)	(0.013)	(0.014)
t+3	-0.446	-0.180	-0.089	-0.152	-0.011	-0.013
	(0.055)	(0.031)	(0.057)	(0.013)	(0.013)	(0.014)
t+4	—	-0.181	-0.080	—	-0.082	-0.104
		(0.032)	(0.059)		(0.013)	(0.014)
t+5	—	-0.162	-0.089	—	-0.128	-0.077
		(0.033)	(0.058)		(0.013)	(0.014)
t+6	—	—	-0.076	—	—	-0.054
			(0.058)			(0.014)
t+7	—	—	0.001	—	—	-0.123
			(0.059)			(0.014)
t+8	—	—	-0.101	—	—	-0.163
			(0.060)			(0.014)
Observations	20,676	23,532	20,664	20,676	23,532	20,664

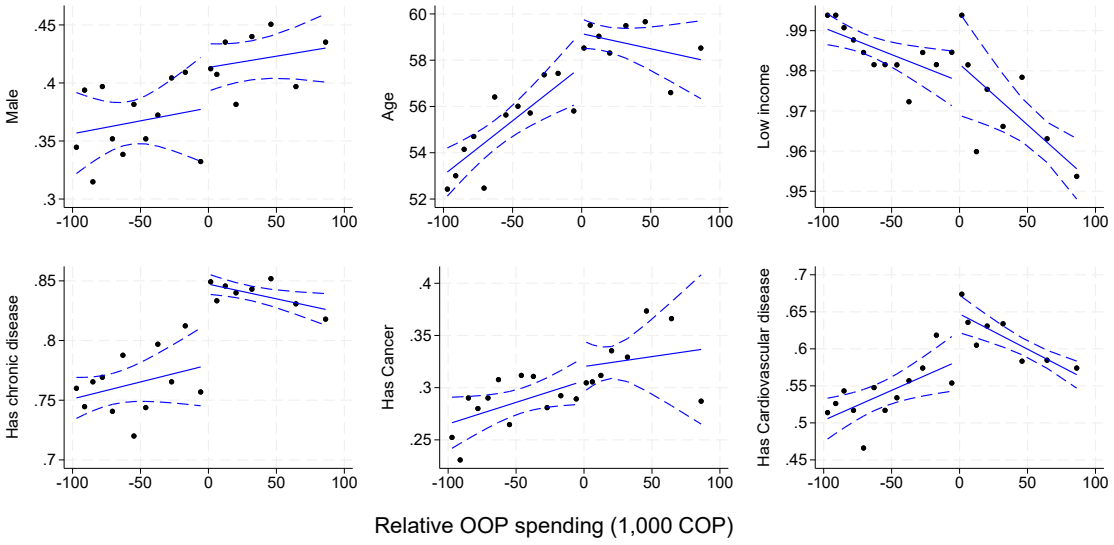
Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (1) for mean claim cost and an indicator for making claims conditional on treated individuals who reach their OOP maximum in September in columns (1) and (4), July in columns (2) and (5), and April in columns (3) and (6).

Appendix Figure 2: Utilization and spending by health service



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (1) for the log of imaging cost in panel A, indicator for making imaging claims in panel B, log of laboratory cost in panel C, and indicator for making laboratory claims in panel D. Estimation uses Callaway and Sant'Anna (2021)'s estimator for staggered treatment. Treated individuals are those who reach their OOP maximum and control individuals are those who do not reach their maximum. Time indicators relative to reaching the OOP maximum are set to −1 for the control group.

Appendix Figure 3: Regression discontinuity on demographics



Note: Regression discontinuity plot on the fraction of males in top left panel, average age in the top middle panel, fraction of individuals making less than 2 times the monthly minimum wage in the top right panel, fraction of individuals with a chronic disease in the bottom left panel, fraction of individuals with cancer in the bottom middle panel, and fraction of individuals with cardiovascular disease in the bottom right panel. Linear regressions are estimated on vigintiles of OOP spending relative to the OOP maximum. Black dots correspond to average outcome in the bin, solid blue lines represent a linear fit, and dashed blue lines represent 95 percent confidence intervals.

Appendix 3 Census Tract Data and Admission Prices

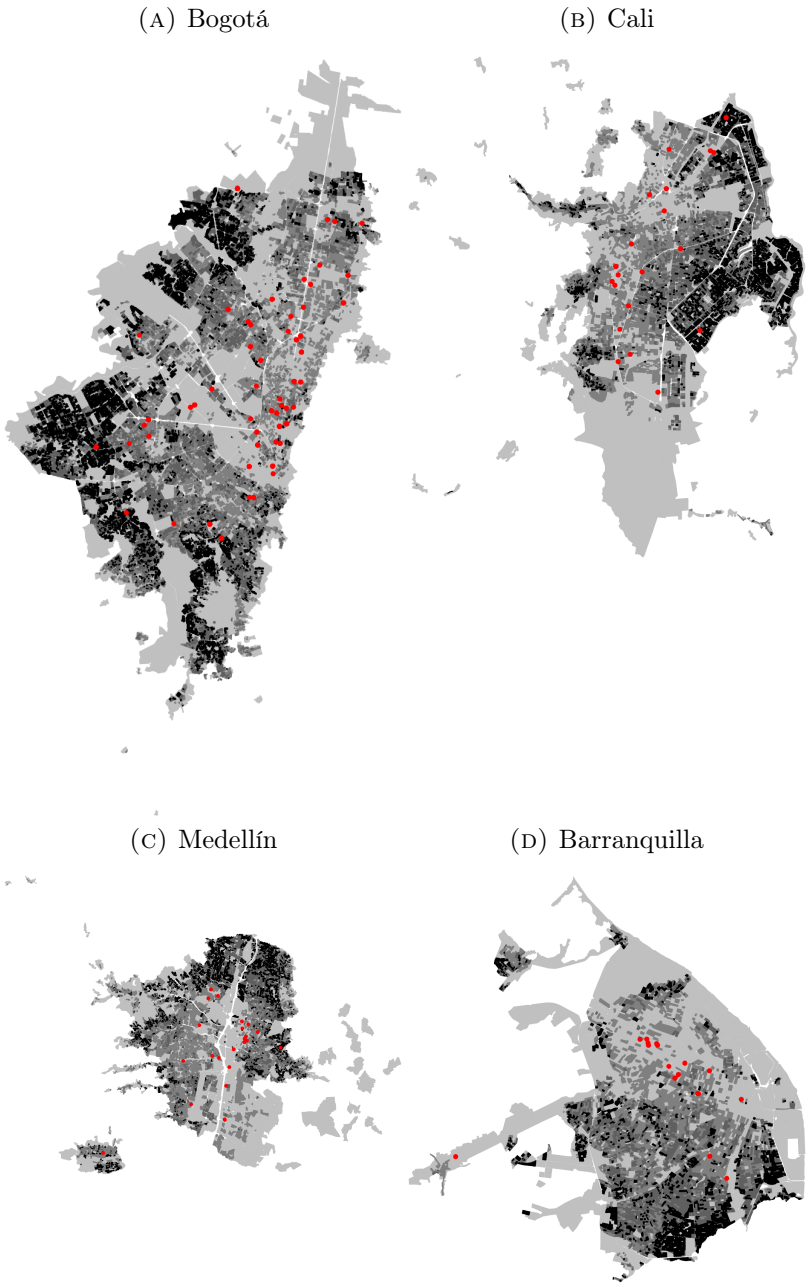
While the claims data reports admission prices that each insurer negotiated with each hospital in its network, these prices sometimes vary with admission characteristics that are unobserved to insurers when they bargain. To average out these characteristics, I estimate the following regression separately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

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

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

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

Appendix Figure 4: Hospital locations and census tracts



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