Health Insurer Gatekeeping

Natalia Serna*

Stanford University

May 1, 2024

Abstract

This paper shows that health insurers can shape the way in which health care is provided to patients beyond the financial characteristics of their contracts by engaging in gatekeeping practices. Using the discontinuity in cost-sharing introduced by the out-of-pocket maximum, I identify effects consistent with insurer gatekeeping even in the presence of information frictions. I find that gatekeeping significantly reduces health care utilization and spending, without effects on individual mortality. Estimates from a structural model of hospital demand indicate that gatekeeping induces patients to travel 1.8 additional kilometers to receive care, while information frictions have negligible effects on consumers' choices.

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

JEL codes: I10, I11, I13, I18.

*e-mail: nserna@stanford.edu. I am deeply grateful to the Colombian Ministry of Health for providing the data for this research. 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 the financial characteristics of their health insurance contracts, such as cost-sharing and coverage (Serna, 2021; Aron-Dine et al., 2013; Eichner, 1998; Newhouse, 1993; Manning et al., 1987). Empirical evidence has also shown that access to health insurance has small impacts on patient health outcomes for certain populations (Finkelstein and McKnight, 2008) and that there are determinants of patient health that are not tied to insurance provision (Finkelstein et al., 2021). Together this evidence seems to point to health insurers only as financial intermediaries between patients and health care providers, analogous to insurance companies in other markets such as car and life insurance. But can health insurers shape the way in which healthcare is provided to consumers unlike these other insurance markets?

This paper explores the role of insurers as health care gatekeepers. Insurers can affect the provision of health care by gatekeeping patients from certain providers – using narrow networks (Ho and Lee, 2017; Buitrago et al., 2024) or prior authorization (Brot-Goldberg et al., 2017)– and certain types of claims –using claim denials (League, 2023).[1] I show that insurer gatekeeping is a much more effective cost-containment mechanism than patient cost-sharing, and that it does not undermine patient health.

I study the effects of gatekeeping on utilization, spending, and where consumers go to seek care in the context of Colombia's contributory health care system. This health system covers individuals who pay payroll 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

[1] Holmes et al. (2024) also show that health insurers can shape productivity and innovation in healthcare.

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 care. Gatekeeping incentives are therefore more salient after patients reach this limit. I show that reaching the OOP maximum in my setting is a random and sudden event, typically a hospitalization, and that individuals cannot preempt it. This provides a unique setting to estimate the causal effect of insurer gatekeeping on different outcomes. To do so, I use enrollment and health claims data from a random panel of 8 million enrollees from 2009 to 2011, who did not switch their insurer over the sample period.

I start by comparing claim prices and likelihood of making claims between individuals who reach their OOP maximum and those who don't, in a dynamic difference-in-differences design. Treated individuals consume significantly cheaper services and have a substantially lower likelihood of making claims than controls after reaching the OOP maximum. 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 as hospitalizations. Instead, results are in line with insurers gatekeeping claims and with information frictions that make consumers unaware of zero prices.

To separate the effect of information frictions from gatekeeping, I estimate my event study specification separately for cohorts of patients that reach their OOP maximum in different months of the year. If information frictions disappear over time, then cohorts who reach their maximum early on should consume more expensive services the closer they are to the end of the calendar year before cost-sharing resets. My findings show no evidence that negative treatment effects vanish for any cohort. Reductions in claim prices and likelihood of making claims are both persis-

tent, suggesting that information frictions are not the main driver of reductions in spending.

Event study results conform to the idea that insurer gatekeeping is an important source of responsiveness to prices. I show that insurers are less likely to gatekeep claims made in an inpatient setting, but are more likely to gatekeep discretionary care made in an outpatient setting, such as imaging and laboratory tests. Because gatekeeping may involve steering patients towards cheaper providers or denying claims altogether, its use raises questions about the impact on patient health. Using a regression discontinuity framework around the patient's OOP spending relative to their OOP maximum, I find no change in individual mortality after reaching the OOP maximum.

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, for example, to derive changes in the marginal disutility of distance that are due to insurer gatekeeping and information frictions.

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 10 percent more than in the observed scenario to reduce commuting distance by 1 kilometer. Put differently, gatekeeping induces individuals to travel on average 1.8 additional kilometers to receive care. Unlike gatekeeping, information frictions have negligible effects on consumers' choices.

Related literature. This paper contributes to the literature on the use of non-price mechanisms to contain health care costs and unnecessary spending, such as spending monitoring programs (Shi, 2024), prior authorization (Roberts et al., 2021; Brot-Goldberg et al., 2023), and claim denials (Gottlieb et al., 2018; League, 2023; Dunn et al., 2024). I add to this literature by quantifying the overall effects of insurer gatekeeping on utilization, spending, mortality, and provider choice.

By demonstrating that gatekeeping is much more effective than cost-sharing at containing spending and that it does not necessarily hurt patient health, I build on prior empirical work finding mixed evidence of patient cost-sharing being effective at these fronts (Chandra et al., 2010; Shigeoka, 2014; Chandra et al., 2014; Baicker et al., 2015; Brot-Goldberg et al., 2017; Chandra et al., 2021; Buitrago et al., 2021). Specifically, I complement prior work describing what happens when health care prices are zero (Chandra et al., 2010; Dague, 2014; Drake et al., 2023). For example, 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 become zero that are incompatible with consumers' behavioral responses.

Finally, my paper is also related to the literature quantifying the relative costs of information frictions in healthcare, of which Handel et al. (2019); Brown (2019) are a few examples of how these frictions impact health insurance pricing and Handel and Kolstad (2015) of how they impact insurance choices. In my setting, I find that information frictions have negligible impacts on which providers consumers choose, hence policies targeted at information provision maybe ineffective at improving the match between patients and providers. Similar findings have been reported in Alpert et al. (2024) in the context of opioid prescriptions.

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: insurance 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.

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
Low: $y < 2$ MMW	1,900	11.5%	28.7%	57.5%
Middle: $y \in [2, 5]$ MMW	7,600	17.3%	115%	230%
High: $y > 5$ 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.

Cost-sharing rules in the contributory system are a function of the enrollee's monthly income level but are standardized across insurers and hospitals. These rules involve a three-tier system of copayments, coinsurance rates, and maximum out-of-pocket (OOP) amounts in the year as seen in table 1. Individuals are assigned specific cost-sharing rules depending on whether they make less than 2, between 2 and 5, or more than 5 times the monthly minimum wage (MMW). 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 doctor or a specialist 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 with the monthly minimum wage and 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 et al., 2021). These studies leverage comparisons across consumers in the different income tiers. Instead, in this paper I study within-patient changes in outcomes before and after they reach their OOP maximum to identify insurer gatekeeping from other sources of price variation (or lack thereof) such as information frictions.

Insurer gatekeeping refers to mechanisms by which the insurer deters, denies, or steers health claims made by its enrollees in order to control costs or risk select. These mechanisms may include requiring prior authorization or having long lines to file claims or get medications. This differs from other definitions where primary care

providers or general practitioners serve as gatekeepers by deciding whether to refer a patient for more specialized care as in the UK.

3 Data and Descriptives

My data consist of all the health claims of a random sample of nearly 8 million enrollees in Colombia's contributory system from 2009 to 2011 who made at least one claim and who did not switch their insurer during the sample period. For every individual I observe 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 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

TABLE 2: Summary Statistics

	Full sample (1)	Above OOP max (2)	Below OOP max (3)
Socio-demographic			
Male	0.48 (0.50)	0.46 (0.50)	0.48 (0.50)
Age	46.9 (17.0)	58.4 (18.1)	46.6 (16.9)
Low income	0.75 (0.43)	0.92 (0.26)	0.75 (0.43)
Medium income	0.19 (0.39)	0.06 (0.25)	0.19 (0.39)
High income	0.06 (0.23)	0.01 (0.10)	0.06 (0.23)
Health			
Cancer	0.17 (0.37)	0.30 (0.46)	0.17 (0.37)
Cardiovascular	0.32 (0.47)	0.64 (0.48)	0.31 (0.46)
Pulmonary	0.05 (0.21)	0.18 (0.39)	0.04 (0.20)
Renal	0.03 (0.17)	0.13 (0.34)	0.03 (0.17)
Health care use			
Mean claim price	12.9 (54.0)	135 (298)	9.71 (16.0)
Total monthly cost	44.5 (175)	645 (866)	28.7 (44.6)
Prescription claims	0.26 (0.93)	1.15 (2.92)	0.23 (0.80)
Outpatient claims	1.13 (1.52)	3.66 (3.70)	1.06 (1.36)
Hospitalization	0.01 (0.03)	0.07 (0.08)	0.00 (0.02)
Observations	600,000	15,393	584,607

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 and price variables are measured in thousands of pesos.

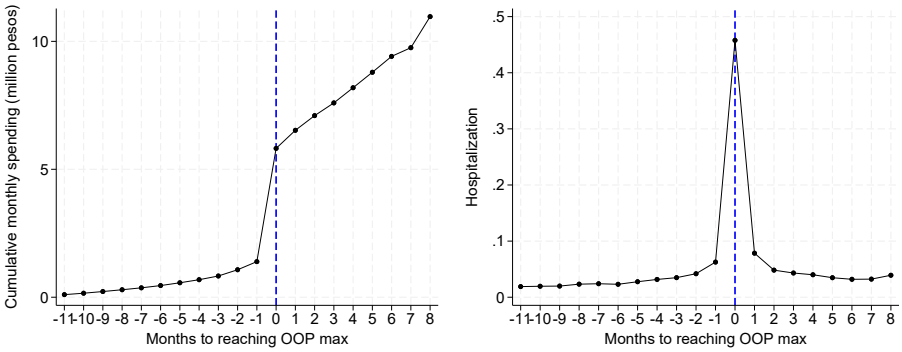
lower income than those who do not reach the limit. 64 percent of those in column (2) have a cardiovascular disease (such as hypertension), while only 31 percent of those in column (3) have this type of diagnosis. Consumers who reach their maximum have higher health care utilization than their counterparts, a difference that is driven only by the health claim made when they reach this maximum. For example, the mean claim price is over 10 times higher and the likelihood of being hospitalized in a month is 7 percentage points higher for those in column (2) relative to column (3).

Reaching the OOP maximum is a sudden event in my setting. Panel A of figure 1 shows that cumulative monthly spending increases smoothly until the month before reaching the OOP maximum and has a sharp discontinuity when this maximum is

reached. This sudden event is typically a hospitalization as seen in panel B 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.

4 Identifying Gatekeeping

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

The empirical strategy of using individuals who face zero prices to identify the magnitude of gatekeeping has several identification threats. The first is a selection bias problem: people who reach their OOP maximum may be unobservably sicker and less responsive to prices compared to those who don't reach it. This type of un-

observed 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 showed that reaching the OOP maximum is a sudden event and therefore there are no reasons to believe that individuals anticipate this event. If this is true, then people who reach their OOP maximum and those who don't should have parallel utilization and spending patterns prior to the event. To characterize these spending patterns more systematically, I compare people who reach the OOP maximum (treated group) and those who don't (control group), before and after they reach the limit in a dynamic difference-in-differences (*did*) design.[2]

The regression specification is as follows:

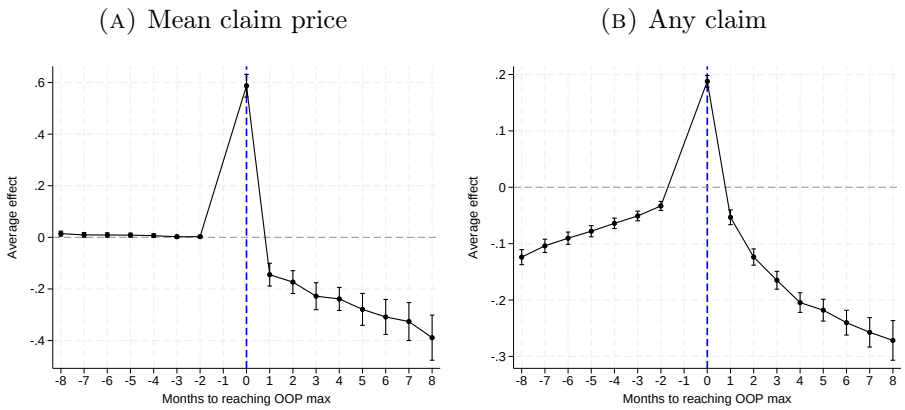
$$y_{it} = \sum_{\substack{k=-11 \\ k\neq -1}}^{8} \beta_k \mathbf{1}\{t - t^* = k\} \times \text{Treated}_i + \theta S_{it} + \alpha_i + \gamma_t + \varepsilon_{it} \tag{1}$$

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, in the style of a regression discontinuity design, imposes a linear trend on the treatment assignment variable; α_i is an individual fixed

[2] 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 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.

effect and γ_t is a calendar month fixed effect.[3] 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 Sun and Abraham (2021)'s estimator to deal with possibly heterogeneous dynamic treatment effects and staggered treatment. Appendix table 2 presents the associated event study coefficients and standard errors and appendix 3 presents robustness checks using a two-way fixed effect estimator and De Chaisemartin and d'Haultfoeuille (2020)'s estimator.

FIGURE 2: Utilization and spending after reaching the OOP limit



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (2) for the mean claim price in panel A, an indicator for making claims in panel B, log of total spending in panel C, and log of total claims in panel D. Regression uses Sun and Abraham (2021)'s estimator. 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.

Controlling for cumulative OOP spending. Recent literature on *did* has devoted attention to the issues that arise when including time-variant and time-invariant covariates (e.g., Caetano et al., 2022). If cumulative OOP spending in my specification is affected by treatment or the true underlying relation between my outcome and this covariate is non-linear, then estimates for β_k will be biased.

[3]This specification is similar to Colonnelli et al. (2020) who use *did* event study regressions that control for the treatment assignment or running variable.

However, note first that cumulative OOP spending determines treatment, not the other way around; and second that outcomes such as the mean claim price contribute linearly to the cumulative OOP spending that is used to determine whether the patient reaches the OOP maximum.

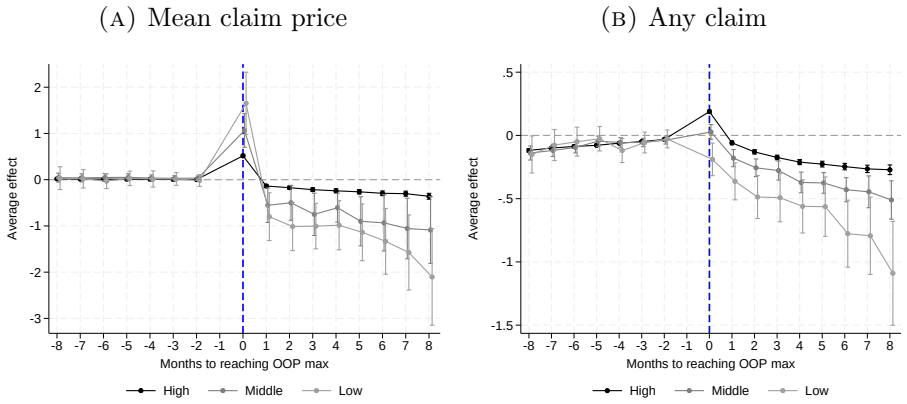
Figure 2 shows event study coefficients and 95 percent confidence intervals using as outcomes the mean claim price in panel A and an indicator for making claims in panel B. I find that trend differences in the mean claim price before the event between treated and controls are negligible, suggestive of parallel pre-trends. More importantly, to the extent that this outcome captures changes in health status, evidence of parallel trends before the event indicates limited selection into reaching the OOP maximum. Although I cannot rule out that treated individuals are on different trends relative to controls for the likelihood of making claims, the fact that trend differences reverse after reaching the OOP maximum is suggestive of substantial dynamic treatment effects. In fact, appendix figure 3 that uses De Chaisemartin and d'Haultfoeuille (2020)'s estimator imposing parallel pre-trends, shows even larger treatment effects on this outcome.

At the time of the event there is a sharp discontinuity in the mean claim price among people who reach the OOP maximum. These individuals claim services that are around 600 thousand pesos ($324) more expensive than individuals who do not reach their limit.[4] The differences in mean claim price become negative over time after reaching the OOP maximum, with treated individuals making claims that are roughly 200 thousand pesos *cheaper* than controls 3 months after the event. In appendix table 1 I find confirming results of this negative treatment effect exploiting the claims level data by estimating a *did* regression of price conditional on service, municipality, and time fixed effects.

[4]The average exchange rate for 2011 was 1,847COP/USD.

While individuals who reach the OOP limit claim relatively cheaper services, their likelihood of making claims does not increase and, in fact, decreases substantially after the event as seen in panel B. One month after the event, individuals who reach the OOP maximum are 5 percentage points (p.p) less likely to make a claim. This difference increases over time, with treated individuals being 25 p.p less likely to make claims between 7 to 8 months after reaching the OOP maximum. These results are robust to excluding individuals who die during the sample period as seen in appendix table 9.

FIGURE 3: Utilization and spending by income group



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (2) for the mean claim price in panel A, and an indicator for making claims in panel B. Estimates in black condition on individuals with incomes above 5 times the monthly minimum wage. Estimates in dark gray condition on individuals with incomes between 2 and 5 times the monthly minimum wage. Estimates in light gray condition on individual with income below 2 times the monthly minimum wage. Regression uses Sun and Abraham (2021)'s estimator. 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.

Because cost-sharing rules are indexed to the enrollee's monthly income, a stacked *did* design, as the one in the previous figure, will compare individuals across income groups for which cumulative OOP spending evolves very differently. This might explain why for outcomes such as the likelihood of making claims, I do not find evidence of parallel pre-trends. To account for the way in which cost-sharing rules are assigned, in figure 3 I replicate my event study specification conditional on each income group. Appendix tables 3 and 4 report the associated coefficients and standard

errors. Results for the mean claim price and the likelihood of making claims all exhibit evidence of parallel trends between treated and control individuals prior to the event. Reductions in both of these outcomes are robust and monotonic with respect to income.

4.1 Heterogeneity Analysis

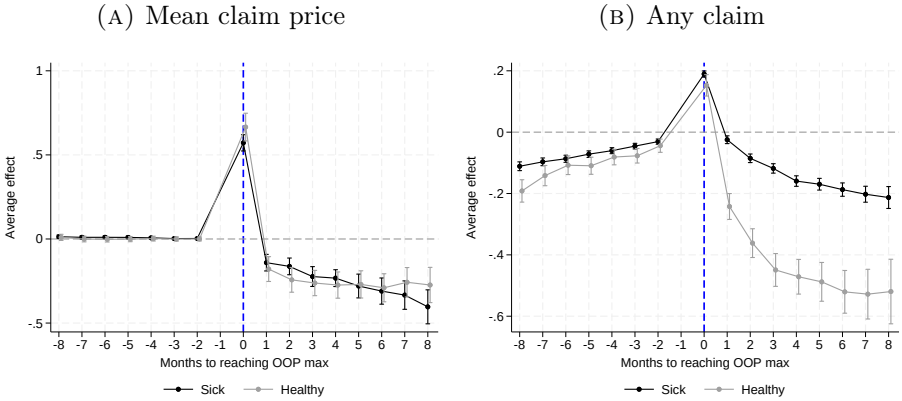
Differences in health care utilization after reaching the OOP maximum are driven by differences health status. Figure 4 reports *did* results on individuals who received a chronic disease diagnosis at any point during the year in black, and conditional on those who never received a diagnosis in gray. Appendix table 5 presents coefficients and standard errors.

Results in panel A show that healthy and sick individuals who reach their OOP maximum have parallel trends in mean claim price relative to controls prior to the event. Between 1 and 6 months after reaching the OOP limit, healthy consumers claim cheaper services compared to individuals who do not reach their limit, and this treatment effect is slightly larger in magnitude than for individuals with chronic diseases. For instance, 2 months after the event, consumers with chronic diseases claim services that are on average 19 thousand pesos cheaper than those claimed by the control group, while healthy consumers claims services that are 25 thousand pesos cheaper.

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 more than three times larger for the former than for the latter every month 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 patients' 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 (2) for the mean claim price 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. Regression uses Sun and Abraham (2021)'s estimator. 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 provide evidence on the type of care that insurers are more likely to deter or steer, I estimate event study designs for subsets of health services. Figure 5 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. Appendix tables 6 and 7 report associated coefficients and standard errors.

Results are suggestive of gatekeeping efforts being present across all types of care but being smaller in magnitude for acute or necessary care. Reductions in the likelihood of making claims are smaller for inpatient services than for outpatient services and prescriptions. However, total spending falls by a greater magnitude for the former because each inpatient claim is more expensive than each outpatient claim.

Panel A shows that treated individuals see a reduction in spending on inpatient services equal to 13 percent relative to controls 3 months after reaching the OOP maximum. The reduction in spending on prescription medications and outpatient services are equal to 1.5 and 9 percent, respectively, in the same month. Instead, panel D shows that individuals who reach their OOP maximum are roughly 12 p.p less likely to make inpatient claims relative to controls 4 months after the event, while they are almost 20 p.p less likely to make outpatient claims in panel F.

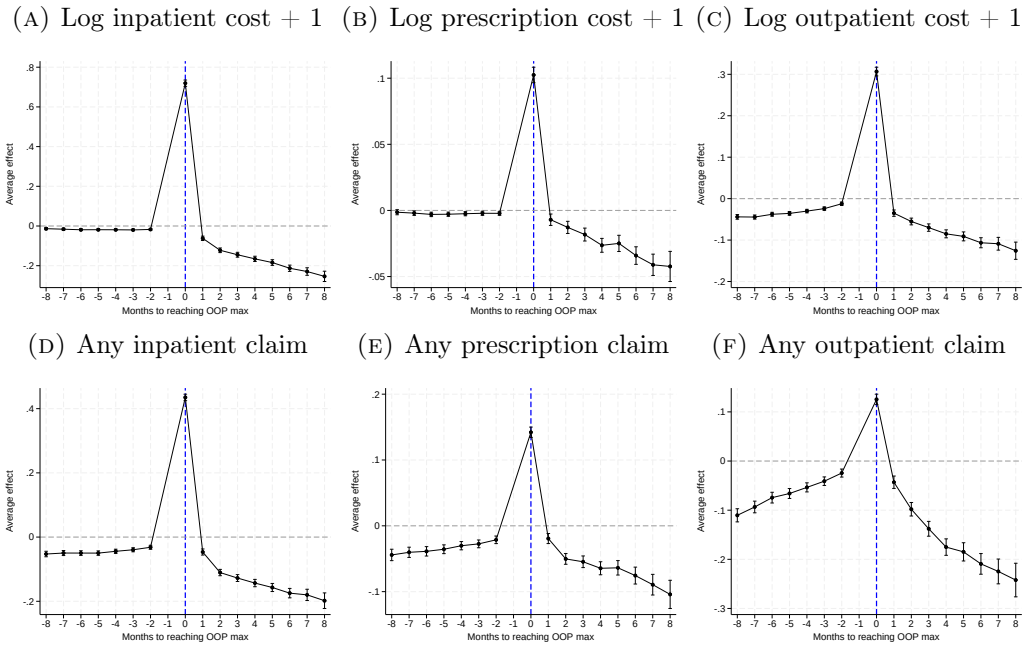
The reduction in the consumption of care that is less acute suggests that, despite their worsened health and their zero cost-sharing, treated individuals behave as if they face non-zero prices for outpatient care. Appendix figure 2 reports event study results for other types of potentially discretionary services such as imaging and laboratory tests, which are consistent with this hypothesis.

4.3 Zero-Price and Health Shock Effects

The event studies in figure 2 generally suggest that after reaching the OOP maximum, treated individuals differ systematically from controls. However, 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 zero prices. These effects are highly correlated in the sense that individuals face sudden hospitalizations and zero prices 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

FIGURE 5: Utilization and spending by type of care



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (2) for the log of inpatient cost in panel A, log of prescription cost in panel B, log of outpatient cost in panel C, indicator for making inpatient claims in panel D, indicator for making prescription claims in panel E, and indicator for making outpatient claims in panel F. Regressions use Sun and Abraham (2021)'s estimator. 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.

compared to when the patient is relatively healthy.

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 health care that individual i faces in period t, and $Y(i,t)$ be the mean claim cost of individual i in period t. The *did* specification using the mean claim price as outcome in figure 2 identifies the average treatment effect on the treated (ATT) as

$$\beta = (E[Y(i,1)|D(i,1) = 0] - E[Y(i,1)|D(i,1) > 0])$$
$$- (E[Y(i,0)|D(i,1) = 0] - E[Y(i,0)|D(i,1) > 0])$$

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

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

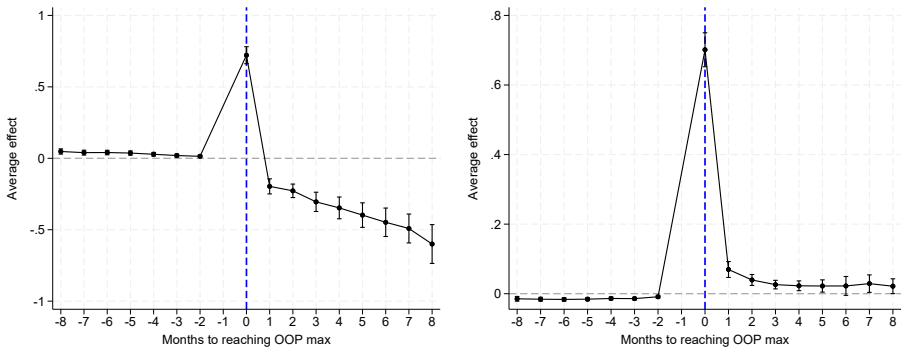
where $x = P(H(i,1) = 1|D(i,1) = 0)$ and $y = P(H(i,1) = 1|D(i,1) > 0)$. The expression above shows that the ATT is the sum of the zero-price effect and the marginal health shock effect on the treated relative to controls.

Results of this decomposition exercise for the mean claim cost are presented figure 6. Appendix table 10 reports associated coefficients and standard errors. Panel A depicts event study coefficients for the zero-price effect, that is, conditional on people who never have a hospitalization in the study period. The regression specification in this case is the same as equation (2).

Panel B depicts event study coefficients for the treated health shock effect, that is, conditional on individuals who reach the OOP maximum. For this effect, time

FIGURE 6: Effect Decomposition

(A) Zero-price effect (B) Health shock effect



Note: Coefficients and 95 percent confidence intervals of the event study specifications for the mean claim price due to the zero-price effect in panel A and due to the health shock effect in panel B. The zero-price effect uses the sample of individuals who are never hospitalized during the sample period. The health shock effect uses the sample of treated individuals comparing those who are hospitalized are those who are not, before and after the hospitalization. Regression uses Sun and Abraham (2021)'s estimator. Time indicators are constructed relative to reaching the OOP maximum for the zero-price effect and are relative to the month when the individual is hospitalized for the health shock effect.

indicators are constructed relative to the month when the individual is hospitalized, h^*, and are equal to those when the individual reaches the OOP maximum, because the two events are perfectly correlated in time. Formally, the regression equation for the health shock effect among those who reach the OOP maximum is:

$$y_{it} = \sum_{\substack{k=-11 \\ k\neq -1}}^{8} \beta_k \mathbf{1}\{t - h^* = k\} \times \text{Hospitalized}_i + \alpha_i + \gamma_t + \varepsilon_{it} \tag{2}$$

Findings show that when the event occurs, the zero-price effect is 20 thousand pesos higher than the treated health shock effect. The zero-price effect on the mean claim price quickly becomes negative one month after reaching the OOP maximum and remains negative thereafter. However, the health shock effect is positive up to 8 months after the individual is hospitalized. The direction of 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 the fact that it

becomes negative, is irreconcilable with health care demand being perfectly inelastic after consumers reach their OOP maximum. This suggests that factors other than consumers' OOP prices may explain why health care demand responds to cost-sharing. I delve into these factors in the next subsection.

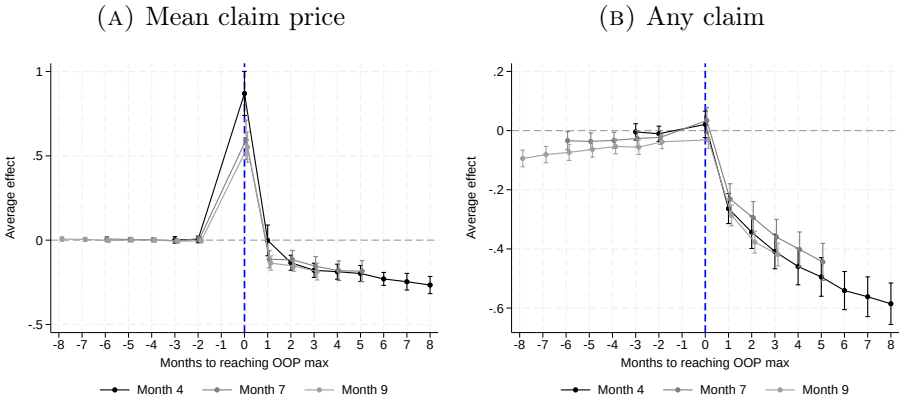
4.4 Dynamic incentives

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 maximum, 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 6 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 maximum, 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.

FIGURE 7: Utilization and spending by cohort



Note: Coefficients and 95 percent confidence intervals of the event study specifications following equation (2) for mean claim price in panel A and an indicator of making claims in panel B conditional on treated individuals who reach their OOP maximum in April (black), July (dark gray), and September (light gray). Regression uses Sun and Abraham (2021)'s estimator. 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.

To get at the role of information frictions in explaining the spending and utilization patterns in figure 2, I estimate separate event study specifications conditional on treated individuals who reach the OOP maximum in different months of the year. Appendix table 11 presents the set of event study coefficients.

Findings in figure 7 show that treated individuals consume significantly cheaper services and are significantly less likely to make claims after reaching the OOP maximum regardless of when the event happens. Comparing across months, we see that individuals who reach the OOP maximum in April consume relatively cheaper services by the end of the calendar year relative to controls, and this reduction in the mean claim price is similar in size for individuals who reach their maximum in July and September. While the homogeneity of effects by treatment timing does not rule out the presence of information frictions, the fact that treatment effects are negative even for consumers who are treated in April suggests a role for insurer gatekeeping

in explaining my results.

My findings are in contrast to Brot-Goldberg et al. (2017), who show no evidence of consumers price-shopping or of consumers responding to the true shadow price of care after they hit their deductible. Figures 2 and 7 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 potentially due to insurer gatekeeping.

The magnitude of gatekeeping. The large declines in utilization and spending after reaching the OOP maximum cannot alone be explained by changes in cost-sharing. Around the OOP maximum and for a low-income consumer, insurers' costs increase 11.5 percent since it moves from covering 88.5 percent to 100 percent of health care costs. Gatekeeping incentives are not likely to change discontinuously on this (intensive) margin, but they are likely driven by changes in the consumer's risk type. If an individual reaches the OOP maximum, they are at risk of being very expensive to the insurer. It is this change in risk (extensive margin) to which insurers respond by restricting the number and the type of services that their enrollees claim.

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, which has been the focus of other recent papers (e.g., Brot-Goldberg et al., 2023; Shi, 2024; League, 2023). Given that insurers in Colombia cannot design their cost-sharing rules nor premiums, they can engage in steering through non-price mechanisms such as claim denials, providing access to narrow hospital networks, or requiring prior authorization for certain services or providers.

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 et al. (2021) show for example that $100 reductions in patient drug budgets among the elderly increases mortality by 13.5 percent. Buitrago et al. (2021) find that mortality increases 0.8 per 1,000 enrollees 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 a *did* specification because, by definition, the outcome does not vary before the event for treated individuals. Let y_{it} be an indicator for whether individual i dies in month t; S_{it} be the individual's relative OOP spending (cumulative OOP spending minus OOP maximum) in month t; and $T_{it} = \mathbf{1}\{S_i - oop_i \geq 0\}$ be an indicator for whether the individual reaches her OOP maximum denoted by oop_i in month t. I estimate the following regression in binned data based on 30 bins of S_{it} and for $S_{it} \in [-100, 100]$ thousand pesos:

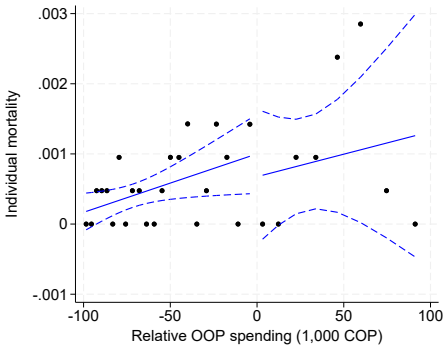
$$y_{it} = \alpha T_{it} + \gamma S_{it} + \beta T_{it} \times S_{it} + \varepsilon_{it} \tag{3}$$

Appendix 4 presents tests of the RD assumptions. For instance, appendix figure 4 shows that the distribution of relative OOP spending is smooth around zero, which suggests potentially no manipulation of the running variable; and appendix figure 5 presents RD plots for covariate smoothness around the cutoff.[5]

Results of equation (3) are presented in figure 8. The main takeaway is that

[5] Although I find a discontinuity in the fraction of low-income individuals, this should lead to lower mortality rates after reaching the OOP maximum.

FIGURE 8: Individual mortality



Note: Regression discontinuity plot for individual mortality in the full sample. Linear regressions are estimated on 30 bins 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.

individual mortality does not change around the cutoff. While the null effect seems to counter results in Buitrago et al. (2021), in my case it speaks to discontinuities in relative OOP spending rather than in copayments as in the authors' case. Both sets of results do align with the notion that consumers are myopic and therefore that they care about the spot price of health care when making health care consumption decisions.

5 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 cannot speak to what would health care consumption or access to care be in absence of gatekeeping given that it cannot completely disentangle the relative magnitude of gatekeeping and information frictions. In this section I develop a structural model of demand for hospitals to answer this question.

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

j. Consumer types are defined by a combination of sex and five-year age group. The consumer chooses a hospital h in the network of her insurer based on the indirect utility in two states of the world, before and after reaching the OOP maximum:

$$u_{ijhm} = \begin{cases} (\alpha_i + \sigma_\alpha \omega_i) r_i p_{jh} + \tau \sum_{l \in m} q_{\theta l} d_{lh} + \xi_h + \varepsilon_{ijhm} & \text{if } c_i + \nu_i \leq oop_i \\ (\beta_i + \sigma_\beta \omega_i) p_{jh} + \tau \sum_{l \in m} q_{\theta l} d_{lh} + \xi_h + e_{ijhm} & \text{o.w} \end{cases} \tag{4}$$

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. The probability that a consumer type θ lives in locality l within municipality m is given by $q_{\theta l}$ and equals the population density of this type of consumer at each locality. d_{lh} is the distance from locality l's centroid to hospital h. Price coefficients are given by $\alpha_i = x_i'\alpha$, $\beta_i = x_i'\beta$, where x_i is a vector of consumer demographics (dummies for sex and age group), diagnoses (indicator for having a chronic disease), and an intercept. Moreover, $\omega_i \sim N(0,1)$ captures unobserved heterogeneity in price sensitivity across consumers with dispersion parameters given by σ_α and σ_β in each state of the world. Finally, ξ_h is a hospital fixed effect representing shared preferences for hospital h across consumers.

States of the world differ on the source of responsiveness to prices. Before reaching the OOP maximum, the consumer responds to prices up to the coinsurance rate. After reaching the OOP maximum, 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 maximum as

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

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

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

I further assume that ν_i, ω_i, ε_{ijhm}, and e_{ijhm} are independent of each other, and that ε_{ijhm} and e_{ijhm} follow a type-I extreme value distribution. I set $\sigma_\alpha = \sigma_\beta \equiv \sigma_p$ to account for the fact that I can only separately identify one dispersion parameter on unobserved preference heterogeneity from the parameter on information frictions. This is a sensible restriction given that we can expect insurers to impact consumer choices only based on observable patient characteristics but not on their unobservables.

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 (4) therefore captures the expected distance to each hospital from each census tract conditional on the patient's age and municipality of residence.

Appendix 5 describes this census data by reporting maps of the 4 largest municipalities in my sample with their localities and hospital geolocations. In this appendix I also explain my methodology for obtaining negotiated admission prices since 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:

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

where

$$P_{ijhm}^s = \int \frac{\exp(\delta_{ijhm}^s)}{\sum_k \exp(\delta_{ijkm}^s)} \phi(\omega) \quad \text{for } s = \{1, 2\} \tag{5}$$

and

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

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

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 in each state of the world. 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.

The unobserved preference heterogeneity parameter σ_p is identified from observationally identical patients that have not reached their OOP maximum but choose

different hospitals. Finally, the impact of information frictions captured by σ_ν is identified from comparing the choices made by patients who reach their OOP maximum and are observationally identical except for their OOP costs prior to the admission.

Estimates. I estimate the hospital demand model using simulated maximum likelihood to approximate the integrals in equation (5). 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, a 10,000 pesos increase in OOP prices reduces the probability of choosing a hospital by 8.54 percent. After reaching the maximum, a 10,000 pesos increase in admission prices reduces the choice probability by 0.57 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 it does not capture the extensive margin effect of gatekeeping.

Interactions of prices with consumer demographics and diagnoses are in line with intuition and previous literature (Ho, 2006). Patients with chronic diseases are significantly less sensitive to OOP prices than patients without diagnoses. Gatekeeping incentives are stronger among older individuals who are potentially more expensive to the insurer. However, insurers are less likely to gatekeep claims from individuals with chronic diseases compared to healthy individuals, consistent with findings in panel B of figure 4. Price sensitivity in both states of the world is substantially heterogeneous across consumers as seen in the estimate for σ_p. I find that patients dislike commuting: if they have to travel one additional kilometer to visit a hospital, the probability of choosing this hospital decreases by 26.5 percent. Moreover, there is no evidence of

TABLE 3: Hospital demand estimates

		coef	se
OOP price		-8.535	(0.287)
Price		-0.574	(0.019)
Distance		-0.267	(0.003)
σ_p		0.234	(0.171)
σ_ν		0.003	(0.0009)
Interactions			
OOP price	Male	2.033	(0.140)
	Age 10-19	-2.407	(0.578)
	Age 20-29	2.035	(0.255)
	Age 30-39	0.848	(0.072)
	Age 40-49	-0.361	(0.118)
	Age 50-59	-1.356	(0.075)
	Age 60-69	1.045	(0.102)
	Age 70 or older	(ref)	
	Sick	6.960	(0.489)
Price	Male	0.421	(0.031)
	Age 10-19	0.360	(0.166)
	Age 20-29	0.349	(0.073)
	Age 30-39	0.026	(0.008)
	Age 40-49	0.422	(0.076)
	Age 50-59	0.360	(0.070)
	Age 60-69	0.202	(0.062)
	Age 70 or older	(ref)	
	Sick	0.725	(0.024)
Observations		596,130	

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

information frictions as seen by the estimate of σ_ν, which goes in line with findings in the model-free section.

6 Partial Equilibrium Analysis

Using my model estimates I conduct two partial equilibrium analyses that reveal the relative importance of gatekeeping and information frictions on access to care.

First, to quantify the extensive margin effect of gatekeeping on commuting I set $\beta_i = \sigma_p = 0$. Second, to quantify the impact of information frictions, I set $\sigma_\nu = 0$. In each scenario, I recompute individuals' choice probabilities and present summary statistics of monetized marginal effects of distance. I specify the monetized marginal effect of distance as:

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

where N_{jh} is the number of patients that have hospital h in their choice set with insurer j. This effect can be interpreted as the average patient's willingness-to-pay to reduce commuting distance to hospital h by 1 kilometer.

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

TABLE 4: Monetized marginal effect of distance

	p25	median	mean	p75
Observed	-4,231	-2,681	-2,427	-1,517
No gatekeeping	-4,208	-2,692	-2,915	-1,495
No information frictions	-4,231	-2,681	-2,427	-1,517

Note: Table shows the mean, median, 25th, and 75th percentiles of the distribution of monetized marginal effect of distance in the observed scenario, the exercise without gatekeeping setting $\beta_i = \sigma_\beta = 0$, and the exercise without information frictions setting $\sigma_\nu = 0$. In each scenario I drop observations with values above the 99th percentile and below the 1st percentile. Values are measured in 2011 COP.

Table 4 presents the mean, median, and 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 would choose hospitals only based on distance and

quality when they reach their OOP maximum. Consistent with this intuition, I find that without gatekeeping the average consumer would be willing to pay 20 percent more to reduce commuting distance by 1 kilometer relative to the observed scenario.

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 the average individual to pay 342 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 forces the average consumer to travel 1.8 additional kilometers to visit a hospital. Unlike gatekeeping, information frictions have no effects effects on commuting distance. This finding goes in line with the model-free evidence that pointed to gatekeeping rather than information as the main source of choice frictions for consumers.

TABLE 5: Heterogeneity in monetized marginal effect of distance

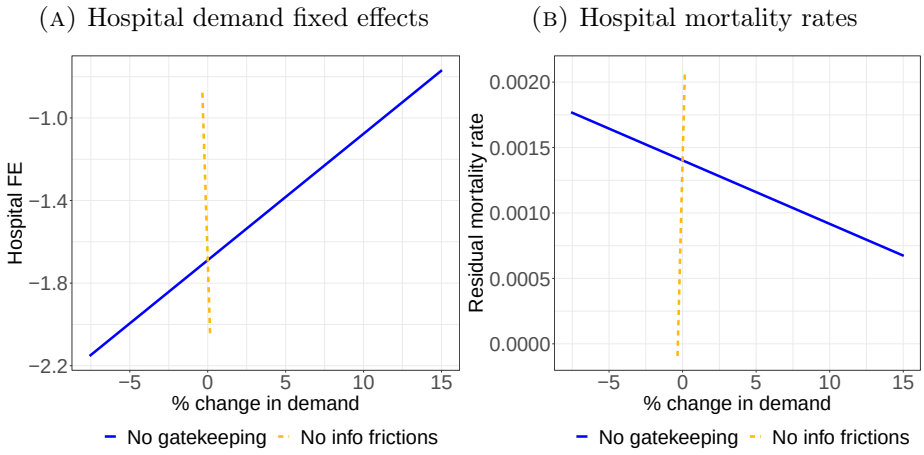
	Observed	No gatekeeping	No information frictions
Panel A			
Aged 65 or older	-2,357	-2,244	-2,920
Aged less that 65	-2,920	-2,956	-2,919
Panel B			
Chronic disease	-923	-1,023	-922
Healthy	-2,342	-2,808	-2,341

Note: Table shows the mean of the distribution of monetized marginal effect of distance across individuals aged less than 65, aged 65 or older, with a chronic disease, and without diseases. Column (1) presents results for the observed scenario, column (2) for the exercise without gatekeeping setting $\beta_i = \sigma_\beta = 0$, and column (3) for the exercise without information frictions setting $\sigma_\nu = 0$. In each scenario I drop observations with values above the 99th percentile and below the 1st percentile. Values are measured in 2011 COP.

In table 5 I explore the heterogeneity of results across age and health status. Panel A shows that for patients aged less than 65 the impact of gatekeeping is larger than for patients aged 65 or older. Without gatekeeping, younger individuals would be willing to pay 1 percent more than in the observed scenario to reduce commuting distance to hospitals by 1 kilometer, however patients aged 65 or older would be willing to

pay 5 percent less. Panel B shows that gatekeeping induces individuals with chronic conditions and healthy consumers to travel 1 and 1.8 additional kilometers relative to the observed scenario, respectively. This result potentially reflects insurers' reduced incentives to gatekeep sick patients relative to healthy ones.

FIGURE 9: Changes in demand



Note: Panel A presents a linear relation between percentage change in demand for every hospital and hospital demand fixed effect. Panel B presents a linear relation between percentage change in demand for every hospital and hospital mortality rate. Both panels use the subsample of patients with chronic diseases. Blue lines correspond to changes in the scenario without gatekeeping relative to the observed scenario, and yellow lines correspond to changes in the scenario without information frictions relative to the observed scenario.

The partial equilibrium analysis reveals that patients would have chosen hospitals that are closer to them on average if it weren't for gatekeeping. What would have health outcomes looked like if consumers had visited these closer hospitals? To provide descriptive evidence towards answering this question, I calculate the correlation between predicted change in hospital demand and measures of hospital quality such as estimated hospital demand fixed effects and hospital mortality rates for the subsample of patients with chronic diseases.[6] Figure 9 presents these correlations, blue lines depict the scenario without gatekeeping and yellow lines the scenario without infor-

[6]To calculate the hospital mortality rate, I estimate the following regression: $y_{ih} = x_i'\beta + \xi_h + \varepsilon_{ih}$, where y_{ih} is an indicator for whether individual i died at hospital h during 2011; x_i is a vector of patient characteristics including sex, dummies for 10-year age categories, an indicator for whether the individual has a chronic disease, and dummies for income group; and ξ_h is a hospital fixed effect. The residual mortality rate is $\hat{\xi}_h$.

mation frictions. The figure shows a higher quality gradient in the scenario without gatekeeping than without information frictions. Without gatekeeping demand from patients with chronic conditions is reallocated towards higher-quality hospitals based on these measures.

7 Conclusions

In this paper I show that health insurers shape the way in which health care is provided to patients by engaging in gatekeeping practices. Gatekeeping has stronger effects on health care utilization and spending compared to demand-side 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, 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 setting reach their OOP maximum as-if-randomly. Those who reach their maximum consume significantly cheaper services and are substantially less likely to make claims afterwards. These results are at odds with behavioral assumptions about consumers when they face zero prices and are in favor of gatekeeping and information frictions driving consumer choices. Estimates from a structural model of hospital demand show that gatekeeping induces patients on average to travel 1.8 additional kilometers to receive care. Finally, I find no impact of gatekeeping on patient health outcomes such as mortality.

In the discussion of how best to deliver health insurance coverage, these findings suggest a role for private insurers as buffers of patient moral hazard. However, recent media attention to cases where gatekeeping has prevented patients from receiving

adequate care for their chronic health conditions in the US, indicates that government regulation of gatekeeping is needed.[7]

References

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

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.

Brown, Z. Y. (2019): "Equilibrium Effects of Health Care Price Information," *Review of Economics and Statistics*, 101, 699–712.

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

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.

Buitrago, G., P. Rodriguez-Lesmes, N. Serna, and M. Vera-Hernandez (2024): "The Role of Hospital Networks in Individual Mortality," *Working paper*.

Caetano, C., B. Callaway, S. Payne, and H. S. Rodrigues (2022): "Difference in Differences with Time-varying Covariates," *arXiv preprint arXiv:2202.02903*.

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

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

——— (2014): "The Impact of Patient Cost-Sharing on Low-Income Populations: Evidence from Massachusetts," *Journal of health economics*, 33, 57–66.

Colonnelli, E., M. Prem, and E. Teso (2020): "Patronage and Selection in Public Sector Organizations," *American Economic Review*, 110, 3071–99.

Dague, L. (2014): "The Effect of Medicaid Premiums on Enrollment: A Regression Discontinuity Approach," *Journal of Health Economics*, 37, 1–12.

De Chaisemartin, C. and X. d'Haultfoeuille (2020): "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110, 2964–2996.

Drake, C., D. Anderson, S.-T. Cai, and D. W. Sacks (2023): "Financial Transaction Costs Reduce Benefit Take-up: Evidence from Zero-Premium Health Insurance Plans in Colorado," *Journal of Health Economics*, 89, 102752.

Dunn, A., J. D. Gottlieb, A. H. Shapiro, D. J. Sonnenstuhl, and P. Tebaldi (2024): "A Denial a Day Keeps the Doctor Away," *The Quarterly Journal of Economics*, 139, 187–233.

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

Finkelstein, A., M. Gentzkow, and H. Williams (2021): "Place-Based Drivers of Mortality: Evidence from Migration," *American Economic Review*, 111, 2697–2735.

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.

Gottlieb, J. D., A. H. Shapiro, and A. Dunn (2018): "The Complexity of Billing and Paying for Physician Care," *Health Affairs*, 37, 619–626.

Handel, B. R. and J. T. Kolstad (2015): "Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare," *American Economic Review*, 105, 2449–2500.

Handel, B. R., J. T. Kolstad, and J. Spinnewijn (2019): "Information Frictions and Adverse Selection: Policy Interventions in Health Insurance Markets," *Review of Economics and Statistics*, 101, 326–340.

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

Ho, K. and R. S. Lee (2017): "Insurer Competition in Health Care Markets," *Econometrica*, 85, 379–417.

Holmes, J., J. Kolstad, and K. Lavetti (2024): "Middlemen or Innovators? The Role of Insurers in Cost and Productivity in Health Care," *Working paper*.

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. (2023): "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.

Roberts, J., R. McDevitt, P. Eliason, J. Leder-Luis, and R. League (2021): "Enforcement and Deterrence of Medicare Fraud: The Case of Non-emergent Ambulance Rides," *NBER Working Paper*.

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

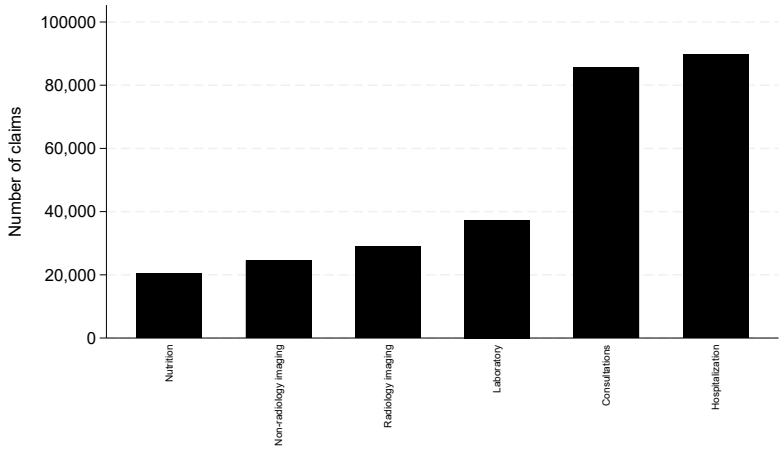
Shi, M. (2024): "Monitoring for Waste: Evidence from Medicare Audits," *The Quarterly Journal of Economics*, 139, 993–1049.

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

Sun, L. and S. Abraham (2021): "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 225, 175–199.

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 Table 1: Claims-level difference-in-differences regression

	Log price	
	(1)	(2)
Treated×Post	-0.075	-0.170
	(0.002)	(0.069)
Treated	0.202	0.407
	(0.001)	(0.100)
Fixed effects		
Service	Yes	No
Municipality	Yes	No
Year	Yes	Yes
Month	Yes	Yes
Observations	8,658,716	

Note: Table shows a regression of log price on an interaction between treatment (defined as reaching the OOP maximum) and post-period indicator (for months after reaching the OOP maximum) controlling for service, municipality, year, and calendar month fixed effects. Estimation uses claims-level data for my main analysis sample. Standard errors in parenthesis are clustered at the service level. Columns (1) and (2) differ in the set of fixed effects.

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 2: Main Event Study Coefficients

	Mean claim cost		Any claim	
	coef	se	coef	se
t-11	0.021	(0.007)	-0.142	(0.014)
t-10	0.020	(0.006)	-0.131	(0.010)
t-9	0.019	(0.005)	-0.128	(0.008)
t-8	0.014	(0.005)	-0.124	(0.007)
t-7	0.010	(0.005)	-0.104	(0.006)
t-6	0.009	(0.004)	-0.090	(0.006)
t-5	0.009	(0.004)	-0.078	(0.005)
t-4	0.007	(0.003)	-0.064	(0.005)
t-3	0.002	(0.003)	-0.051	(0.004)
t-2	0.003	(0.002)	-0.033	(0.004)
t-1	(ref)	(ref)	(ref)	(ref)
t	0.587	(0.022)	0.188	(0.005)
t+1	-0.145	(0.023)	-0.053	(0.007)
t+2	-0.173	(0.023)	-0.124	(0.007)
t+3	-0.228	(0.027)	-0.165	(0.008)
t+4	-0.239	(0.023)	-0.205	(0.009)
t+5	-0.279	(0.032)	-0.218	(0.010)
t+6	-0.309	(0.035)	-0.240	(0.011)
t+7	-0.326	(0.037)	-0.257	(0.013)
t+8	-0.389	(0.045)	-0.272	(0.018)
Observations	7,200,000		7,200,000	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (2) for the mean claim cost and an indicator for making claims, using Sun and Abraham (2021)'s estimator. 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 for Mean Claim Cost by Income Group

	(1) High		(2) Middle		(3) Low	
	coef	se	coef	se	coef	se
t-11	-0.124	(0.206)	0.083	(0.069)	0.027	(0.005)
t-10	0.058	(0.131)	0.020	(0.057)	0.027	(0.005)
t-9	0.049	(0.116)	0.046	(0.054)	0.024	(0.004)
t-8	0.033	(0.126)	0.045	(0.049)	0.019	(0.004)
t-7	0.019	(0.102)	0.039	(0.045)	0.014	(0.004)
t-6	-0.033	(0.081)	0.048	(0.043)	0.013	(0.003)
t-5	0.030	(0.080)	0.051	(0.040)	0.010	(0.003)
t-4	0.016	(0.089)	0.036	(0.034)	0.008	(0.003)
t-3	0.015	(0.070)	0.030	(0.029)	0.003	(0.002)
t-2	-0.017	(0.065)	0.035	(0.022)	0.002	(0.002)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	1.652	(0.344)	1.065	(0.186)	0.517	(0.017)
t+1	-0.800	(0.264)	-0.552	(0.190)	-0.137	(0.018)
t+2	-1.013	(0.266)	-0.500	(0.193)	-0.173	(0.018)
t+3	-1.004	(0.251)	-0.749	(0.233)	-0.216	(0.020)
t+4	-0.985	(0.271)	-0.606	(0.160)	-0.240	(0.022)
t+5	-1.138	(0.313)	-0.901	(0.271)	-0.262	(0.024)
t+6	-1.333	(0.363)	-0.933	(0.310)	-0.294	(0.026)
t+7	-1.575	(0.414)	-1.055	(0.335)	-0.302	(0.028)
t+8	-2.104	(0.532)	-1.087	(0.369)	-0.359	(0.033)
Observations	411,936		1,364,040		5,424,024	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (2) for mean claim cost conditional on individuals with incomes above 5 time the monthly minimum wage in column (1), with incomes between 2 and 5 times the monthly minimum wage in column (2), and with incomes below 2 times the monthly minimum wage in column (3). Estimation uses Sun and Abraham (2021)'s estimator. 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 for Making Any Claim by Income Group

	(1) High		(2) Middle		(3) Low	
	coef	se	coef	se	coef	se
t-11	-0.218	(0.153)	-0.165	(0.055)	-0.134	(0.014)
t-10	-0.050	(0.107)	-0.098	(0.037)	-0.129	(0.010)
t-9	-0.057	(0.076)	-0.143	(0.033)	-0.124	(0.008)
t-8	-0.150	(0.075)	-0.139	(0.028)	-0.119	(0.007)
t-7	-0.076	(0.063)	-0.118	(0.025)	-0.100	(0.006)
t-6	-0.049	(0.058)	-0.095	(0.022)	-0.087	(0.006)
t-5	-0.026	(0.049)	-0.045	(0.021)	-0.078	(0.005)
t-4	-0.120	(0.049)	-0.056	(0.019)	-0.062	(0.005)
t-3	-0.056	(0.042)	-0.056	(0.017)	-0.049	(0.005)
t-2	-0.028	(0.036)	-0.037	(0.016)	-0.032	(0.004)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	-0.189	(0.066)	0.027	(0.030)	0.188	(0.006)
t+1	-0.363	(0.075)	-0.179	(0.035)	-0.059	(0.008)
t+2	-0.487	(0.087)	-0.256	(0.035)	-0.130	(0.009)
t+3	-0.493	(0.097)	-0.278	(0.039)	-0.174	(0.009)
t+4	-0.561	(0.107)	-0.372	(0.042)	-0.211	(0.010)
t+5	-0.563	(0.120)	-0.376	(0.042)	-0.227	(0.012)
t+6	-0.777	(0.135)	-0.429	(0.049)	-0.246	(0.013)
t+7	-0.794	(0.156)	-0.446	(0.064)	-0.264	(0.015)
t+8	-1.090	(0.210)	-0.510	(0.077)	-0.272	(0.019)
Observations	411,936		1,364,040		5,424,024	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (2) for making any claim conditional on individuals with incomes above 5 time the monthly minimum wage in column (1), with incomes between 2 and 5 times the monthly minimum wage in column (2), and with incomes below 2 times the monthly minimum wage in column (3). Estimation uses Sun and Abraham (2021)'s estimator. 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 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	0.031	(0.018)	0.016	(0.007)	-0.190	(0.039)	-0.133	(0.015)
t-10	0.004	(0.011)	0.020	(0.006)	-0.225	(0.026)	-0.114	(0.010)
t-9	0.015	(0.011)	0.018	(0.006)	-0.218	(0.022)	-0.113	(0.008)
t-8	0.010	(0.009)	0.013	(0.005)	-0.192	(0.019)	-0.111	(0.007)
t-7	0.000	(0.009)	0.010	(0.005)	-0.142	(0.017)	-0.097	(0.006)
t-6	-0.002	(0.008)	0.010	(0.004)	-0.108	(0.016)	-0.087	(0.006)
t-5	-0.001	(0.008)	0.009	(0.004)	-0.109	(0.014)	-0.072	(0.005)
t-4	0.002	(0.007)	0.007	(0.004)	-0.081	(0.013)	-0.060	(0.005)
t-3	0.000	(0.007)	0.003	(0.003)	-0.077	(0.012)	-0.046	(0.005)
t-2	0.000	(0.006)	0.003	(0.003)	-0.044	(0.011)	-0.031	(0.004)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	0.665	(0.042)	0.572	(0.025)	0.153	(0.018)	0.190	(0.005)
t+1	-0.179	(0.038)	-0.141	(0.025)	-0.242	(0.022)	-0.025	(0.006)
t+2	-0.243	(0.038)	-0.163	(0.025)	-0.362	(0.024)	-0.085	(0.007)
t+3	-0.262	(0.038)	-0.223	(0.030)	-0.449	(0.027)	-0.118	(0.008)
t+4	-0.274	(0.040)	-0.233	(0.026)	-0.471	(0.029)	-0.159	(0.009)
t+5	-0.270	(0.041)	-0.280	(0.036)	-0.488	(0.032)	-0.170	(0.010)
t+6	-0.290	(0.043)	-0.310	(0.040)	-0.521	(0.036)	-0.187	(0.011)
t+7	-0.257	(0.045)	-0.334	(0.043)	-0.528	(0.041)	-0.202	(0.013)
t+8	-0.274	(0.053)	-0.404	(0.052)	-0.520	(0.054)	-0.213	(0.018)
Observations	3,576,708		3,623,292		3,576,708		3,623,292	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (2) 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 Sun and Abraham (2021)'s estimator. 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: Event Study Coefficients for Log Cost by Service Category

	Prescriptions		Outpatient		Inpatient	
	coef	se	coef	se	coef	se
t-11	-0.001	(0.001)	-0.048	(0.005)	0.001	(0.005)
t-10	-0.002	(0.001)	-0.046	(0.004)	-0.007	(0.004)
t-9	-0.001	(0.001)	-0.044	(0.003)	-0.009	(0.003)
t-8	-0.001	(0.001)	-0.044	(0.003)	-0.013	(0.003)
t-7	-0.002	(0.001)	-0.044	(0.003)	-0.016	(0.003)
t-6	-0.003	(0.001)	-0.038	(0.002)	-0.019	(0.003)
t-5	-0.003	(0.001)	-0.036	(0.002)	-0.019	(0.002)
t-4	-0.002	(0.001)	-0.030	(0.002)	-0.019	(0.002)
t-3	-0.002	(0.001)	-0.024	(0.002)	-0.020	(0.002)
t-2	-0.002	(0.001)	-0.012	(0.002)	-0.017	(0.002)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	0.102	(0.003)	0.307	(0.005)	0.719	(0.008)
t+1	-0.007	(0.002)	-0.035	(0.004)	-0.062	(0.006)
t+2	-0.013	(0.002)	-0.055	(0.004)	-0.122	(0.006)
t+3	-0.018	(0.002)	-0.070	(0.005)	-0.145	(0.006)
t+4	-0.026	(0.003)	-0.085	(0.005)	-0.166	(0.007)
t+5	-0.025	(0.003)	-0.091	(0.006)	-0.184	(0.007)
t+6	-0.034	(0.003)	-0.106	(0.006)	-0.213	(0.008)
t+7	-0.041	(0.004)	-0.109	(0.008)	-0.229	(0.010)
t+8	-0.042	(0.006)	-0.126	(0.011)	-0.254	(0.013)
Observations	7,200,000		7,200,000		7,200,000	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (2) for log total spending in prescriptions in column (1), outpatient services in column (2), and inpatient services in column (3). Estimation uses Sun and Abraham (2021)'s estimator. 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 for Making Any Claim by Service Category

	Prescriptions		Outpatient		Inpatient	
	coef	se	coef	se	coef	se
t-11	-0.041	(0.009)	-0.135	(0.014)	-0.029	(0.008)
t-10	-0.050	(0.006)	-0.118	(0.010)	-0.041	(0.006)
t-9	-0.042	(0.005)	-0.116	(0.008)	-0.042	(0.005)
t-8	-0.044	(0.004)	-0.110	(0.007)	-0.053	(0.004)
t-7	-0.040	(0.004)	-0.093	(0.006)	-0.050	(0.004)
t-6	-0.039	(0.004)	-0.074	(0.006)	-0.050	(0.004)
t-5	-0.036	(0.003)	-0.066	(0.005)	-0.050	(0.004)
t-4	-0.030	(0.003)	-0.054	(0.005)	-0.044	(0.003)
t-3	-0.027	(0.003)	-0.041	(0.004)	-0.039	(0.003)
t-2	-0.021	(0.003)	-0.024	(0.004)	-0.032	(0.003)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	0.142	(0.004)	0.125	(0.005)	0.435	(0.005)
t+1	-0.019	(0.004)	-0.043	(0.006)	-0.046	(0.005)
t+2	-0.050	(0.004)	-0.098	(0.007)	-0.110	(0.005)
t+3	-0.055	(0.005)	-0.138	(0.008)	-0.128	(0.005)
t+4	-0.065	(0.005)	-0.175	(0.009)	-0.143	(0.006)
t+5	-0.064	(0.006)	-0.185	(0.009)	-0.157	(0.006)
t+6	-0.076	(0.007)	-0.209	(0.011)	-0.175	(0.008)
t+7	-0.089	(0.008)	-0.225	(0.013)	-0.180	(0.009)
t+8	-0.104	(0.011)	-0.242	(0.017)	-0.198	(0.012)
Observations	7,200,000		7,200,000		7,200,000	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (2) for making any prescription claims in column (1), outpatient claims in column (2), and inpatient claims in column (3). Estimation uses Sun and Abraham (2021)'s estimator. 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 8: Event Study Coefficients for Laboratory and Imaging Services

	Log cost + 1				Any claim			
	(1) Laboratory		(2) Imaging		(3) Laboratory		(4) Imaging	
	coef	se	coef	se	coef	se	coef	se
t-11	-0.005	(0.002)	-0.015	(0.003)	-0.103	(0.013)	-0.071	(0.012)
t-10	-0.008	(0.001)	-0.014	(0.002)	-0.108	(0.009)	-0.079	(0.008)
t-9	-0.008	(0.001)	-0.013	(0.002)	-0.108	(0.008)	-0.086	(0.007)
t-8	-0.007	(0.001)	-0.015	(0.001)	-0.101	(0.007)	-0.090	(0.006)
t-7	-0.006	(0.001)	-0.015	(0.001)	-0.086	(0.006)	-0.073	(0.005)
t-6	-0.005	(0.001)	-0.013	(0.001)	-0.078	(0.005)	-0.060	(0.005)
t-5	-0.005	(0.001)	-0.012	(0.001)	-0.075	(0.005)	-0.055	(0.005)
t-4	-0.005	(0.001)	-0.010	(0.001)	-0.060	(0.005)	-0.047	(0.005)
t-3	-0.004	(0.001)	-0.008	(0.001)	-0.048	(0.005)	-0.031	(0.004)
t-2	-0.003	(0.001)	-0.007	(0.001)	-0.030	(0.005)	-0.025	(0.004)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	0.057	(0.002)	0.103	(0.002)	0.163	(0.006)	0.199	(0.005)
t+1	-0.014	(0.001)	-0.021	(0.001)	-0.084	(0.006)	-0.078	(0.005)
t+2	-0.020	(0.001)	-0.030	(0.002)	-0.116	(0.006)	-0.117	(0.006)
t+3	-0.022	(0.001)	-0.033	(0.002)	-0.127	(0.007)	-0.124	(0.006)
t+4	-0.026	(0.002)	-0.035	(0.002)	-0.148	(0.007)	-0.134	(0.007)
t+5	-0.028	(0.002)	-0.038	(0.002)	-0.159	(0.008)	-0.157	(0.007)
t+6	-0.034	(0.002)	-0.042	(0.002)	-0.193	(0.009)	-0.167	(0.008)
t+7	-0.034	(0.002)	-0.045	(0.003)	-0.178	(0.011)	-0.174	(0.010)
t+8	-0.038	(0.003)	-0.047	(0.004)	-0.219	(0.015)	-0.171	(0.014)
Observations	7,200,000		7,200,000		7,200,000		7,200,000	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (2) for log total spending and an indicator for making claims in laboratory services and imaging services. Estimation uses Sun and Abraham (2021)'s estimator. 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 9: Event Study Coefficients Excluding Deaths

	Mean claim cost		Any claim	
	coef	se	coef	se
t-11	0.015	(0.007)	-0.127	(0.016)
t-10	0.010	(0.006)	-0.122	(0.012)
t-9	0.010	(0.005)	-0.118	(0.009)
t-8	0.008	(0.005)	-0.111	(0.008)
t-7	0.005	(0.005)	-0.094	(0.007)
t-6	0.005	(0.004)	-0.084	(0.007)
t-5	0.005	(0.004)	-0.077	(0.006)
t-4	0.003	(0.003)	-0.059	(0.006)
t-3	0.000	(0.003)	-0.054	(0.005)
t-2	0.001	(0.003)	-0.034	(0.005)
t-1	(ref)	(ref)	(ref)	(ref)
t	0.467	(0.022)	0.183	(0.006)
t+1	-0.117	(0.019)	-0.047	(0.007)
t+2	-0.133	(0.017)	-0.115	(0.008)
t+3	-0.175	(0.024)	-0.154	(0.009)
t+4	-0.203	(0.026)	-0.196	(0.010)
t+5	-0.219	(0.028)	-0.206	(0.011)
t+6	-0.240	(0.031)	-0.236	(0.013)
t+7	-0.255	(0.033)	-0.256	(0.015)
t+8	-0.295	(0.040)	-0.275	(0.021)
Observations	4,796,616		4,796,616	

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (2) for mean claim cost and an indicator for making claims excluding individuals who die during the sample period. Regression uses Sun and Abraham (2021)'s estimator. 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 10: Zero-price and Health Shock Effects Event Study Coefficients

	Zero-price		Health Shock	
	coef	se	coef	se
t-11	0.064	(0.013)	-0.018	(0.005)
t-10	0.061	(0.011)	-0.016	(0.005)
t-9	0.062	(0.010)	-0.017	(0.004)
t-8	0.048	(0.009)	-0.015	(0.003)
t-7	0.040	(0.009)	-0.016	(0.003)
t-6	0.040	(0.008)	-0.016	(0.003)
t-5	0.037	(0.008)	-0.016	(0.002)
t-4	0.028	(0.007)	-0.014	(0.002)
t-3	0.019	(0.006)	-0.014	(0.002)
t-2	0.014	(0.005)	-0.009	(0.002)
t-1	(ref)	(ref)	(ref)	(ref)
t	0.721	(0.030)	0.701	(0.025)
t+1	-0.196	(0.027)	0.070	(0.012)
t+2	-0.228	(0.024)	0.039	(0.008)
t+3	-0.305	(0.034)	0.026	(0.006)
t+4	-0.347	(0.039)	0.023	(0.007)
t+5	-0.397	(0.044)	0.022	(0.009)
t+6	-0.448	(0.051)	0.022	(0.014)
t+7	-0.491	(0.052)	0.029	(0.013)
t+8	-0.600	(0.069)	0.022	(0.011)
Observations	6,818,472		91,428	

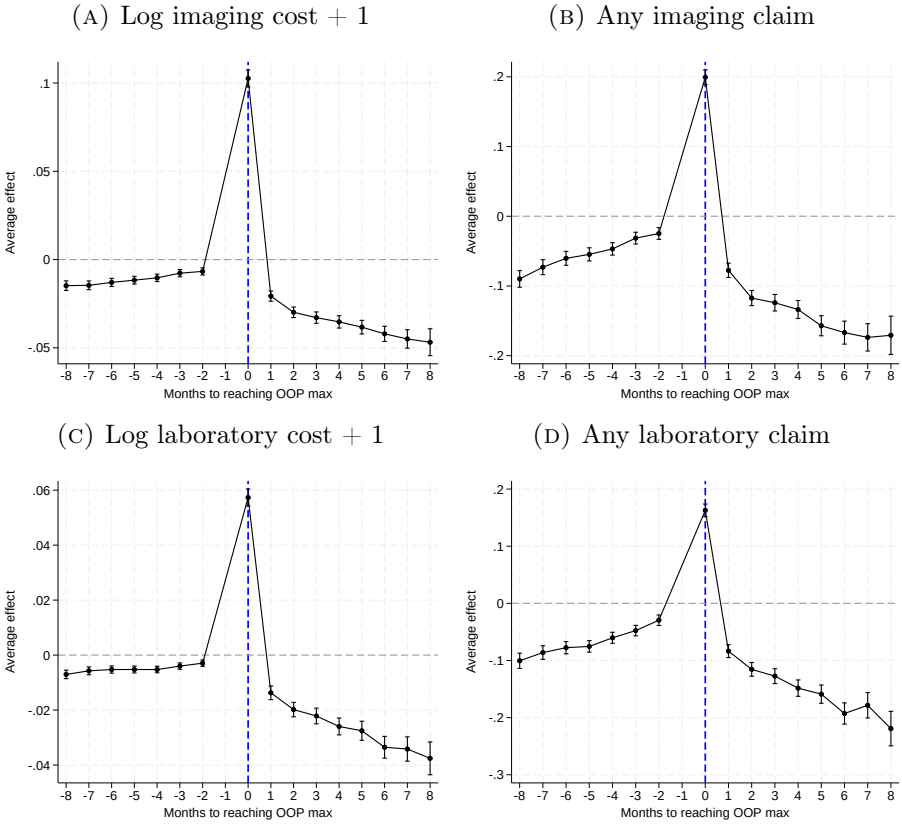
Note: Coefficients and standard errors in parenthesis of the event study specifications for the mean claim cost due to the zero-price effect and due to the health shock effect. The zero-price effect uses the sample of individuals who are never hospitalized during the sample period. The health shock effect uses the sample of treated individuals. Estimation uses Sun and Abraham (2021)'s estimator. Time indicators are constructed relative to reaching the OOP maximum for the zero-price effect and are relative to the month when the individual is hospitalized for the health shock effect.

Appendix Table 11: Event Study Coefficients by Cohort

	Mean claim cost			Any claim		
	(1) Month 4	(2) Month 7	(3) Month 9	(4) Month 4	(5) Month 7	(6) Month 9
t-8	—	—	0.007	—	—	-0.094
			(0.007)			(0.014)
t-7	—	—	0.004	—	—	-0.081
			(0.006)			(0.014)
t-6	—	0.007	-0.002	—	-0.034	-0.074
		(0.007)	(0.006)		(0.016)	(0.014)
t-5	—	0.004	0.001	—	-0.037	-0.063
		(0.007)	(0.006)		(0.015)	(0.014)
t-4	—	0.002	0.001	—	-0.033	-0.054
		(0.006)	(0.006)		(0.014)	(0.013)
t-3	0.002	-0.004	-0.006	-0.004	-0.027	-0.056
	(0.010)	(0.006)	(0.005)	(0.014)	(0.013)	(0.013)
t-2	0.004	0.010	-0.005	-0.011	-0.023	-0.038
	(0.010)	(0.007)	(0.005)	(0.013)	(0.012)	(0.012)
t-1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
t	0.870	0.595	0.552	0.021	0.034	-0.031
	(0.067)	(0.060)	(0.046)	(0.023)	(0.022)	(0.015)
t+1	-0.001	-0.113	-0.136	-0.264	-0.231	-0.287
	(0.047)	(0.026)	(0.021)	(0.026)	(0.026)	(0.018)
t+2	-0.134	-0.116	-0.153	-0.343	-0.295	-0.377
	(0.023)	(0.028)	(0.016)	(0.028)	(0.028)	(0.019)
t+3	-0.179	-0.155	-0.186	-0.409	-0.360	-0.419
	(0.021)	(0.029)	(0.026)	(0.030)	(0.031)	(0.020)
t+4	-0.187	-0.181	—	-0.459	-0.402	—
	(0.023)	(0.030)		(0.032)	(0.030)	
t+5	-0.198	-0.184	—	-0.495	-0.444	—
	(0.024)	(0.032)		(0.033)	(0.032)	
t+6	-0.230	—	—	-0.541	—	—
	(0.020)			(0.033)		
t+7	-0.247	—	—	-0.562	—	—
	(0.025)			(0.034)		
t+8	-0.266	—	—	-0.585	—	—
	(0.026)			(0.036)		
Observations	7,032,744	7,035,552	7,037,808	7,032,744	7,035,552	7,037,808

Note: Coefficients and standard errors in parenthesis of the event study specifications following equation (2) for mean claim cost and an indicator for making claims conditional on treated individuals who reach their OOP maximum in April in columns (1) and (4), July in columns (2) and (5), and September in columns (3) and (6). Estimation uses Sun and Abraham (2021)'s estimator. 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 2: Utilization and spending by health service



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (2) 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. Regression uses Sun and Abraham (2021)'s estimator. 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.

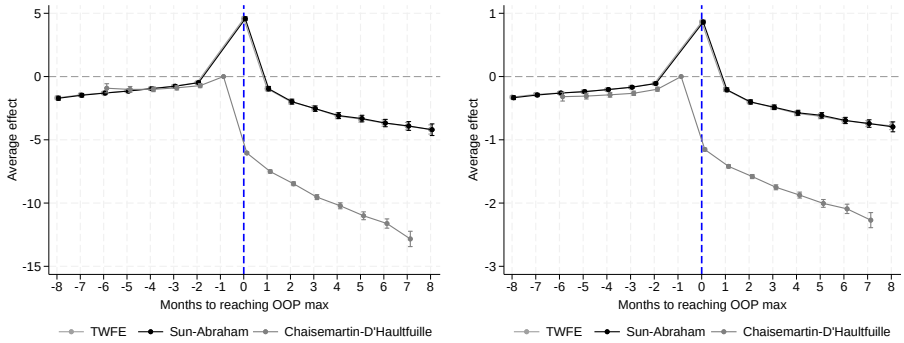
Appendix 3 Robustness Checks

This appendix presents event study coefficients and 95 percent confidence intervals for my main outcomes using Sun and Abraham (2021)'s estimator, two-way fixed effects, and De Chaisemartin and d'Haultfoeuille (2020)'s estimator.

Appendix Figure 3: Utilization and spending after reaching the OOP limit

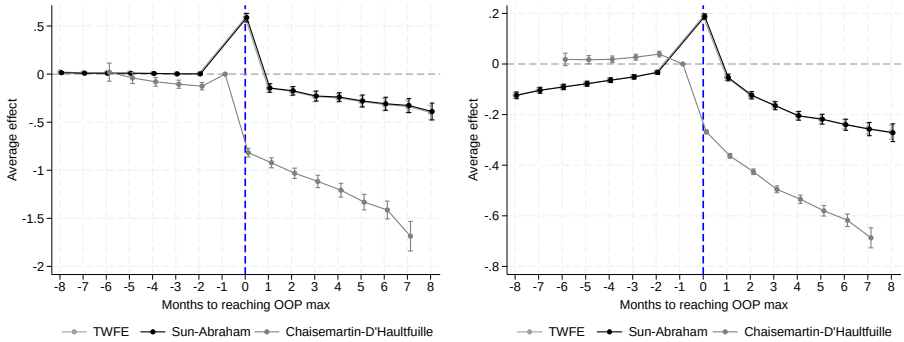
(A) Log total spending + 1

(B) Log total claims + 1



(C) Mean claim cost

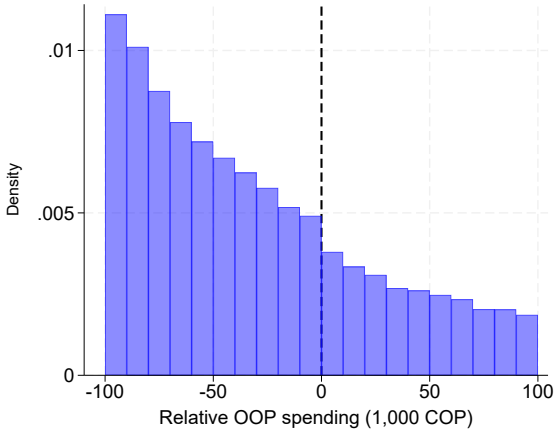
(D) Any claim



Note: Figure shows coefficients and 95 percent confidence intervals of the event study specifications following equation (2) 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. Regression uses Sun and Abraham (2021)'s estimator in black, De Chaisemartin and d'Haultfoeuille (2020)'s estimator in dark gray, and two-way fixed effects in light gray. Time indicators relative to reaching the OOP maximum are set to −1 for the control group.

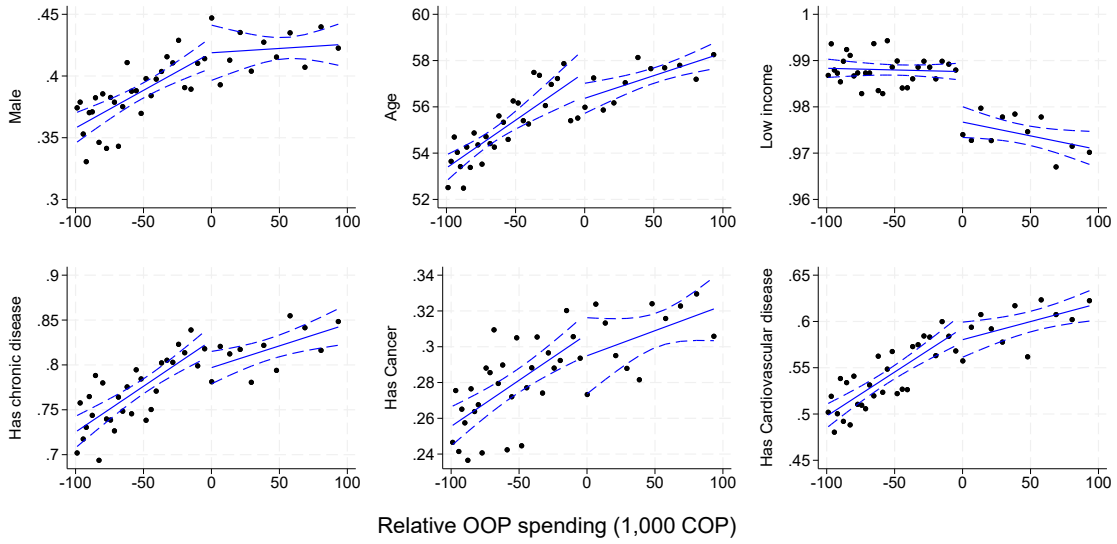
Appendix 4 Regression Discontinuity Assumptions

Appendix Figure 4: Histogram of relative OOP spending



Note: Histogram of OOP spending relative to the OOP maximum in the full sample.

Appendix Figure 5: Regression discontinuity on demographics



Note: Regression discontinuity plot using the full sample, 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 5 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:

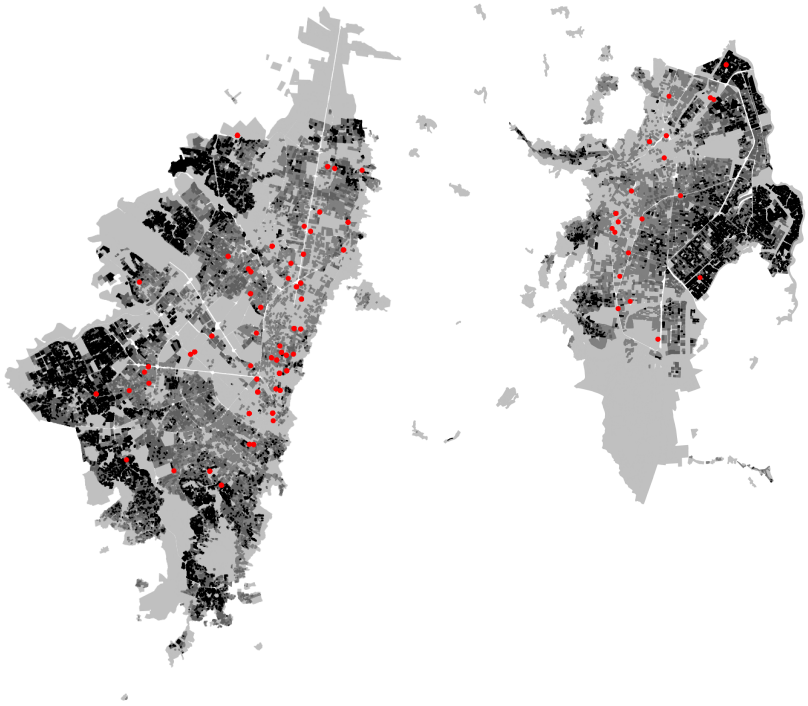
$$p_{cjh} = \lambda_1 + x_c'\lambda_2 + \lambda_h + \upsilon_{cjh}$$

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 6 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 6: Hospital locations and census tracts

(A) Bogotá (B) Cali



(C) Medellín (D) Barranquilla

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.