**Title**
Essays in Health Economics



**Author**
Atal Chomali, Juan Pablo

**Publication Date**
2016

Peer reviewed|Thesis/dissertation

**Essays in Health Economics**

by

Juan Pablo Atal Chomali

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Emmanuel Saez, Co-Chair
Professor Benjamin Handel, Co-Chair
Professor Patrick Kline
Professor Jonathan Kolstad

Spring 2016

**Essays in Health Economics**

## Abstract

Essays in Health Economics

by

Juan Pablo Atal Chomali

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Emmanuel Saez, Co-Chair

Professor Benjamin Handel, Co-Chair

These essays study how private incentives affect the functioning of three dimensions of health care markets: health insurance, prescription drugs, and the delivery of health care by physicians.

In the first chapter, I study the workings of long term health insurance, a form of contracts with the potential to efficiently insure individuals against reclassification risk, but at the expense of other limitations like provider lock-in. I empirically investigate the workings of long-term guaranteed-renewable contracts, which are subject to this tradeoff. Individuals are shielded against premium increases and coverage denial as long as they stay with their initial contract, but those that become higher risk are subject to premium increases or coverage denials upon switching, potentially leaving them locked-in with their original network of providers. I provide the first empirical evidence on the importance of this phenomenon using administrative panel data from the universe of the private health insurance market in Chile, where competing insurers offer guaranteed-renewable plans. I fit a structural model to yearly plan choices, and am able to jointly estimate evolving preferences for different insurance companies and supply-side underwriting in the form of premium risk-rating and coverage denial. To quantify the welfare effects of lock-in, I compare simulated choices under the current rules to those in a counterfactual scenario with no underwriting. The results show that consumers would be willing to pay around 13 percent more in yearly premiums to avoid lock-in. Finally, I study a counterfactual scenario where guaranteed-renewable contracts are replaced with community-rated spot contracts, and I find only minor general-equilibrium effects on premiums and on the allocation of individuals across insurers. I argue that these small effects are the result of large levels of preference heterogeneity uncorrelated to risk.

In the second chapter, David Silver and I study worker interactions among the medical staff in the emergency department. Using rich administrative case-level data from two hospital-based emergency departments, we start by documenting peer effects among physicians. We find that physicians are 1.5 percent faster when working with peers who are 10 percent faster. We devise a test for random patient-physician assignment and we provide a

number of tests to discern the mechanisms underlying these spillovers. The evidence points to spillovers that are driven primarily by faster peers responding negatively to working with slower peers. Utilization of shared resources accounts for little of the spillover, and event-study evidence points to spillovers that come into effect as soon as slower peers begin their shifts.

In the third chapter, José Ignacio Cuesta, Morten Sæthre and I study regulations to pharmaceutical laboratories in the form of bioequivalence (BE) requirements – the most prevalent tool used in developed economies to ensure the effectiveness of generic drugs allowed in the market. The main goal is to empirically investigate how the market reacts to BE requirements, and the consequences in prices, market shares, and product availability for branded and generic drugs. In particular, this study is an early exploration of the experience of Chile, where BE requirements were adopted for 172 molecules, leading to the BE approval of 642 generic drugs between March 2009 and March 2015. We show that the introduction of the requirements lead to a significant increase in BE approvals and in the share of BE drugs in the market. However, prices and market shares of other competing drugs were not significantly affected during the period we analyze. Other outcomes, like number of products, and market concentration are also found to be unaffected.

*Para May y Fernando*

*Para Daniela*

# Contents

# List of Figures

# List of Tables

# Acknowledgments

I would not have been able to finish this dissertation without the help of numerous people who contributed in many different and important ways. I extend my deepest gratitude to all of them in the following lines.

To my advisor Ben Handel, for sparking my interest in the field of Health Economics, and for his invaluable contribution to this project and to my development as a researcher. His constant encouragement and insightful suggestions made this dissertation possible, in spite of all the difficulties I encountered.

To Emmanuel Saez, for his guidance, and enormous generosity; to Patrick Kline for his invaluable advise particularly in the most difficult moments; and to Jon Kolstad for his support and thoughtful insights.

To all the people that also contributed to this project with useful feedback and suggestions. In particular to Allyson Barnett, Marika Cabral, David Card, William Dow, Igal Hendel, Kei Kawai, and Jen Kwok.

This research would not have been possible without generous financial support from the Center for Equitable Growth, CONICYT, and the Center for Labor Economics.

To all the people that helped me with datasets and institutional details. In particular to Alexis Aceituno, Roberto Arce, Joaquín Brahm, Manuel Espinoza, Patricio Huenchuñir, Paula Jara, Ernesto Muñoz, Marcela Pezoa, Eduardo Salazar, Marlene Sánchez, Niccolo Stagno, María Teresa Valenzuela, and Suley Vergara.

My friends and co-authors José Ignacio Cuesta, Morten Sæthre and David Silver deserve special recognition. I am extremely grateful for the opportunity to work with them, for all the hard work they have put into our projects, and for all they have taught me along the way.

To Adam, Danny, Justin, and MJ, for making these '10,000' days more enjoyable.

To all the amazing friends I was surrounded with during this period of my life, for your smiles, words of encouragement and all the fun moments. In particular to Alisa, Anders, Dorian, Miguel, Moisés, Pierre, Simon, Torsten, and Youssef. To the *Lokos* and the *Goal Diggers*. A los roommates de la 'Milvia House' Andrés, Nati, Darko; a la 'Berkeley Family', Gabi, Pato, Paulina y Juan; y a Mirentxu.

A Philippe, por acompañarme en cada paso de este proceso, así como lo ha hecho en todos los aspectos de mi vida.

Gracias Mamá, Papá, Raimundo, Vicente, Ignacio y Gabriela, por su amor incondicional, y por estar siempre apoyándome en esta búsqueda, aún cuando nos ha mantenido alejados físicamente más de lo que pensábamos.

A Daniela, por haber sido durante este proceso fuente inagotable de sonrisas, versos y amor.

# Chapter 1

# Lock-in in Dynamic Health Insurance Contracts : Evidence from Chile

## 1.1 Introduction

When health insurance contracts are of limited duration, changes in health status might lead to substantial increases in premiums or subsequent coverage denial.[1] Ensuring affordable coverage for individuals reclassified by the market into a higher risk group was among the most important goals of the Affordable Care Act (ACA). The ACA eliminates barriers to purchasing insurance faced by sick individuals by prohibiting all forms of differential pricing or coverage denial based on preexisting conditions.[2] However, ruling out this type of discrimination (known as "community rating") leads more costly individuals to sort into more comprehensive plans, which in turn prices out lower risk individuals (Akerlof, 1970, Handel, Hendel, and Whinston, 2015a). The ACA tackles this adverse-selection problem by making the purchase of health insurance mandatory –one of the most controversial aspects of the regulation.[3]

In theory, an unregulated marketplace with properly designed long-term contracts can also deliver protection against reclassification risk without adverse selection. In particular, Pauly, Kunreuther, and Hirth (1995) show that long-term individual agreements with guaranteed renewability can provide full insurance against medical expenses and reclassification risk. By paying a premium higher than what they would otherwise pay under a short-term contract, individuals acquire the guarantee of affordable coverage in the future, regardless of any potential negative health shocks. However, if individuals leave their long-term contract to buy another in the spot market, they lose that guarantee and must pay premiums according to their current health status. This financial incentive to remain in the long-term

[1] Hendren, 2013 cites that between 2007 and 2009, prior to the implementation of the Affordable Care Act, one in seven applications to the four largest insurance companies in the US non-group market were rejected. See Hendren's paper for the theoretical rationale for coverage denial.

[2] Under the ACA premiums can only be adjusted by age and geographic location

[3] The individual mandate requires that most individuals obtain health insurance or pay a tax penalty

contract potentially leaves participants inefficiently tied to their original plan and/or health service provider. For the remainder of this paper, I will use the term "locked-in" to refer to an individual who would prefer to switch provider networks if the offers she faced across insurers were not differentially risk-rated, but is unable (due to coverage denial) or unwilling to do so given the discrepancy between the premium she pays under her guaranteed renewable contract and those faced in the spot market.

This paper quantifies the inefficiency from insurer lock-in generated by guaranteed renewable contracts. Consider an individual that enters a contract with a narrow network of providers while healthy, but later develops a chronic condition (such as cancer) that would be better treated in a specialty clinic outside the network. Absent reclassification in the spot market, the individual would switch to a contract with a more appropriate network. In the presence of reclassification, however, the individual might find it prohibitively expensive to switch, and might even be denied coverage outside of her current contract. An individual who does not switch because of this underwriting is effectively locked-in with her network, generating a welfare loss.

Although health insurance contracts in most markets are short term, there are a couple of practical experiences of markets with guaranteed-renewable (GR) arrangements. Prior to the ACA, GR contracts were particularly common in the US individual health insurance market, and were required by the federal Health Insurance Portability and Accountability Act (HIPAA) of 1996 (Herring and Pauly (2006), Pauly and Herring (1999)). However, lack of regulation with respect to the evolution of non-price features made these contracts generally unappealing to sick customers (Handel, Hendel, and Whinston, 2015b). GR contracts are also present in the German private health insurance market (Hofmann and Browne (2013)).

Proponents of guaranteed renewability have recognized that the inability of riskier individuals to switch insurers is potentially welfare-reducing (Patel and Pauly (2002)). However, quantifying its importance is empirically challenging, and requires individual-level panel data on health insurance purchases and claims, with the ability to observe individuals after they switch insurers. This paper, which is the first to my knowledge to quantify the welfare effects of lock-in, takes advantage of a unique panel data set from the universe of the private health insurance market in Chile in which contracts are guaranteed-renewable. The data contains individualized claim records for all enrollees (and their dependents) in each of the private health insurance companies that operate in Chile. Importantly, the claims data contains detailed procedure codes and identifiers of each health service provider.

It is *a priori* plausible that valuations over private insurance companies in Chile differ across individuals, and that individual's preferences may change over time, as they develop specific health conditions. Four stylized facts from the data support this idea. First, individuals enrolled in different insurance companies access different heath care providers. Second, networks differ in the extent to which they give access to different providers for different types of claims. For example, two networks can have similar providers for routine claims but different providers for cancer-related procedures.[4] Third, switching companies strongly

[4]For instance providers labeled as "$P_1$", "$P_5$" and "$P_{11}$" in the data are the major three providers for

predicts seeing a new health care provider : the probability that the same individual sees a new health care provider after switching insurance company increases by around 40% the month after switching. Fourth, high-risk switchers generally switch to different companies than low-risk switchers.

Individuals's valuations for these companies are also likely to change as individuals move geographically. The market share of each of these companies varies substantially across geographical regions, and is strongly correlated with the presence of in-network providers (details in section 4.1).

In order to show why welfare-reducing lock-in could occur in this market, I begin by developing a simple two-period model of guaranteed-renewable contracts with evolving preference heterogeneity. This model will also allow me to identify the data objects needed to empirically assess the magnitude of this inefficiency. The shape and evolution of preferences for companies over time will determine the share of individuals that would eventually switch companies absent reclassification. The extent of premium risk-rating and coverage denial in the spot market, and the path of health expenditures over an individual's lifetime determine the share of potential switchers that are effectively locked-in.

I then show that the main features of guaranteed-renewable contracts are present in the Chilean private insurance market, and provide reduced-form evidence of lower switching rates among the highest risk individuals. This result is implied by the theory of guaranteed-renewable contracts, since healthier individuals are not subject to coverage denial or premium increases due to risk-rating in the spot market, and thus are not constrained in switching insurance contracts.

In order to quantify the welfare losses resulting from lock-in, I use a structural model that jointly estimates plan choices and underwriting in the spot market. The model incorporates heterogeneous and evolving preferences for firms that are correlated with health status and geographic location. Individuals also evaluate the potential benefits of switching plans by the degree of overlap between their current provider network and that of the alternative offers they face. I account for guaranteed-renewability by allowing the plan that an individual chooses in year $t$ to always be available to that individual in year $t+1$ regardless of any changes in health status. However, individuals also receive risk-rated offers in the spot market each year. I estimate the level of risk-rating in the spot market by empirically linking each plan's characteristics to the health risks of those who switch into that plan. The arrival rate of offers from competing insurers in the spot market depends on an individual's preexisting conditions, consistent with the possibility that insurance companies may deny coverage based on these conditions. Finally, the model also allows for "choice inconsistencies" (*à la* Jackson Abaluck and Gruber, 2011) and "inertia" –two well-known behavioral biases affecting health insurance purchase (see e.g., Handel, 2013 and Jason Abaluck and Gruber, 2013). Inertia

routine claims for insurance companies "A" and "B", with an accumulated share of approximately 40% routine claims in each company. For cancer-related treatment, another provider "$P_4$" is the most frequent provider for company $B$ with a share of 33 percent, although $P_4$ represents only 9 percent of those claims in Company A (These numbers were computed using the claims data for the providers in the Metropolitan Region of Santiago during 2011.).

is potentially an important factor in choice persistence in this market where search costs are likely to be high. In the spirit of Ching, Erdem, and Keane, 2009 and Grubb and Osborne, 2015, I model inertia as inattention that restricts the choices that individuals actually consider in each period. Finally, I estimate age and gender specific health transition matrices that allow me to simulate the path of health expenditures over an individual's lifetime.

I estimate the parameters of the structural model by adapting the Geweke-Hajivassiliou-Keane (GHK) simulator (Keane, 1993; Keane, 1994, Hajivassiliou, McFadden, and Ruud, 1996, and J. Geweke and Keane, 2001) to an environment where, as a result of guaranteed renewability, the evolution of options available in each period is dependent on the choices made in the preceding periods, and where the first period choice is not necessarily observed (left-censored choices). To quantify the welfare loss resulting from lock-in, I use the estimated parameters of the model along with the estimated risk profiles to simulate the path of choices over time. Then, I compare predicted choices under the current underwriting rules against a counterfactual scenario in which coverage denial based on preexisting conditions is banned and there is no premium risk rating in the spot market (as recently regulated by the ACA)

The results suggest that around 5% of individuals end up locked-in with their insurance company, but an individual on average would be willing to pay 13% of the average yearly premium to avoid the possibility of this lock-in. Locked-in individuals would be willing to pay on average 2.5 times the premium to switch plans. I find that most of the lock-in is due to coverage denial based on preexisting conditions, and only mildly caused by spot premiums which have been adjusted to reflect current health status. The results depend mostly on the estimated level of coverage denial in the spot market. My estimates suggest that one in five individuals with a preexisting condition is denied coverage in the spot market. Next, I investigate potential market unraveling following endogenous sorting across insurance companies if a community rating policy were adopted. I simulate a counterfactual situation without underwriting in the spot market, but allowing for the overall level of premiums in each company to be adjusted in order to maintain profits per enrollee at their current level. In this counterfactual scenario, guaranteed-renewable contracts are replaced with community-rated spot contracts, but premiums are adjusted at the insurer level to reflect the changes in each insurer's risk pool. I find only minor general-equilibrium effects on prices and allocation of individuals across insurers.

I use the estimated parameters to interpret these findings and to shed light on how different aspects of preference heterogeneity affect the desirability of guaranteed-renewability. Although time-varying preferences generate lock-in, preferences that are persistent over time have the opposite effect, showing another margin of interaction between preference heterogeneity and health insurance design (Bundorf, Levin, and Mahoney, 2012 and Geruso, 2013). I use the stylized model of Einav and Finkelstein, 2011 to formalize this insight: in this static setting, preference heterogeneity within health status reduces the share of high-risk individuals willing to enroll in a contract, compared to a situation in which health status and preferences are perfectly correlated. This limits the mechanical effects of banning coverage denial and the resulting adverse selection, as the share of risky individuals that enroll

when allowed is smaller than in the case in which health status is the only determinant of preferences.[5]

This paper is related to two strands of literature. First, it draws from the theoretical literature of pricing in one-sided commitment contracts with adverse selection. It highlights the advantage of "time consistent" contracts suggested by Cochrane, 1995, which eliminate reclassification risk and ensure access to health insurance for sick people through a severance payment payable after the diagnosis of a long-term illness. Time-consistent contracts are fully portable (and thus do not generate lock-in), but at the expense of potentially large up-front borrowing (Handel, Hendel, and Whinston, 2015b). By quantifying the presence of lock-in and resulting welfare loss, this paper puts in perspective the advantages of time consistent contracts relative to guaranteed-renewability in relation to their portability.

Second, the paper contributes to the sparse empirical literature on the dynamics of health insurance contracts. Koch, 2011 and Handel, Hendel, and Whinston, 2015a focus on the tradeoff between adverse selection and reclassification risk by analyzing consecutive short-term marketplaces, like the ACA. Instead, I focus on long-run guaranteed renewable arrangements. These contracts have been empirically studied by Herring and Pauly, 2006 and Marquis et al., 2006, who demonstrate that the relationship between premiums and health claims shows the patterns predicted by the theory in the individual US market for health insurance. Handel, Hendel, and Whinston, 2015b study the welfare implications of the financial aspects of different long-term arrangements. In particular, they quantify the welfare loss resulting from the inability to effectively smooth consumption which is implied by front-loading vis-à-vis other market arrangements in a setting with imperfect capital markets. Instead, my focus is on lock-in due to evolving preference heterogeneity, assuming away capital market imperfections. Also, Crocker and Moran, 2003 show that employment-based health insurance can facilitate the existence of long-term contracts even in the absence of front-loading. Adverse selection is reduced since healthy individuals would have to switch jobs to drop their health insurance contract and buy another that does not cross-subsidize the riskier.

Relatedly, Hendel and Lizzeri, 2003 provide evidence of front-loading in life-insurance contracts, and show that more front-loading is associated with higher retention rates (less "lapsation") which allows for the retention of a healthier risk pool. Finkelstein, McGarry, and Sufi, 2005 focus on the long-term care insurance market, and show that lower risk individuals are most likely to leave the contracts. This selective lapsation generates dynamic inefficiencies, as individuals who stay in the contract have to pay a premium consistent with dynamic adverse selection, which undermines the protection against reclassification risk. This paper contributes to this line of research by providing further evidence of selective switching, and explicitly using data on both supply and demand for health insurance to estimate how it relates to welfare.

[5]Studies that find high levels of market unravelling produced by community rating generally analyze environments in which insurance contracts are purely a financial product (e.g., Handel, Hendel, and Whinston, 2015a).

Finally, this paper is also related to Pardo and Schott, 2012; Pardo and Schott, 2013 who use survey data to analyze enrollment decisions across the private and public sector in Chile, and the role of preexisting conditions in these decisions. Their results, complementary to this paper, suggest that few individuals would switch sectors if restrictions on preexisting conditions were eliminated. They argue that most of the sorting across sectors can be explained by heavy asymmetries in the premium structure, since the public sector offers important cross-subsidies for families, women, and low-income users. This paper focuses instead in lock-in within the private sector, which permits me to estimate heterogeneity in preferences for insurance companies with the same structure of financial incentives. Also, I have access to detailed administrative claims data which permit good assessment of health conditions and to study the general-equilibrium effects of counterfactual policies.

The remainder of the paper is organized as follows. Section 1.2 intuitively explains the working of guaranteed-renewable contracts and the sources of welfare loss implied by these contracts when preferences vary over time. In section 1.3, I provide an overview of the main institutional details of the Chilean health insurance system. Section 1.4 describes the data and provides reduced-form evidence to support the idea that preferences for Chilean private health insurance companies are heterogeneous and evolve over time. Section 1.5 presents empirical evidence that the main features of guaranteed-renewability are present in the Chilean market, making it a suitable environment for empirical analysis. I also provide reduced form evidence of the link between health risk and switching rates across companies. Section 1.6 presents a structural model of plan choice that incorporates the main features of guaranteed-renewability and underwriting in the spot market. Section 1.7 discusses the parameter estimates, quantifies lock-in and its welfare implications, and simulates a counterfactual policy with community rating. Section 1.8 concludes.

## 1.2 Economics of guaranteed-renewable contracts and welfare consequences of lock-in

In this section, I use a simple model with two firms and two periods to highlight the main welfare consequences of lock-in in markets with guaranteed renewable contracts. In start with a simple adaptation of the seminal work on guaranteed renewability by Pauly, Kunreuther, and Hirth, 1995 that assumes no preference heterogeneity (section 1.2), and then incorporate the features that generate lock-in (section 1.2).

### No preference heterogeneity

There are two periods and two health types $(L, H)$, with corresponding probability of an adverse health event $p_L$ or $p_H > p_L$. An adverse event entails a monetary loss equal to $C$. In period $t = 1$ everyone is of type $L$. Individuals become type $H$ in period $t = 2$ if they have a negative health event in period one, otherwise they stay type $L$. Importantly, there is one side-commitment (i.e., individuals cannot be forced to stay in the contract in period two) and

symmetric information between insurers and consumers (such that all insurers can observe the resolution of uncertainty regarding the period-2 status). Individuals are risk-averse, and there is no discount rate.[6]

In this setting, Pauly, Kunreuther, and Hirth (1995) show that a competitive marketplace can offer long-term contracts that provide full insurance against medical expenditures and against reclassification risk. In period one, insurers will sell contracts to healthy individuals at a premium that is equal to the expected value of their period-two claims, and guarantee the renewability at a premium equal to the expected value of their period-1 claims. That is, the sequence of premiums $(P^{GR}_{t=1}, P^{GR}_{t=2})$ offered in $t = 1$ is front-loaded: $P^{GR}_{t=1} = p_L C(1+(p_H - p_L))$ and $P^{GR}_{t=2} = p_L C < P^{GR}_{t=1}$.

Insurance companies selling these guaranteed-renewable contracts can provide full coverage each period and break-even in expectation, since the price in period one covers the expected loss in period two and the price in period two covers the expected loss in period one. Although several premium profiles would make the firms break-even, this particular profile also complies with *no-lapsation constraints*. The lack of consumer commitment requires that, in every state in the second period, the consumers prefer to stay in the long-term contract rather than switching to a competing insurance company. In fact, individuals receive offers for short-term (spot) contracts in period two. Because of perfect competition, these contracts are offered at a premium that is equal to period-two expected claims conditional on each individual's updated type. Formally, spot contracts are offered in period $t = 2$ for individual $i$ at $P^{SPOT}_{i,t=2} = p_i C \geq P^{GR}_{t=2}$. With prices $(P^{GR}_{t=1}, P^{GR}_{t=2})$, all individuals (and in particular the healthy) have incentives to purchase the long-term contract in period one and remain in the contract in period two, since spot contracts in period two are (weakly) more expensive.

The above results explain why guaranteed-renewable contracts attain the first-best allocation when there is one-sided commitment, and non-binding borrowing constraints.[7] In the absence of non-financial plan characteristics, individuals do not switch in equilibrium, effectively eliminating reclassification and adverse selection. In the next section I extend this model by incorporating evolving preference heterogeneity.

## Evolving preference heterogeneity

As already suggested by Patel and Pauly, 2002, guaranteed renewable contracts are not perfect

> "[...] since some high risk individuals could find themselves locked in with an insurance they *have come to dislike*" (Patel and Pauly, 2002, pp. 289, emphasis added).

[6] A more general theoretical analysis is provided by Krueger and Uhlig, 2006. In particular they show that the relative discount rates between the principal (insurance company) and the agent (individual) is crucial in determining the level of insurance for the agent in the long run.

[7] When it is costly for individual to pay this front-loaded set of premiums, long-term contracts provide only imperfect insurance against reclassification risk (Hendel and Lizzeri (2003)) and Handel, Hendel, and Whinston (2015b)

Those individuals that become high risk are reclassified in the spot market, and would be required to pay a potentially much higher premium if they were to switch insurers.

To formalize Patel and Pauly's observation I extend the previous model by allowing for preferences for insurers that are heterogeneous across the population, and are permitted to change from period one to period two. In this extension, there are two firms, $A$ and $B$. Each firm offers a guaranteed-renewable contract in period $t = 1$ (when everyone is healthy) and risk-rated (type-specific) spot contracts in $t = 2$. Contracts are distinguished only by the premium and the firm that offers them (all other characteristics are the same).

Panel (a) of Figure 1.1 shows the distribution of the relative willingness to pay for firm $A$ in the population in each period. The $y$-axis represents the willingness to pay in period one, and the $x$-axis the willingness to pay in period two. Evolving preference heterogeneity is represented by an imperfect (although positive) correlation between the willingness to pay in each period, leading to an ellipsoidal shape oriented to the north-east.

I assume that preferences and the competition environment are such that the price (and marginal costs) in period one and period two of the guaranteed renewable contract are the same for both firms. This simplification is not essential but it facilitates the graphical analysis. Also, I assume individuals do not take into account the uncertainty about period-two taste shocks in their period-one decision, or that the distribution of taste shocks in period two is such the optimal strategy in period one corresponds to the optimal strategy of a myopic individual. [8]

Under these assumptions, individuals in the two upper quadrants choose firm A in period one. Because of guaranteed-renewability, those individuals have the option of staying with firm A in period two by paying a premium equal to the expected cost of type-$L$ individuals (expected loss from period one), regardless of their health status.

The offer that enrollees in A get from B in period two depends on their health status. Panel (a) of Figure1.1 represents the case of type-$L$ individuals (healthy enrollees), who are also offered a spot contract in firm B at the same price as the price they pay in period two under the long-term contract in A. Absent any choice frictions (like search or switching costs), healthy individuals in the left quadrant will choose firm A in period two, whereas those in the right quadrant will choose firm B. By the same argument, each quadrant representing two-period preferences also represents the choices of health individuals. Since marginal costs are the same, it follows that healthy individuals are efficiently allocated across companies.

Consider now those individuals who had an adverse health event in period one and therefore are reclassified as type $H$ (risky) in period two, as represented in Panel (b) of Figure 1.1. These individuals pay an extra premium $\Delta P$ to switch across companies, equal to difference between the premium of the spot contract for the risky and the guaranteed-renewable contract, $\Delta P = (p_H - p_L)C$. In the figure, the distance between the dashed lines and the x-axis corresponds to this difference.

Risky individuals in the upper left quadrant chose firm A in period one and would switch

[8] This could happen for instance if individuals are fully myopic or if the distribution of taste shocks in period 2 is symmetric around zero. I will return to discuss the plausibility of this assumption in Section 6.

to firm B in period two if they were charged the guaranteed-renewable price in both firms. However, as shown in the figure, a share of those individuals will stay with firm A even they prefer firm B in period two, since they are not willing to pay the extra premium to switch. The dashed area of the left-upper quadrant represents these individuals, who I define as being "locked-in to firm A". Similarly, individuals represented by the dashed area in the lower-right quadrant are "locked-in to firm B": they chose B in period one and would have switched to firm A had they been charged the same price in both companies.

As it is typically done in the literature of guaranteed-renewable contracts, I have assumed an environment in which health type is revealed over time symmetrically to all market participants. In this case, a firm and a risk-averse agent will always be willing to trade at a premium equal to the expected cost, regardless of the risk type. However, Hendren, 2013 shows that failures to this assumption (i.e., when applicants to an insurance contract hold private information) can explain enrollment denial. Another potential reason for coverage denial, and also potentially relevant for this paper - is firm's inability to fully risk-rate their plans through legal impediments or menu costs. I analyze the effect of denying enrollment on lock-in in Panel (c).[9] In that case, all individuals in the upper left and bottom-right quadrant cannot switch away from the firm they picked in period one even if it would be efficient for them to do so.

In this simple analysis, the degree of inefficiency produced by lock-in depends on the following factors: First, it is contingent on the shape of preference heterogeneity dynamics, as depicted by the shape of the shaded area in Figure 1.1. It also depends on the share of individuals who are subject to reclassification (i.e those for which the relevant situation is described in panel (b) of Figure 1.1), and the extra premium they pay upon switching, $\Delta P$. Finally, inefficiency resulting from lock-in depends on the share of individuals who are denied coverage if they were willing to switch, as represented by the share of individuals in situation described by panel (c) of Figure 1.1. In section 1.6 I describe the main empirical framework to quantify these objects in the data.

The existence of welfare losses in guaranteed-renewable contracts is at odds with previous theoretical analysis that has shown that they achieve pareto-optimality. That analysis makes the strong assumption of no preference heterogeneity for providers, so it abstracts from any non-financial motive to switch plans over time. This is also true if preferences are heterogeneous but constant over time, as would be the case when the shaded area of Figure 1.1 shrinks to a line, as shown in panel (d) of Figure 1.1. In that case, there are no individuals willing to switch companies over time and thus no possibility of lock-in.

[9] See Cabral, 2015 for dynamic inefficiencies caused by dynamic asymmetric information.

## 1.3 Institutional framework

The Chilean health-care system is divided into a public and private system.[10] The public regime, FONASA, is a pay-as-you-go system financed by the contributions of affiliates and public resources. The private sector –operated by a group of insurance companies known as "Instituciones de Salud Previsional" or ISAPRES –is a regulated health insurance market.[11] FONASA covers more than two thirds of the population (about 11 million people), while ISAPRES covers around 17 percent. The remainder of the population is presumed to be affiliated with special healthcare systems such as those of the Armed Forces or to not have any coverage at all (Bitran, Escobar, and Gassibe (2010)).

Workers and retirees have the obligation to contribute 7 percent of their wages to the public system, or to buy a plan that costs at least 7% of their wages in the private system.[12] The two systems differ in many respects, including provider access, premiums, coinsurance structure, insurer payment caps, exclusions, and quality. Affiliates of FONASA are classified into four groups based on wages and family composition. These groups determine copayment for each service (which ranges from 0-20 percent), but otherwise benefits are unrelated to income. Unlike the private sector, there are no exclusions based on preexisting conditions, nor pricing based on age or gender, and there is no additional contribution for dependents. As a consequence, the private sector serves the richer, healthier, and younger portion of the population (Pardo and Schott, 2012).

The private health insurance market is comprised of 13 ISAPRES, which are classified into two groups: six "open" (available to all workers) and seven "closed" (available only to workers in certain industries). Open ISAPRES account for almost 95 percent of the private market. When workers enroll in a health insurance contract under an ISAPRE, they must immediately select a specific plan. Contracts in the private sector are, for the most part, individual arrangements between the insured and the insurance company. The contracts are yearly, although those who have already been enrolled for one year may switch to another ISAPRE or to the public sector at any time.

The monthly premium $P$ is a combination of a base price $P_B$ and a risk-rating factor $r$ so that

$$P = P_B \times r \tag{1.1}$$

where $r$ is a gender-specific and discontinuous (step) function of age.

Several features of the plan determine the base price $P_B$. A plan has two main coverage features: coinsurance rates (one for inpatient care and another for outpatient care) and coverage caps (insurer payment caps). Every plan assigns the insurer a per-service payment cap, and these caps apply to each visit. Coinsurance rates and the insurer payment caps

[10]The details of the Chilean health care system have already been described elsewhere, in particular Duarte (2012) and Dague and Palmucci, 2015. I draw from those papers heavily in this section.

[11] From this point forward, I will refer to a private insurer that is part of this group as an "ISAPRE", and the group of insurers collectively as "ISAPRES".

[12]With a cap of 186 USD per month

remain constant across visits and do not accumulate over time. For any particular claim, a person pays her coinsurance rate until the amount that the insurance company contributes reaches the cap for that service. After hitting the cap, the patient pays the rest. The basic formula to determine the copay is therefore:

$$copay = price - minimum([price \times (1 - coinsurance)], cap)$$

Base prices are indexed to inflation, and adjustments to the base price in real terms can be made once a year. Three months before the end of the contract year, ISAPRES must inform the regulator of their projected price increases for the year. Each company must also inform their clients about these increases, justify their reasons for the changes and offer alternative contracts to their clients that maintain monthly premiums but that often imply lower coverage.

A couple of features of the market restrict the extent to which private firms can risk-rate their plans. First, base prices are set at the plan (and not the individual) level. Also, since a major reform to the system in May 2005 –the "ley larga de ISAPRES" –each firm can have at most two $r$ functions. However, there is a large number of plans in the market (around 52 thousand with active enrollees and around 18 thousand actually offered in the market at a given point in time), so the effective number of insureds per plan is fairly small (on average 28). A large share of plans has only one insured (40 % as of January of 2011). Although the spirit of the regulation is not to allow risk-rating through the base price, I evaluate this issue empirically.

The "ley larga" introduced a major restriction that limits the extent of reclassification of individuals already in a contract: the price increase of each particular plan $j$ in ISAPRE $k$ cannot be higher than 1.3 times the average price increase of all plans of ISAPRE $k$.[13] Formally,

$$\Delta\%P_{jk} \leq 1.3 \times \overline{\Delta\%P_k} \tag{1.2}$$

where $\Delta\%P_{jk}$ is the percentage change in the base price of plan $j$ in ISAPRE $k$ and $\overline{\Delta\%P_k}$ is the average price increase for all plans in ISAPRE $k$.

As shown in section 4, rule (1.2) effectively limits the variation in price increases and therefore the extent of reclassification risk.

**Preexisting Conditions** Each new potential insured has to fill a "Health Declaration" before signing a new contract with a private firm. The companies are allowed to deny coverage of any preexisting condition during the first 18 months of enrollment, or even to reject the prospective enrollee altogether. Although there is no available data on the extent to which ISAPRES deny coverage, anecdotal evidence and conversations with industry actors suggests that this is a regular practice. Note that preexisting conditions are relevant for

[13]There is a small share of plans (5%) that are not subject to this rule. In the empirical part I use only the sample of plans for which this rule applies

switching across ISAPRES or into an ISAPRE from the public sector, but not within a given ISAPRE.[14]

**Networks** Individuals have access to different types of plans with respect to the provider network. "Preferred-provider" plans are tied to a specific network, although enrollees can use providers outside of their insurers network at a higher price (similar to PPO in the US). Individuals can also choose - at a higher premium - plans with an unrestricted network of providers. Under these "free choice" plans, coverage is not tied to the use of a particular clinic or health care system, similar to a traditional fee for service indemnity plan in the United States. Companies also offer a small share of "closed network" plans, where enrollees can only use the services of the plan providers or must pay full price (the equivalent of the U.S. HMO).

## 1.4 Data

I have access to an administrative dataset containing the universe of insureds in the private market for the period 2009-2012. This dataset is sent by ISAPRES to the regulatory agency (*Superintendencia de Salud*) and was made available through a research partnership. The data contains the stock of policyholders each month (around 1.5 million per month), including basic demographics (age, gender, number of dependents, district of residency), wage (capped to the contribution limit) and plan choice. I have access to the universe of claims in each month for each individual and his or her dependents. Claim information includes total cost, insurer cost, copayment, provider identification and geographical location, and a claim (procedure) code. The data also includes biannual information on the stock of all "active" plans in the market, which is updated in January and July. Active plans are defined as those that are either currently sold in that month or that have been discontinued but are still held at least by one enrollee. The information on the plans include the company, premiums, the adjustment factor $f$, and the date at which the plan was introduced in the market. I provide the main descriptive statistics of the data in section 6.5 when I describe the construction of the data for estimation.

### Provider differentiation and provider switching

In this section I provide empirical evidence suggesting that the valuation for private insurance companies in Chile differ across individuals, and that individual's preferences change over time, in particular as they develop specific health conditions and as they move geographically. This evidence supports the notion that evolving preference heterogeneity –the main driver

[14]Other important characteristics of the plans are : a) Capitation scheme: Plans can either be capitated or not, b) Maternity-related expenses: Some plans do not have coverage for maternity-related expenses. Policyholders can opt for these plans and pay a lower premium.

of lock-in –is likely to be important in the context analyzed in this paper, and motivates the remainder of the paper.

First, enrollees to health insurance companies that are subject to this study access different networks, particularly for some types of conditions. To illustrate this fact with a concrete example, in Table 1.1 I list the most frequent providers of cancer treatment for each of the six companies and the shares of cancer-related claims of each company treated by these providers in the Metropolitan region of Santiago[15]. For each company, the list includes the largest cancer treatment providers in descending order up to the point where providers jointly account for 80 % of the treatment or more. I also include, for each provider in each company, the share of claims related to other procedures. Companies are labeled as $A$, $B$, .., $F$ and providers as $P_1$, $P_2$, ..., $P_{16}$.

Although there are some discrepancies across companies, a few common patterns emerge from this table: First, cancer-treatment procedures are concentrated in a handful of providers: more than 80% of the claims are treated by 5-7 providers, depending on the company. These are most likely big hospitals with a high degree of complexity. While these hospitals treat an important share of cancer-related procedures, their participation in treating other types of claims is on average significantly smaller. The cumulative share of other procedures treated by these providers is fairly small in some of the companies (12 % percent in company $F$ and 10 % in company $D$), although higher in companies $C$ and $E$ (58% and 41 % respectively).

The relative importance of each provider varies across insurance companies and depends on the type of claim. Consider for instance the case of companies $E$ and $F$, and how they are linked to providers $P_7$ and $P_4$. Provider $P_7$ is the main provider seen by patients enrolled in a plan under $E$ for cancer-related procedures: 55 percent of such procedures are performed by provider $P_7$ (along with 14 % of other types of procedures). Provider $P_7$ is, however, an infrequent destination of enrollees in company $F$: it treats 5 % of cancer-related claims and 3 % of non-cancer claims. Most of the cancer claims of enrollees from company $F$ (56 %) are treated by another provider, $P_4$, that does not treat a significant share of other types of claims in $F$, nor a significant share of claims of enrollees in $E$. Thus, differences between $P_7$ and $P_4$ are likely to be especially relevant in shaping the preferences over companies $E$ over $F$ for individuals in need of cancer treatment. However, a healthy individual comparing these two companies upon entering a contract might not consider her taste for $P_4$, since she is unlikely to utilize it unless she develops cancer.

This table suggests that the provider networks vary across insurance companies, and that network differences are specific to the nature of the treatment. However, it does not rule out that individuals sort perfectly across insurance companies in relation to their contingent preferences for a given network.

In practice, though, individuals *do* switch companies. Do individuals that switch companies also switch providers? I answer this question by constructing a monthly panel data sample of individuals from 2009-2012, and follow their enrollment as well as the providers

[15]This region corresponds to nearly 2/3 of the Chilean private market. These figures were computed using all the claims data from 2011

they see. This dataset contains a 10% random sample of enrollees in companies $A$-$E$ by january 2009 (including enrollees in the entire country). For reasons I explain later, I exclude enrollees from company $F$.[16] To evaluate whether individuals that switch companies also switch providers, I run an event-study specification, where the event corresponds to an insurance-company switch, and the outcome of interest is whether the individual sees a provider she hasn't seen before.[17] Specifically, I run:

$$newprovider_{it} = \sum_k \beta_k D_{it}^k + \psi_t + \theta_i + \epsilon_{it} \tag{1.3}$$

where $newprovider_{it}$ is a dummy variable equal to 1 if $t$ is the first time that individual $i$ sees the provider seen in month $t$. Since the records of provider utilization are left-censored in january 2009, I construct the variable $newprovider_{it}$ using the entire time span but estimate the parameters only on data from the years 2011 and 2012.[18] In $\psi_t$ I include month-year time dummies, $\theta_i$ is an individual effect, and $\epsilon_{it}$ is an error term.

The $D_{it}^k$ are a series of "event-time" dummies that equal one when an individual switches company $k$ periods away. Formally,

$$D_{it}^k = \mathbb{1}(t - s_i = k)$$

where $s_i$ is the month individual $i$ switched.

The $\beta_k$ coefficients represent the time path of the probability of seeing a new provider relative to event of switching insurance company. [19] I estimate equation 1.3 by ordinary least squares and I normalize $\beta_{-1} = 0$, since the inclusion of individual fixed-effects make the $D$'s perfectly collinear. I also place the following endpoint restrictions:

$$\beta_k = \begin{cases} \bar{\beta} & \text{if } k \geq 7 \\ \underline{\beta} & \text{if } k \leq -6 \end{cases}$$

Figure 1.2 plots the estimated $\beta_k$ coefficients from a regression of the form given in Equation 1.3, where the dependent variable is the dummy for new provider, with the corresponding 95% percent cluster-robust confidence intervals. The results show a significant increase in the probability of seeing a new provider after switching insurance companies. From a baseline probability of 35%, the probability increases by 13 percentage points the month after switching. The probability of seeing a new provider continues to be higher than the baseline for 5 months after switching, after which it stabilizes to its pre-switching level. This result emphasizes that individuals do switch companies, and when they switch, they

[16] Since an individual may see more than one provider in a month, I use the provider with the largest amount of claims in the month per individual.

[17] This specification follows closely Kline, 2012.

[18] Although this issue should be largely addressed with the time dummies, the results are robust to running OLS only on the 2012 data.

[19] For simplicity, I drop all the observations after an individual switches companies for a second time during the period, which corresponds to about 4 % of observations.

do see different providers. This is evidence against the simplifying assumptions that make guaranteed-renewable contracts "perfect", which require no heterogeneity across companies or stable preferences for them over time.

As another piece of reduced-form evidence showing that lock-in with a given insurer is potentially important, I show that destination companies among switchers are different across individuals with different pre-switching health status. That is, high-risk switchers generally switch to different companies than those preferred by the low-risk when they switch. First, I show this evidence by calculating differences by health status in the net flows into each company. I define net flow into company $k$ as the difference between the number of people switching into $k$ and the number of people switching out of $k$, as a share of total switchers. The net flows for the monthly panel sample are in Table 1.2. For instance, during the sample period, healthy individuals switch in net out of company B, with a net flow of -8.5 % . On the other hand, individuals with preexisting conditions disproportionately switch into company B, with a net flow of 6.1 %. Differences in net flows across health status are statistically significant for all companies except for company A. As an additional test, I show in Table A.1 the results of estimating a multinomial logit on the sample of switchers for the probability of choosing each of the companies, as a function of health status and other demographics. Several specifications robustly show that pre-switching health condition is a statistically significant determinant of the destination company.

Finally, the data shows that geographical location is also a potentially important determinant of individual valuation's for each company. Chile is split in 346 districts, which belong to one of 53 provinces, which in turn belong to one of 15 regions. Panel (a) of Figure 1.3 plots the market share of the 6 open ISAPRES by region for the 10 largest regions in terms of number of insureds. For instance, while company 1 has around 20 percent of the market share in the largest region, it has only 10 percent of the market in the second largest. In panel (b) I show the market shares of each company also varies substantially across districts of region 1 (Metropolitan Area). Company 1 has a market share of around 30 percent in the largest district, 18 percent in the second largest and around 12 percent in the third largest. Overall, these pictures suggests that there are substantial differences across districts in the relative valuation for difference insurance companies. To understand the variation of preferences across geographic in section A.2 I show that market shares at the district level are positively correlated with the presence of in-network providers in each district. The presence of an in-network provider in the district is associated with a 12 higher market share.

## 1.5 Empirical evidence on guaranteed renewability

In this section I show evidence that the main features of guaranteed-renewable contracts are present in the Chilean health plans. The findings in this section motivate the remainder of the paper that estimates the welfare consequences of these contracts using a structural choice model. In particular I show evidence that 1) there is low reclassification for individuals who

remain in their contract, 2) premiums are front-loaded, and 3) individuals buy contracts in a "spot market". In the estimation section, I also show that premiums in the spot market are correlated with an individual's previous health expenditures even after controlling for plan generosity.

**Reclassification risk** Private firms are not permitted to unilaterally cancel an individual's contract. Moreover, they cannot change the contract's characteristics. However, in principle, they could effectively force out an enrollee by a large enough increase in her premium. The rule described by equation 1.2) aims to eliminate this possibility, by constraining the variance of price increases: the price increase of a single plan cannot exceed 1.3 times the average price increase in the corresponding company.

There are many ways in which ISAPRES could effectively comply with this constraint. To show how this constraint works in practice, in Figure 1.4 I show histograms with the distribution across plans of yearly (real) price increases (in percentage points) for the period 2010/2011, which is representative of the pattern for all years in the sample.[20] Although there are many possible ways to comply with 1.2, Figure 1.4 shows that in practice companies pick only a handful of price increases to apply to most of the contracts. This practice limits the correlation between individual health shocks and individual price increases, which implies very limited reclassification.

**Front-loading** I show evidence of front-loading by looking at the evolution of premiums relative to health claims for individuals who stay with their insurance company. Let $h_{it}$ be the total claims (insurer cost) in period $t$ of individual $i$, and $P_{it}$ the corresponding premium. I show that the ratio $rat_{it} = h_{it}/P_{it}$ is increasing in $t$.[21] This test is equivalent to Hendel and Lizzeri, 2003's, who report that the ratio of yearly premium to probability of death in the life insurance market shows a decreasing pattern over time. Similarly, Marquis et al., 2006 show evidence of front-loading in California's individual market by showing that among longer-term enrollees, families that include an adult who contracts a chronic medical condition after enrolling in the individual market pay less than families with a chronically ill adult at enrollment.

Still, decreasing markups is a *strong* test of front-loading, as even in its the presence, markups could increase over time if individuals display enough inertia (as is often the case in health insurance markets, see e.g., Handel, 2013 and Jackson Abaluck and Gruber, 2011;

[20] In the interest of space, I do not show the distribution for other years, but they are available upon request

[21] As Herring and Pauly, 2006 argue, front-loading does not necessarily imply a decreasing premium schedule. Premiums can increase only to reflect the increase in the spot price of the healthy individuals. Instead, front-loading means that the (expected) markup decreases as individuals stay in the contract. Since the theory predicts full insurance, there is no distinction between total cost and insurer cost. However, since individuals that stay in the same contract keep their coverage rates, the distinction is not relevant for testing the dynamics of either one relative to premiums. The results of Table 1.3 are robust to using total cost instead of insurance costs.

Jason Abaluck and Gruber, 2013). In markets with consumer inertia, firms are expected to use an "invest-then-harvest" pattern for prices, i.e., start charging a low price and increase it over time (Farrell and Klemperer, 2007).[22] In the context of one-sided commitment, guaranteed renewable contracts combined with an "invest-then-harvest" strategy do not imply unambiguous price patterns. Intuitively, inertia relaxes the no-lapsing constraint that is needed to incentivize the healthier to stay. Therefore, firms can charge in period two a price that is above the actuarially fair premium for the healthy type. This increased revenue in period two is passed on to the first period in the form of lower premiums.[23] Moreover, the evidence I provide is limited to the first 4 years of enrollment.

I test the hypothesis of increasing markups using the monthly panel of the sampled individuals enrolled in January 2009 and followed until December 2012.

As is common in health expenditures data, $h_{it}$ (and therefore $rat_{it}$) has a significant zero mass and is heavily skewed. In this setting, the use of generalized linear models (GLM) has become popular to deal with the undesirable properties of standard OLS methods or two-part models.[24] I estimate the model using the method of generalized estimating equations (GEE), which extends GLM to take into account potential within-individual correlation (Blough, Madden, and Hornbrook, 1999 and Nedler, 1989). I specify a log link and gamma distribution with an AR(1) process. In a first specification I estimate

$$log(E(rat_{it})) = \alpha + \beta \times T_{it} \tag{1.4}$$

with $r_{it} \sim \Gamma$. I also investigate whether the parameter $\beta$ varies across age groups by interacting $T_{it}$ with three age groups (as of January 2009) ; 20-35, 36-45, and 45+ :

$$log(E(rat_{it})) = \sum_g \mathbb{1}(agegroup_i = g)(\alpha_g + \beta_g \times T_{it}) \tag{1.5}$$

The parameter estimate pooling age groups is $\hat{\beta} = .078(0.007)$ corresponding to a marginal effect of an extra year enrolled of .050(0.005). The results allowing different slopes for each age group, in the second column of Table 1.3, indicate that the slope of $rat_{it}$ is not statistically different across groups.

Overall, these results indicate that markups decrease over time as individuals stay enrolled in the same plan, as suggested by the theory of guaranteed renewable contracts.

**Spot markets** One key aspect of markets with guaranteed-renewability is that consumers buy contracts in a spot market, where contracts are tailored to their risk. I argue that in the Chilean market environment this spot market exists.

Plans are constantly created. Between January 2009 and December 2011, ISAPRES created on average more than 5400 plans per year. The constant creation of plans allows

[22] Indeed, Ericson, 2012 shows that premiums in Medicare Part D plans follow this pattern.

[23] Thus, a test for invest-then-harvest pricing strategies in the context of guaranteed renewability would look for the presence of potential savings that healthy enrollees would make conditional on switching

[24] See Buntin and Zalavsky, 2004 for a review of the methods handling skewed health care cost.

ISAPRES to potentially have plans with slightly different features but different coverage rates and premiums.

A couple of features of this market support the notion that insureds are not free to shop across all the plans, and their choice set is restricted to the offers made by insurance companies to them. First, there is no centralized resource where price quotes are available from all the plans offered by different companies. Survey data on plan choice in this market shows that around 70 % of individuals chose among a few options offered by a sales agent (Criteria Research, 2008).

The dynamics of plan purchase are also consistent with the existence of an active spot market. To show how the purchase of plans relate to the date at which plans launched in the market, I split plans into different "plan cohorts", defined as a function of the date when they were created. In particular, I split plan into 6 different groups: group 1 contains all plans created before July 2007, group 2, plans created between July 2007 and June 2008, and so on, until group 6, which contains plans created between July 2011 and June 2012.

In Figure 1.5 I show the share of switchers in date $t$ that switch to plans of each cohort. This exercise is conducted for the same random sample of enrollees in January 2009. It shows, for instance, that almost 90% of individuals who switched in January 2009, did so to plans created between july 2007 and june 2008 and almost 10% switched to plans created after July 2008 (a negligible share of individuals switched to older plans). The pattern is repeated over time: the majority of switchers in a given moment in time switch only to relatively new plans. Figure 1.6 shows a similar pattern for switchers within insurance companies.

## Switching and health status

It is expected in GR contracts that switching rates will be decreasing in health expenditures. To test this hypothesis, I estimate a proportional hazard model for the probability of switching companies using the monthly panel of individuals enrolled in January 2009.[25]

I identify individuals with a preexisting condition using the claims data. The data contains a detailed procedure code that I link to medical conditions that are typically considered by ISAPRES as preexisting conditions.[26] In Table 1.4 I show a list of the six conditions considered, as well as their prevalence in the data (column 2). As shown in Table 1.4, the prevalence rates compare well with self-reported prevalence rates derived from survey data shown and shown in column (1).[27]

I estimate a proportional Cox model for the hazard rate of switching companies, $\lambda_{it}$, as a function of a set of covariates $X_{it}$

[25] In the subsample of individuals for whom I observe a complete spell within an insurance company in the period.

[26] For instance "Simple vascular access for hemodialysis" is considered to indicate Renal Insufficiency.

[27] I use the 2009 wave of Social Protection Survey, which is a nationally representative survey on a variety of issues related to social protection. This survey asks individuals if they were diagnosed with a variety of conditions, as well as health insurance enrollment. More information in `http://www.previsionsocial.gob.cl/subprev/?pageid=7185`

$$\lambda_{it} = \lambda_{0t} \times exp(\beta X_{it}) \tag{1.6}$$

where $\lambda_{0t}$ is a time-specific baseline hazard rate The results are found in Table 1.5, for different specifications. In Column (1) I only include a dummy for preexisting conditions, which is equal to 1 for every period after the first realization of a procedure related to conditions listed in Table 1.4, $\mathbb{1}(preex_{it} > 0)$. The estimated hazard ratio is 0.74, indicating that individuals with preexisting conditions are 26% less likely to switch companies. Column (2) shows that this result is robust to including a quadratic term on age and a dummy for gender. In Column (3) I add controls for contemporary measures of healthcare utilization. I compute three-month moving averages of health expenditures and create a) the logged amount of spending on preexisting conditions, $\mathbb{1}(h_{it}^{preex} > 0) \times log(h_{it}^{preex})$, and b) an indicator for any health expenditure $\mathbb{1}(h_{it} > 0)$ and its interaction with the logged value of all health expenditures $\mathbb{1}(h_{it} > 0) \times log(h_{it})$. The results reflect interesting patterns: the presence of a preexisting condition is strongly (negatively) correlated with switching rates even after controlling for the amount of contemporary expenditures. Total expenditures on preexisting conditions among individuals with preexisting conditions are slightly correlated with switching rates. On the contrary, other types of health expenditures only have an effect in the extensive margin : positive health expenditures do not predict lower switching rates, but among those with positive expenditures, higher total expenditures does predict lower switching. I interpret these results as reflecting that higher health care utilization does in general cause lower switching rates, but that preexisting conditions cause lower switching rates beyond the increased expenditures they produce. This is consistent with a model in which individuals who have claims do not want to switch companies (so as not to lose their provider) and/or higher expenditures imply higher prices in the spot market, but also that insurance companies also limit the extent to which individuals with preexisting conditions are able to switch. I will incorporate all these possibilities into the structural analysis in the next section.

## 1.6 Structural estimates

The reduced form evidence revealed differences in switching rates across health groups. In this section I turn to a structural model to jointly estimate the demand-side decision to enroll in a plan and the supply side plan offers from insurance companies depending on health status. The structural model allows me to quantify the welfare consequences of lock-in and simulate counterfactual scenarios, at the cost of several modeling assumptions.

### Discrete Choice Model

The demand-side of the model –plan enrollment given a choice menu –follows mainly Jackson Abaluck and Gruber, 2011, who provide the micro-foundations that allow to conveniently specify utility as a linear function of plan's characteristics, and at the same time incorporate

realistic departures from full optimization under full information. In addition, I incorporate heterogeneous and time-varying preferences for plans, which is the fundamental source of lock-in.

Individuals face different money lotteries and maximize flow expected utility. Let $OOP_{it}^{jk}$ be out-of-pocket costs for individual $i$ in period $t$ under plan $j$ of company $k$. $E\left(OOP_{it}^{jk}\right)$ and $Var\left(OOP_{it}^{jk}\right)$ are a function of the individual's health risk and the financial characteristics of plans (copays and caps). Assuming CARA utility with risk-aversion parameter $\gamma$ and normally-distributed cost process, Jackson Abaluck and Gruber (2011) show that an individual's utility for plan $(j,k)$ in period $t$, $U_{i,t}^{j,k}$ can be approximated by (ignoring heterogeneity in non-financial characteristics across plans):

$$U_{it}^{jk} \simeq -P_{it}^{jk} - E\left(OOP_{it}^{jk}\right) - \frac{\gamma}{2} Var\left(OOP_{it}^{jk}\right) \tag{1.7}$$

Equation 1.7 implies three main restrictions to the parameters governing utility, namely (1) a one dollar decrease in premiums is equivalent to a one dollar decrease in expected out-of-pocket expenses, (2) a $\gamma/2$ dollar increase in premiums is equivalent to a one dollar increase in the variance of out-of-pocket expenditures, where $\gamma$ is the risk-aversion parameter, and (3) financial characteristics of plans should not matter beyond their effect on the mean and variance of out-of-pocket expenditures.[28]

Following Jackson Abaluck and Gruber (2011), I use a more flexible alternative that allows for "choice inconsistencies" by relaxing the above restrictions on the parameters.

$$U_{it}^{jk} = -\beta_0 P_{it}^{jk} - \beta_1 E\left(OOP_{it}^{jk}\right) - \beta_2 Var\left(OOP_{it}^{jk}\right) + \lambda \mathbf{f^{jk}} \tag{1.8}$$

where $\mathbf{f^{jk}}$ are the financial characteristics of the plans. I also allow for individual-specific tastes for each insurance company ("brand intercepts", Berry, 1996). These intercepts, which I denote by $\alpha_{it}^{jk}$, capture consumer heterogeneity in tastes for each firm based on factors that are unobserved. These include individual-specific heterogeneity over tastes for the provider network of each company, among other factors that make an insurance company more attractive to an individual but are not directly observed in the data. By including other observable dimensions of heterogeneity in plans, $X_{it}^{jk}$, and idiosyncratic taste shocks, $u_{it}^{jk}$, the final demand specification is:

$$U_{it}^{jk} = \alpha_{it}^{k} - \beta_0 P_{it}^{jk} - \beta_1 E\left(OOP_{it}^{jk}\right) - \beta_2 Var\left(OOP_{it}^{jk}\right) + \delta X_{it}^{jk} + \lambda \mathbf{f^{jk}} + u_{it}^{k} \tag{1.9}$$

In general, the brand intercepts $\alpha_{it}^{k}$ are aimed to capture unobserved attributes for which people have heterogeneous tastes that are constant over time (Keane, 2013). It is particularly important to incorporate these factors when analyzing lock-in because stable preference

[28] Jackson Abaluck and Gruber (2011) and Jason Abaluck and Gruber (2013) show that all of these restrictions are violated in the market of Medicare Part D.

heterogeneity for ISAPRES decreases the extent to which individuals would like to switch over time. However, I also allow these intercepts to vary *deterministically* as a function of time-varying covariates $Z_{it}$, in order to capture potential sources of lock-in associated with changes in these observables. In particular, I allow health expenditures and individual geographic location to be included in $Z_{it}$, so that preferences for different firms are allowed to depend on an individual's region of residency and risk profile. These brand intercepts $\alpha_{it}^k$ are assumed to be normally-distributed, with

$$\alpha_{it}^k \sim \mathcal{N}(\alpha_0^k + \sum_{s=1}^{S} \lambda_s^k Z_{its}, (\sigma^k)^2)$$

where $\lambda_s^k$ is the differential valuation for company $k$ after a one-unit increase in the characteristic $Z_{ts}$.

The idiosyncratic taste shocks $u_{it}^{jk}$ are assumed to arise from unobserved attributes of plans for which people have heterogeneous tastes that vary over time Keane, 1997. This interpretation motivates an AR(1) specification (Keane, 2013). I allow for autocorrelation within insurance companies as well as autocorrelation within plan, such that $u_{it}^{jk} = \kappa_2 u_{it-1}^{jk} + \sqrt{1-\kappa_2} v_{it}^{jk}$ and $u_{it}^{j'k} = \kappa_1 u_{it-1}^{jk} + \sqrt{1-\kappa_1} v_{it}^{j'k}$ with $v_{it}^{jk} \sim \mathcal{N}(0,1)$ and $v_{it}^{j'k} \sim \mathcal{N}(0,1)$.

In $X_{it}^{jk}$ I include an individual-specific measure of the utility derived from the provider network of the plan. I assume that the utility of having a plan with a restricted network of providers instead of an unrestricted network is given by $-\beta_{RN}$. When switching plans from a restricted network-plan $(j,k)$ to a restricted network $(j',k')$, I assume that the disutility of doing so, denoted by $\psi$, is proportional to a "provider-distance" measure $d$ (to be defined in more detail in the following section)

$$\psi\left(N_{it-1'}^{j'k}, N_{it}^{jk}\right) = -\beta_{RN}\mathbb{1}(RN_{it}^{jk}) - \beta_d \mathbb{1}(RN_{it}^{jk}) \times \mathbb{1}(RN_{it-1}^{j'k'}) \times d^{jk,j'k'}$$

The term $\beta_{RN}$ captures the disutility of having a restricted provider network, whereas $\beta_d$ captures the extra disutility of switching to a plan that has a different provider network than the source plan. Finally, I allow for variation in the coefficients $\beta_{RN}$ and $\beta_d$ as a function of health status, to capture the fact that the "distance" across providers could matter differentially for individuals with different levels of utilization of care. Specifically, I allow $\beta_{RN,it} = \beta_{RN0} + \delta_{RN} \times log\,(1+h_i^t)$ and $\beta_{d,it} = \beta_{d0} + \delta_d \times log\,(1+h_i^t)$.

## Forward-looking behavior

This demand model abstracts away from forward-looking behavior. Forward-looking generates an option value that may affect current choices, since they affect the set of feasible future choices. Specifying a dynamic demand model would require to specify individual's perceptions about the distribution of their future preference shocks, supply-side behavior regarding reclassification, and discount rates.[29]

[29]The complexity of choice in health-insurance as well as the evidence showing choice inconsistencies in this market is arguably a main reason why most recent papers estimating health insurance demand in

I mostly worry about individuals that predict being locked-in because of high future health expenditures decide on their insurance company accordingly. As a reduced-form test for this behavior, I look at active enrollment decisions (choice of ISAPRE) in 2009 and how they correlate with future health expenditures in 2010 and 2011 controlling for current health expenditures using a multinomial logit with score $s_{it}^k = \beta X_i + \gamma_0 log(1 + h_{i,t}) + \gamma_1 log(1 + h_{i,t+1}) + \gamma_2 log(1 + h_{i,t+2}) + \beta X^k$. I cannot reject the null that $\gamma_1 = 0$ and $\gamma_2$=0 (see Table A.4).

## Construction of key explanatory variables

Some of the key variables that enter in the demand model described above are not directly observed in the data. In this section I briefly explain how I construct each of them.

**Financial characteristics: effective coverage rate** Plan-specific coverage rates are only partially observable in the dataset. Plans typically specify outpatient and inpatient copayment rates as well as per-service caps that can depend on the provider and the specific service. Since I only have access to a general copayment rate for outpatient and inpatient rate, I calculate an "effective coverage rate" $c$ (or "actuarial value"), as the share of health care costs that a health plan effectively covers using the claims data. Specifically, for each plan I calculate $c$ as the sum of all copayments and divide by total claims (insurer cost + copayment):

$$c_p = 1 - \frac{\sum_{i \in I_p} \sum_{s \in S_i} Copayment_{is}}{\sum_{i \in I_p} \sum_{s \in S_i} (Copayment_{is} + InsurerCost_{is})}$$

where $I_p$ is the set of individuals $i$ enrolled in plan $p$

and $S_i$ is the set of claims of individual $i$. For plans with a restricted provider network, I calculate a different coverage rate for in-network providers ($c_{p,in}$) and out-network providers ($c_{p,out}$). In the choice model, I estimate the weight $w$ that individuals put on each coverage rate. Specifically, financial characteristics of the plan enter as

$$\mathbf{f} = \beta_c \times (\mathbb{1}(RN = 0) \times c_p + \mathbb{1}(RN = 1) \times (w \times c_{p,in} + (1 - w) \times c_{p,out}))$$

**Out-of-pocket expenditures** I use a rational-expectations assumptions to model out-of-pocket expenditures for each individual in each plan (see e.g Jackson Abaluck and Gruber (2011) and Handel (2013)). Under this assumption, individuals predict their future health shocks based on current information available. For new enrollees, I calculate the mean and variance of health expenditures within each decile for the year following enrollment,

dynamic settings do not incorporate forward-looking behavior (e.g., Handel, 2013 or Jason Abaluck and Gruber, 2013). A recent literature uses Medicare part D dynamic pricing incentives to estimate discount factors and myopia in drug purchases, and finds strong levels of myopia (see e.g., Dalton, Gowrisankaran, and Town, 2015, Jason Abaluck, Gruber, and Swanson, 2015)

conditional on 10 deciles of health expenditures during the month of enrollment, within gender and 5-year age bins. For incumbents, I calculate the mean and variance in year $t$ conditioning on health expenditures during year $t-1$.[30]

**Network - distance across plans** Claims data allow me to determine the set of in-network providers for a given plan, as each claim identifies the provider and whether it corresponds to an in-network or an out-of-network claim. Let $N^{jk}$ be the set of in-network providers for plan $j$ in company $k$. I define the "distance" between two RN plans, in terms of their provider networks, as the number of providers that are in $(j,k)$ and $(j',k')$ over the number of providers that are in $(j,k)$ or in $(j',k')$. Formally,

$$d\left(N^{jk}, N^{j'k'}\right) \equiv 1 - \frac{\left|N^{jk} \bigcap N^{j'k'}\right|}{|N^{jk} \bigcup N^{j'k'}|}$$

The measure $d$ is equal to 1 when all the plans in the network of $(j,k)$ are also in the network of $(j',k')$. On the contrary, if there are no plans in the network of both plans, $d$ is equal to 0. In to visualize the outcome of this exercise, Figure 1.7 graphs each plan in a Euclidean two-dimensional space, where the coordinates have been calculated to reflect pairwise distances across plans.[31]

The average $d$ across distinct plans in the sample is 3.1%, but there is substantial variation in the data, both within and across insurance companies. The company with the least diversified network is company F with an average $d$ within its plans of 20%, while the company with the most diversified network is company C with an average $d$ of only 3%. Although generally plans within the same company are closer to each other than plans across companies, companies 3 and 5 have similar networks. The variation of $d$ within and across companies allows me to identify the role of the provider network separately from individual-specific brand-related tastes for insurance companies.

## Choice sets

As suggested by the survey evidence regarding plan choice in Criteria Research, 2008, the large number of plans and the impossibility of searching across all potential plans in the market means that, for most individuals, the choice set is *de facto* restricted to the menu offered by the sale agents. In this scenario, allowing individuals to choose among all plans available in the market is unlikely to recover consistent demand parameters. I handle this problem by explicitly restricting each individual's choice set before estimating demand.

First, I form each choice set to comply with the guaranteed-renewable environment. As such, the plan chosen by individual $i$ in year $t$ is always available to individual $i$ in $t+1$ at the corresponding guaranteed price. On top of their guaranteed-renewable plan, individuals

[30] I do not observe cohort "0" expenditures in 2008 to predict their claims in 2009. However, since for this cohort I only estimate choices in 2010 and 2011 conditional on the choices in 2009, this is not problematic.

[31] The representation is only determined up to location, rotations and reflections.

receive offers in the spot market that depend on their observable characteristics $\mathbf{D}$. The vector $\mathbf{D}$ includes age and gender to account for risk-rating through the $f$ function, and wage, to comply with the 7 % rule. I also allow potential offers to depend on an individual's geographic location and family composition.

Offers received in the spot market are subject to underwriting. By including age and gender in $\mathbf{D}$, I account for risk-rating through the $r$ risk-rating function. In order to allow for potential risk-rating through the base price $P_B$, I also include in $\mathbf{D}$ the overall health status, captured by individual's health expenditure quintile between $t-1$ and $t$, $h_{it}$. The extent to which firms are able to discriminate through the base price is an empirical question, but the myriad of plans available and constantly created suggest that this possibly cannot be ruled out *ex-ante*.

I allow coverage denial as a second form of underwriting, to account for the possibility that individuals with preexisting conditions are not offered plans from other firms in the spot market because of their health declaration. Specifically, incumbent individuals enrolled in company $k$ receive spot offers from other companies $k' \neq k$ with a probability that depends on whether the individual has a preexisting condition, denoted $\rho^s(\mathbb{1}(Preex_{it})$. In particular, I parametrize $\rho^s$ as:

$$\rho^s\left(1(Preex_{it}\right) = 2 \times (1 - \Phi(\theta_s \times \mathbb{1}(Preex_{it}))$$

The parameter $\theta_s$, to be estimated, captures the degree of dependency of offer rates on the presence preexisting conditions. The offer rate is equal to 1 (plans are offered to anyone regardless of preexisting conditions) when $\theta_s$ is equal to 0, and it decreases as $\theta_s$ increases. In the interest of reducing the number of parameters to be estimated, this specification makes two simplifying assumptions. First, $\rho^s$ is not permitted to depend on the insurance company, even if there might be differences in the underwriting procedures across these firms. Also, I estimate a single parameter for all preexisting conditions, although I expect that some conditions classified as preexisting, like depression, entail lower levels of coverage denial compared to conditions like cancer that are more expensive to treat. As such, $\rho^s$ reflects the average offer rate across ISAPRES for the average individual with a preexisting condition.

I summarize the supply behavior of each company $k$ as an "offer-policy",

$$\mathbb{M}^k\left(\mathbf{D}, h, \mathbb{1}(Preex_{it})\right) = \left\{P^k\left(\mathbf{X}^k, \mathbf{D}, h\right), \mathbf{X}^k\left(\mathbf{D}, h\right), \rho^s\left((Preex_{it})\right)\right\} \tag{1.10}$$

The offer policy is a function that maps demographics and health status to the characteristics of offers in the spot market. In each period $t$, an incumbent individual is confronted with an offer from each company $k$ complying with the corresponding offer policy, and the guaranteed renewable plan, $M_{it}^{GR}$. The argument $\rho^s(Preex_{it})$ is meant to capture that an individual enrolled in company $k$ receives spot offers from companies $k' \neq k$ with probability $\rho^s$.

Finally, the model allows for inertia, which is a widely-documented phenomenon in health insurance purchase (see for instance Handel (2013) and Jason Abaluck and Gruber (2013)).

In this market, inertia is potentially important considering the large number of plans available. Therefore, I model inertia as arising from "inattention", such that individuals may not necessarily see all the choices in their potential choice set when deciding to renew their plan. Arguably, other reasons besides inattention (or search costs) might cause inertia, such as habit formation, learning, or real switching costs (like hassle costs of paperwork involved) (see Handel, 2013). In practice, it is empirically difficult to disentangle among alternative explanations without direct measures of these costs or strong assumptions. In practice, I model inertia as the probability of making an active choice in every period after the first enrollment period, similar to Grubb and Osborne (2015) and Ching, Erdem, and Keane, 2009. Specifically, in every period after the first, individuals actively choose between guaranteed-renewable plans and the spot offer received from their insurance company with probability $\rho^w$. Also, among those that actively choose within their insurance company, some also search across all other insurance companies with probability $\rho^a$.[32] This specification of inertia thus operates through the plans that individuals consider in their choice set. In one of the specifications I allow $\rho^w$ and $\rho^a$ to depend on the individual's potential savings for switching in terms of premiums, $\Delta P$, so that $\rho^w = \rho_0^w + \rho_{sav}^w \times \Delta P$ and $\rho^a = \rho_0^a + \rho_{sav}^a \times \Delta P$. I also allow $\rho^a$ to depend on age, so that $\rho^a = \rho_0^a + \rho_a \times (age/\overline{age} - 1)$, where $\overline{age}$ is the average age in the sample.

Putting together the supply and demand features, I have a probabilistic choice set model for incumbent individuals. In period $t = 1$, an entrant individual makes an active choice considering offers from all the companies. However, in every period $t > 1$, an individual previously enrolled in plan $j$ of company $k$ will have a choice set that matches one of three mutually exclusive possibilities $C_{it}^1$, $C_{it}^2$, or $C_{it}^3$:

- $C_{it}^1 = M_{it}^{j1} \cup M_{it}^{j2} ... \cup M_{it}^{jK} \cup M_{it}^{GR}$: guaranteed renewable contract and spot contracts within and across insurance companies. This happens if the individual searches within and across insurance companies and is offered a plan in each. All new clients are assumed to have this choice set when they pick a plan for the first time.

- $C_{it}^2 = M_{it}^{jk_{t-1}} \cup M_{it}^{GR}$ guaranteed renewable contract and spot contract within their insurance company. This occurs if the individual (a) searches only within her insurance company or (b) searches within and across insurance companies but is not offered a plan in the other companies.

- $C_{it}^3 = M_{it}^{GR}$: only guaranteed renewable contract.

  Let $\mathbb{C}_i$ be the set of all potential choice sequences and $C_i^s$ an element of $\mathbb{C}_i$.[33] Since the probability of a given choice sequence depends on the choice set $C_i$ drawn by the individual, the overall choice probability is

[32] This is equivalent to assuming that individuals have infinite search cost with probability $\rho^w$ and $\rho^a$ for switching within and across, respectively.

[33] For instance, an individual entering the market in 2009 and observed in 2009, 2010 and 2011 has 9 potential choice set sequences; the combination of the 3 options listed above in 2010 and the same 3 options in 2011.

$$Pr\left(\mathbf{d_i}\right) = \sum_{\mathbb{C}} Pr\left(\mathbf{d}|C_i = C_i^s\right) Pr\left(C_i = C_i^s\right) \tag{1.11}$$

## Identification

Here I discuss identification of the key parameters of the model. A common identification issue in choice models in panel data is how to disentangle between the roles of state dependence from autocorrelation. I do so with functional form assumptions and exclusion restrictions that leverage time-varying covariates. Autocorrelation, modeled as an AR(1) process, implies that the probability of repeating a choice that has small observed utility decreases over time. The main exclusion restriction that allows more robust identification is that lagged premiums do not affect current utility. In the absence of state dependence, a transitory change in premiums causes at most a transitory change in the outcome, while in the presence of state dependence a transitory shock has a persistent effect in the outcome (see Hyslop, 1999).

Another identification issue is separately identifying preference heterogeneity from state dependence. In this setting, individual-specific plan characteristics help to identify preference heterogeneity using the cross sectional data because of the presence of alternative-specific premiums and coverage rates (that vary within insurance companies which is the level at which I allow unobserved preference heterogeneity). Also, the covariate-specific brand intercepts are identified from the presence of different plans in a given firm and individual-specific prices. Still, I also impose a parametric model as it is typically done to identify state dependence from preference heterogeneity Keane, 1997. In this paper, I model state dependence by assuming that individuals choose actively with a probability that is constant over time (or depends deterministically on age and potential savings) as in Grubb and Osborne, 2015. The preference parameters that enter in the flow utility are identified from the choices of individuals who enter the market (see e.g., Handel (2013)), under the assumption that unobservables are uncorrelated with premiums and characteristics.

The inertia parameters $\rho^w$ and $\rho^a$ are identified by the switching rates of healthy individuals within firms and across firms. The parameter governing offer rates $\theta$ is identified based on the assumption that inertia does not depend on having a preexisting condition, which is the key identification assumption of the model. If individuals become more aware about their plans and their incentives to search increase after acquiring a preexisting condition, the $\theta$ would be negatively biased. On the other hand, if individuals with preexisting conditions are discouraged to search because they correctly predict lower offer rates, $\theta$ would be positively biased. I discuss the sensitivity of the results to the estimated $\theta$.

As is common in these models, incumbent individuals are assumed to have the same preferences as individuals who are new entrants to the market, so that inertia is identified from differences in choice between observationally equivalent incumbents and new entrants. Although covariates of incumbent individuals do differ from new individuals entering the market, empirically there is a substantial degree of overlap. In particular, I show the kernel

density estimates of age for incumbents and new entrants (by pooling all new entrants across years) in Figure 1.8.

Finally, $\delta_{RN}$ and $\delta_d$ are identified from the gradient in switching rates across health expenditures and plan distances, using the variation in network distances within and across insurance companies.

To set the level of utility, I normalize the intercept $\alpha_{it}^5 = 0$ in equation (1.9). Normalizing the scale requires normalizing the variance of one the composite error terms, which I achieve by setting $\sigma^4 = 0$, so that $var(\epsilon^4) = 1$.

## Construction of estimation sample and descriptive statistics

I construct a yearly panel, where the year $t$ is defined to begin in September of each corresponding calendar year. I define three cohorts of "new clients" that enter the system between October of $t-1$ and September of year $t$ for years $t = 2009, 2010, 2011$.[34]

From a universe of approximately 1.5 million enrollees, there are approximately 120 thousand new enrollees each year. I perform a few sample restrictions among the new clients: I keep only individuals that have individual plans, with contracts under "open" ISAPRES (so enrollment is not limited to specific industries), and whose plans are subject to the standard pricing regime, which correspond to around 100 thousand of new enrollees per year. I only keep individuals older than 25 and younger than the corresponding legal retirement age (60 for females and 65 for males) leaving around 70 thousand new enrollees each year. Besides these sample restrictions, I drop observations with invalid or missing wages, or plan characteristics. Due to miscoded plan identifiers that resulted in difficulties in matching plans to their characteristics for one of the 6 insurance companies included in the analytical dataset, individuals buying their first plan with that company are eliminated from the sample. [35]

To the universe of new clients described above, I add a 10% random sample of incumbent clients as of September 2009 who are followed until 2011. I label these as "cohort 0". Cohort 0 is subject to the same sample restrictions detailed above. I also drop those that enrolled in the system before july 2005, before the major law ("ley larga described in section 1.3) substantially changed the pricing rules of plans. The inclusion of cohort 0 permits a richer and more representative support on the health expenditures distribution. However, as explained in Section A.4, choices of cohort 0 in 2009 cannot be estimated with an autocorrelated

[34]For each individual, I the date when she entered into a contract with her current insurer, but the date when she entered the overall system. I identify individuals that enter the system for the first time in period $t$, as those that entered a contract with a firm in $t$ and cannot be found contributing to the system in any earlier period. This method yields a cohort in 2009 that is substantially larger than those of 2010 and 2011. Since there are no structural reasons for this to be the case, I use a matching technique to get a subsample of cohort 2009 of the same size as of cohort 2010 with similar demographic distribution.

[35]Individuals switching to that company, or to other plans not considered in the sample are treated as "leaving the sample"

error structure. The likelihood of the choices of cohort 0 in 2010 and 2011 are estimated conditional on their choices in 2009

The final sample consists of approximately 313,000 individual-year pairs. The main demographic characteristics are summarized in Table 1.6.

Around 60% of the individuals in the sample are men, and the average age is 34 years old. Almost two-thirds of individuals live in Santiago. In the estimation sample, between 13 and 20 percent of individuals have a preexisting condition, depending on the cohort and year. Cohort 0 is older, has higher wages, and a higher share of individuals with preexisting conditions. On average, around 80 percent of individuals remain in the same plan from $t$ to $t+1$ and around 13 percent of individual switch within the same insurance company, so that on average 7 percent switch across insurance company from year to year.

The panel dataset contains plan choices for each individual in the estimation sample. Each individual has one (potential) spot offer from each insurance company, with the exception of those living in seven (out of fifteen) "special regions", where insurer "C" has a negligible market share. Individuals living in one of these regions are assumed to receive offers only from the other 4 companies.[36] Also, the guaranteed-renewable plan is always in the choice set of incumbent individuals. The main characteristics of the plans in the choice set are described in the Table 1.7

Spot offers made to each individual in the sample are constructed by assigning to each individual, in each period, a plan within the set of plans to which an individual with her same characteristics $D$ switched during a window of 12 months. In practice, I assign a spot offer to each individual from each insurance company by finding an exact match on gender, age, region, and health expenditure quintiles, and a nearest-neighbor match on wage, where the neighbor of individual $i$ is found among those individuals with weakly higher wages, to be consistent with the "7% rule".

With the sample of spot offers I can evaluate empirically the presence of risk-rating *via* spot prices. I do so with the OLS estimates of log of premium on health expenditure quintiles, after controlling for demographics and for plan's characteristics as in the following specification,

$$log\left(P_{it}^{jk}\right) = \alpha_j + \beta_h h_{it} + \beta_D D_{it} + \beta_X X_{it}^{jk} + \epsilon_{it}^{jk} \tag{1.12}$$

where $h_{it}$ are health expenditures quintiles during year $t-1$, $X_{it}$ are plan characteristics besides premium, and $\mathbf{D_{it}}$ is the set of demographic characteristics.

Column 1 of Table 1.8 shows the results of a specification in which only plan characteristics are added as controls: insurer dummies, effective coverage, quality, and a dummy for unrestricted network. The fit of this model is only 20%, and the restricted network coefficient has the incorrect sign. When age category and gender interactions are included, the fit of the model increases substantially, and the unrestricted network coefficient has the expected sign. This reflects the risk-adjustment on age and gender though the $r$ function. Column (3)

[36] regions 1-4,11,12, and 15

includes health expenditure quintile dummies, and shows an increasing relationship between health expenditures in the previous period, suggesting some moderate level of risk-rating via spot prices: individuals in the highest quintile of health expenditures pay on average 9.0% more in premiums for plans with equivalent characteristics. I will incorporate this empirical level of risk-rating when simulating choice sets under the current scenario.

## Estimation Procedure

I estimate the model using the Geweke-Hajivassiliou-Keane (GHK) multivariate normal simulator (J. Geweke and Keane, 2001, Keane, 1993; Keane, 1994, and Hajivassiliou, McFadden, and Ruud, 1996). GHK is convenient over standard accept/reject simulators since it requires the simulation of choice probabilities only for the chosen sequence. The algorithm consists of drawing the composite errors for each of the alternatives in each period from a normal distribution that is consistent with the chosen sequence. GHK permits the incorporation of cross-sectional correlation (present because an individual in period $t$ might face two offers from the company she picked in $t-1$) and time-series correlation (because of the AR(1) structure of the error term). The details of this procedure are in section A.4, but a few important modifications to the standard procedure are worth mentioning here. First, the guaranteed-renewability of contracts makes the choice set in a given period dependent on past choices. In particular, the company chosen in period $t$ defines the company of the guaranteed-renewable contract in period $t+1$ and therefore the correlation of its random components with those of each spot contracts. In practice, this makes the structure of the variance-covariance matrix of the composite error term $\alpha_i + u_{it}$ to be individual-specific. Also, the standard GHK procedure assumes that the econometrician observes the entire choice sequence. In order to incorporate "cohort 0" into the analysis (those incumbents individuals in the first period), I adapt the algorithm to allow for a truncation in the observed past choices. In section A.4 I show that a simple extension to the standard procedure, writing the likelihood for the choices in 2010 and 2011 conditional on the observed choice in 2009, allows for the use of the information provided by the choices of this cohort.[37]. Finally, in order to incorporate random choice sets, I jointly simulate the choice set and the error terms. In each repetition of the simulation, I simulate a choice set sequence and then the set of random normal terms. I use standard maximization techniques with $R = 200$ repetitions.

[37]This solution is in the same spirit of Wooldridge, 2005's solution to the initial conditions problem in dynamic panels. However, in this case, the problem arises from the guaranteed-renewability environment that makes individual-specific covariance matrices in year $t$ to depend on the choices in year $t-1$.

# 1.7 Results

## Parameter Estimates

Table 1.9 lists the estimates for the main parameters of the model for three different specifications. The first specification, in Column (1), restricts $\rho^a = \rho^w = \rho^s = 1$, so that the choice set the individual confronts is always the full set of potential choices. Column (2) shows the results of a specification in which $\rho^a$, $\rho^w$, and $\rho^s$ are estimated. The third specification allows $\rho^a$ and $\rho^w$ to depend on potential savings from search, and $\rho^a$ to also depend on age.

The price coefficient is between -0.25 and -0.18 depending on the specification. This implies a premium elasticity that is lower than what has been found in previous work.[38] However, a likelihood-ratio test strongly rejects the first specification where the price elasticity is the smallest.

The estimated probability of searching within is $\hat{\rho}_0^w = 0.43$, and a probability of searching across (conditional on searching within) is estimated to be $\hat{\rho}_0^a = 0.80$. Lower potential savings in the spot market decrease the probability of searching, but the result is not economically meaningful: the estimated probability of searching is higher by 1 percentage point for an individual with no potential savings than for an individual with the average (negative) potential savings.

To interpret the rest of the demand coefficients I calculate the marginal effects using simulation. I simulate the market shares for each company across four groups defined by health status and geographic region: healthy individuals v.s. individuals with preexisting conditions, and Santiago v.s. other regions. The effect of health status on preferences, as estimated in the model, results from by comparing the predicted market shares in column (a) to those in column (b). This effect varies across firms. The effect is larger for company $D$ that is predicted to have a 12.3 market share among the risky in Santiago compared to 8.8 among the healthy in Santiago. The effect is the smallest in company $E$ where both market shares differ by less than 1 percentage point. Since the model does not allow for interaction in the $Z$ variables (health status and region), the pattern described above also holds for market shares in other regions. On the other hand, market shares are predicted to vary significantly across regions, particularly for company $D$ that has a market share around 30 percentage points larger in regions different than Santiago.

On the supply side, the average individual with a preexisting condition is offered a contract with probability $\rho^s = 0.82$, which implies that on average one in five individuals with preexisting conditions is denied coverage in the spot market.

[38] Jackson Abaluck and Gruber, 2011 find an elasticity close to -1. I estimated the model using the control function approach of Petrin and Train, 2009 , with a "marginal cost" instrument derived from the average covered expenditures for individuals in the plan. The price coefficient is not altered significantly in the IV models, and since I am not confident that the exclusion restriction is satisfied, I continue to estimate the model without an instrument

## Health status by age

Along with the structural parameters estimated above, a key input for quantifying lock-in is the share of individuals subject to underwriting in the spot market. In this section I explain my empirical approach to simulate the evolution health status over and individual's lifetime, to determine the type of offers they receive in the spot market.

As stated in Section 1.10, offer policies map an individual's health status (as well as other demographics) to a potentially offered premium and plan characteristics, as well as a coverage decision. Specifically, I have modeled throughout that premiums and characteristics depend on 5 health expenditure quintiles, and the coverage decision depends on the presence of a preexisting condition. Therefore, the supply response is determined by 10 different and mutually exclusive states for status, corresponding to the combination of 5 different health expenditure quintiles and a preexisting condition indicator.

I estimate the probability of being in each of the 10 states during an individual's lifetime by assuming that the health process is a Markov Chain with transition probabilities that depend on age and gender. For each age and gender, I use the actual transition rates across the 10 states as estimates of the transition probabilities. I calculate a separate transition matrix for each age and gender. Then, I use each of these age-gender specific 10 by 10 transition matrices for simulating health paths from age 25 until retirement (60 for females and 65 for males).[39].

Tables 1.11 and 1.12, present, as an example, the transition matrices at age 25 for females and males respectively, and tables 1.13 and 1.14, the corresponding transition matrices at age 55. States 1 to 5 correspond to 5 quintiles of health expenditures with no preexisting conditions, with states 6 to 10 corresponding to 5 quintiles of health expenditures but with preexisting conditions. On-diagonal entries reflect persistence in health status. For instance, the first element in Table 1.11 shows that 41 % of 25 year-old females that are in the healthiest group are expected to remain in that category at age 26. States 6-10 (with preexisting conditions) represent only a small share of cases at age 25, and the vast majority transition to states 1-5 (without preexisting conditions) in the next period. On the other hand, persistence at the sickest states is high at age 55 : 40 % of females and 48 of males in the sickest state at age 55 are expected to remain in state 10 at age 56.

The most important outcome from these tables is the predicted share of individuals with preexisting conditions at each age. To assess the accuracy of this procedure in forecasting the prevalence of such conditions over time, Figure 1.9 compares the empirical share of males and females with preexisting conditions by age (full line), and the simulated share of males and females with preexisting conditions (dashed line). Overall, this procedure achieves a good fit. Males start with a prevalence of preexisting condition of around 6% at age 25. The prevalence among men rises to around 50 % by the age of 65. The prevalence among females at 25 is slightly higher than for males, starting at around 13 %, but it increases less steeply than for males. At age 60, I estimate that around 38 % of females have a preexisting

[39]Handel, Hendel, and Whinston, 2015b take a similar approach

condition.[40] Note that total health expenditures enter in the demand model, in particular in individual's valuation for each company. Total health expenditures are also simulated non-parametrically, by drawing a random number from the empirical distribution of health expenditures conditional on each state.

## Lock-in

With simulated health expenditures, the estimated level of underwriting, and the estimated preferences, I can now quantity the share of individuals locked-in because underwriting in the spot market. I use the estimated preferences and underwriting rules, to simulate the choices of individuals over their lifetime. Each individual's health status, which impacts their preference and potential choices, are simulated with the methodology described in the previous section.

Formally, let $K_i\left(\Theta_{it}, \mathbb{M}, H_{it}\right) = \{k_1, k_2, ..., k_T\}$ be the sequence of companies chosen by individual $i$ under the current offer policy $\mathbb{M}$, preference parameters $\Theta_{it}$, and health-status process $H_{it}$. Let $\mathbb{M}'$ be a different policy and $K_i'\left(\Theta_{it}, \mathbb{M}', H_{it}\right) = \{k_1', k_2', ..., k_T'\}$. I define an individual's *willingness to pay for policy* $\mathbb{M}$*' over policy* $\mathbb{M}$ *in period* $t$ as $w_{it}$:

$$w_{it}\left(\mathbb{M}', \mathbb{M}, \Theta_{it}, H_{it}\right) \equiv max\left(\frac{u_{it}^{k_t'} + \beta_P P^k - \left(u_{it}^{k_t} + \beta_P P^{k'}\right)}{\beta_P}, 0\right) \tag{1.13}$$

The willingness to pay $w_{it}$ can be interpreted as the dollar amount that makes agent $i$ indifferent between her current policy $\mathbb{M}$ and paying $w_{it}$ for receiving offers from policy $\mathbb{M}'$. I quantify this object by simulating $K_i$ and $K_i'$ for a representative sample of individuals that enter in the market at 25 years old and stay for 35 years.

To quantify lock-in, I calculate the willingness to pay for an offer policy $\mathbb{M}'$ that eliminates risk-rating in the spot market, so that a) premium risk-rating is eliminated and b) coverage denial is eliminated. In this exercise I assume away any potential changes to the overall level of premiums associated with the new policy $\mathbb{M}'$. In that sense, calculating $w_{it}$ does not answer a full welfare analysis question, but it is instructive to calculate how much a single individual would be willing to pay to eliminate her underwriting while keeping everyone else's. I return to the question of full welfare analysis in general equilibrium in the next section, when I allow prices to adjust to the new policy.

The simulation procedure to recover $K_i$ for a given offer policy $\mathbb{M}$ is as follows:

1. Set t = 0 and set $D_{i0} = (age_0, gender_0, region_0)$ equal to the empirical $D_i$ for individuals entering in the market at age 25 in 2009.

2. Draw $H_{it} = (h_{it}, Preex_{it})$ and $Z_{it}$ from the empirical (joint) distribution $F_H(D_{it})$ and each $z_i \in Z_i$ from the empirical distribution $F_z(D_i)$

[40] In the interest of space I do not show this comparison for all 10 states, but they all show a good fit. These figures are available upon request

3. Construct the choice menu by drawing offers from each company, $M_{ik} \in \mathbb{M}^{\mathbb{k}}(D_{it}, h_{it}, Preex_{it})$, by

   a) Drawing $X^k(D_{it}, h_{it}, Preex_{it})$ from the empirical distribution of spot offers in each company and

   b) calculating spot prices using the estimates of spot prices from equation 1.12.

4. If $t > 0$, add $k_{i,t-1}$ to the choice menu.

5. Draw $u_{it}^{jk}$ and choice $k_{it}$ for each $i$ among her choice menu.

6. Update $D_{it}$ and return to step 2.

The left Y-axis of Figure 1.10 shows the share of individuals with positive willingness to pay for the policy described above over the current policy. I also include the average willingness to pay in the right Y-axis.

Individuals with $w_{it} > 0$ are those locked-in to their plans: they are enrolled in plan $k$ under policy $\mathbb{M}$ but would be enrolled in plan $k' \neq k$ under an alternative policy $\mathbb{M}'$ that bans underwriting. The share of locked-in individual reaches around 5 percent after 35 years and is increasing over time, as preexisting conditions become more prevalent. On average, an individual would be willing to pay around 13 % of the current average premium for policy $\mathbb{M}'$.

I use the estimated parameters of the model to shed light on the relative importance of the two potential sources of underwriting in the market: coverage denial and premium risk-rating. To calculate the level of lock-in produced only by risk-rating of premiums, I simulate the share of individuals with $w_i^t > 0$ and willingness to pay for a policy $\mathbb{M}''$ where everyone gets risk-rated offers. In practice, I leave the current level of risk rating but use $\rho^s = 1$, to shut down the coverage-denial mechanism. The purpose of this exercise is only to describe the relative importance of both sources of lock-in rather than answering a general-equilibrium question, so I leave the level of risk-rating at the original parameter estimates. The result of this simulation is shown in the blue dashed line in Figure 1.12. I find that most of the lock is due to preexisting conditions: the simulations predict that less than one percent of individuals would be locked-in if I set $\rho^s = 1$ while keeping risk-rating in spot premiums.

Thus, the level of lock-in results is mostly sensitive to the estimated coverage denial rates rather than the level of premium risk-rating. My estimates indicate that 1 in 5 individuals with preexisting conditions are denied coverage in the spot market. Since around 50 % of males and 40 % of females are expected to end-up with a preexisting condition by age 60, mechanically, the share of individuals facing coverage denial in the spot market is 10 % of males and around 8% for females.

To show how higher coverage-denial rates translate into higher lock-in, I simulate the economy assuming that everyone with a preexisting condition is denied coverage in the spot market, that is by setting now $\rho^s = 1$. The results are in the black dashed line in Figure

1.12. Under this alternative assumption, the share of locked-in individuals would be around 16 % at the age of 60.

## Repricing effects

The share of locked-in individuals calculated in the previous section do not necessarily correspond to the share of switchers if policy $\mathbb{M}'$ is implemented. Mechanically, 5 % of individuals would switch if prices of plans remain fixed at their original level. However, the prices of contracts are expected to change in response to those switchers, creating a general-equilibrium effect in the allocation. Prices of contracts to which individuals with preexisting conditions switch are expected to increase, decreasing the share of those who would effectively want to switch, and also potentially generating the result that some healthy individuals would want to switch out of those contracts.

Predicting price responses to policy $\mathbb{M}'$ would require that I specify and estimate a full supply model, which is outside of the scope of this paper. Instead, I make use of simple supply-side assumptions that allow me to use the already estimated parameters to quantify these effects. I find the equilibrium under $\mathbb{M}'$ by assuming that the average markup per enrollee of each company does not change after the policy. This simulates, for instance, a scenario in which all extra payments made by enrollees in the counterfactual scenario go to a common pool that is distributed to each company accordingly (so there is "risk-adjustment" relative to the original policy). Also, I assume that the change in markup at the company level is compensated with a uniform price increase of all plans at the company.

Formally, let $\mathbf{A}_t(\mathbb{M}, \Theta)$, the $I \times K$ allocation matrix whose element $A_t\left(i, k, \mathbb{M}, \Theta\right)$ is equal to 1 if individual $i$ is enrolled in company $k$ in period $t$ and 0 otherwise, given policy $\mathbb{M}$ and demand parameters $\Theta$. Let $J\left(k\right)$ be the collection of plans of firm $k$. The average markup in period $t$ of company $k$ under allocation $\mathbf{A}_t$ is given by

$$\mu_k^t\left(\mathbf{A}_t|\Theta, \mathbb{M}\right) = \frac{\sum_{i=1}^{N}\sum_{j\in J(k)} A_t\left(i, k|\mathbb{M}, \Theta\right) \times \left(P_{it}^{jk}\left(\mathbb{M}\right) - c^k(h_{it}, X_{it}^{jk})\right)}{\sum_{i=1}^{N}\sum_{j\in J(k)} A_t\left(i, k\right)}$$

I define an allocation $\mathbf{A'}_t$ as an *equilibrium allocation under a counterfactual offer policy* $\mathbb{M}'$ if

1. Markups are equal to current markups

$$\mu_k^t\left(\mathbf{A'}_t|\Theta, \mathbb{M}'\right) = \mu_k^t\left(\mathbf{A}_t|\Theta, \mathbb{M}\right)$$

2. Individuals choose company/plan optimally given $\Theta$ and $\mathbb{M}'$, so that for all plans $\tilde{j}, \tilde{k} \in \mathbb{M}'$

$$A_t(i, j, k|\Theta, \mathbb{M}') = \mathbb{1}(U_{it}^{jk} > U_{it}^{\tilde{j}\tilde{k}})$$

The difference between the share of locked-in individuals and the share of switchers under the equilibrium allocation quantifies the general equilibrium effect of the policy. I calculate the equilibrium allocation using the following algorithm:

1. Set $t = 0$
2. Set $r = 0$, $P_{it}^{jk,(r)} = P_{it}^{jk}$ and $\mathbf{A}_t^{(r)} = \mathbf{A}_t$, i.e., start with prices and allocations under current offer policy $\mathbb{M}$.
3. Simulate $\mathbf{A}_t^{(r+1)}$ given offer policy $\mathbb{M}'$, and prices $P_{ik}^{(r)}$.
4. Construct $\delta^{(r)} = |\mathbf{A}_t^{(r+1)} - \mathbf{A}_t^{(r)}|$ where $||$ is a norm. If $\delta^{(r)} < \epsilon$: stop. Else,
   a) calculate

$$\Delta\mu_t^{k,(r)} \equiv \frac{\sum_{i=1}^{N}\sum_{j\in J(k)} A_t^{(r+1)} \times \left(P_{it}^{jk,(r)} - h_{it}\right)}{\sum_{i=1}^{N}\sum_{j\in J(k)} A_t} - \frac{\sum_{i=1}^{N}\sum_{j\in J(k)} A_t^{(r)} \times \left(P_{it}^{jk,(r)} - h_{it}\right)}{\sum_{i=1}^{N}\sum_{j\in J(k)} A_t}$$

   b) update prices $P_{ik}^{(r+1)} = P_{ik}^{(r)} + \Delta\mu_t^{k,(r)}$
   c) go back to step (2) with $r + 1 \to r$

Figure 1.12 compares the share of locked-in individuals under $\mathbb{M}$ to the share of individuals that would switch under policy $\mathbb{M}'$, after prices adjust. The results show that both are almost indistinguishable.

## Preference heterogeneity, lock-in, and adverse selection

A recent literature has focused on the interaction between preference heterogeneity and regulation in static health insurance contracts (Einav and Finkelstein, 2011, Bundorf, Levin, and Mahoney, 2012 and Geruso, 2013). In guaranteed-renewable contracts, preference heterogeneity plays two opposing roles in determining the level of lock-in. As discussed in Section 1.2, evolving preference heterogeneity is the main source of lock-in. On the other hand, when preferences are stable, individuals are less prone to lock-in in the guaranteed-renewable environment, since it reduces the share of individuals for whom reclassification in the spot market is relevant.

The general equilibrium effects of transitioning to community rating also depend on preference heterogeneity.[41] In the situation analyzed in this paper, preference heterogeneity

[41] As an example, Einav and Finkelstein, 2011 show that if risk-aversion is negatively correlated with health risk, a uniform pricing may induce advantageous selection. Contrary to the standard model of health insurance markets, health insurance is more valuable for healthier individuals.

over companies may arise from several reasons uncorrelated to health expenditures. I discuss this issue in more detail using static framework of Einav and Finkelstein, 2011, where I analyze the impact of stable preference heterogeneity in both the mechanical and general equilibrium effect of banning underwriting.

As in Bundorf, Levin, and Mahoney, 2012, an individual's relative valuation for insurance is given by $u(h, \epsilon)$ where $h \in [0, \infty)$ is health risk and $\epsilon \in (-\infty, \infty)$ summarizes other determinants of valuation that are orthogonal to $h$, with $E(\epsilon) = 0$. The presence of $\epsilon$ in the utility function is intended to capture preference heterogeneity. The degree of preference heterogeneity is captured by $\partial u/\partial \epsilon$, which is assumed to be weakly positive. I assume that there is adverse-selection, so that $\partial u/\partial h > 0$. Here $h$ is private information, in the sense that firms cannot price based on $h$.

Preference heterogeneity within health status decreases the average cost at any price, as show in panel (a) of Figure 1.13. Intuitively, starting with a situation in which everyone is enrolled in a plan, the marginal enrollees that would drop out of the contracts after a price increase are unambiguously the healthiest if preferences are perfectly correlated with health status (so that the healthiest are those that have the lowest valuation for the contract). On the contrary, with preference heterogeneity, some marginal enrollees are high-risk individuals with low valuation for the plan because of reasons uncorrelated to health (I provide a simple formal proof in section A.4).

The competitive equilibrium is found at the intersection of the demand curve and the respective average cost curve (AC). For a given demand curve, higher preference heterogeneity will therefore imply lower premiums and a higher number of enrollees in equilibrium. [42] As shown in the shaded area in panel (b), the standard marginal cost curve (depicted by $MC_1$) is replaced by a marginal cost *correspondence* ($MC_2$), to reflect the heterogeneous costs of the marginal enrollees[43]

Assume that preexisting conditions take the form of denying coverage to anyone with $MC > c^*$. The share of individuals who are locked-in corresponds to the number of individuals with preexisting conditions who are not allowed to enroll in the plan if preexisting conditions are introduced, but that would otherwise enroll. The mechanical effect is represented by a leftward shift of the demand curve of a magnitude that is equal to the number of

[42] Even if the demand and the cost curve are tightly linked in insurance markets, the degree of this linkage depends on the degree of preference heterogeneity. Thus, two different utility functions, with different degrees of preference heterogeneity, can yield the same demand curve and different average cost curves. For instance $u_1 = h + \epsilon$ and $u_2 = 2h$ with $h \sim U[0, 1]$ and $\epsilon \sim U[0, 1]$ produce the same demand curve but different AC curves.

[43] This figure provides an explanation complementary to Bundorf, Levin, and Mahoney, 2012's to why preference heterogeneity makes a uniform price policy inefficient at any price. Any price above $P_h$ does not produce the efficient outcome because a lower price would generate marginal enrollees who have a marginal cost below their willingness to pay. On the other hand, any price below $P_l$ is also inefficient because increasing the price will make individuals who have a marginal cost above their willingness to pay opt out of the contract. Moreover, all prices between $P_l$ and $P_h$ do not yield the efficient outcome either, since under these prices there are some individuals who are inefficiently enrolled in the plan and some inefficiently not enrolled.

individuals with preexisting conditions who would have bought the plan at any price. Panel (b) of Figure 1.13 shows that preference heterogeneity decreases the share of such individuals, and therefore produces a smaller leftward shift in demand when preexisting conditions are introduced. Panel (c) shows the leftward shift in demand, represented by the new demand curve $D'$, in the case of no preference heterogeneity. Panel (d), shows a smaller shift, corresponding to the case with heterogeneity. The mechanical effect is represented by the decrease in quantity from the original equilibrium $Q^*$ to the new quantity $Q_{mec}$, corresponding to the new curve and the original average cost curve (AC). The "GE" effect is represented by a movement along the new demand curve, toward its intersection with the new AC curve (AC').

In my empirical model, stable preference heterogeneity is captured by the terms $\alpha_{i0}$, $\sigma_k$, and the autocorrelation terms $\kappa_1$ and $\kappa_2$. When these terms are higher (in an absolute value sense for $\alpha_{i0}$), individuals are predicted to switch less during their lifetime, even in the absence of underwriting.

To shed light on the importance of stable preference heterogeneity in reducing lock-in, I simulate the economy assuming that $\alpha_{i0} = 0$ and $\kappa_1 = 0$ and $\kappa_2 = 0$ instead of the estimated parameters. The results are in the red dashed line of Figure 1.12, which shows the share of locked-in individuals under these new assumptions, for the case of $\rho^s = 1$. The share of locked-in individual increases from 16 percent to 19 percent at age 60.

## 1.8 Conclusions

This paper contributes to the literature with an empirical evaluation long-term health insurance contracts. Specifically, I evaluate the workings of guaranteed-renewable contracts in the Chilean private health insurance market, where individuals potentially receive offers from different health insurance companies.

Theoretically, guaranteed-renewable contracts have the potential to fully eliminate adverse selection and reclassification risk as long as individuals do not have incentives to switch across these companies for non-financial reasons. However, in reality, contracts have non-financial characteristics –like the provider network –which vary across companies. Individuals switch every year –7 % on average in the Chilean market –but switching rates are significantly higher among the healthier. Sick individuals that come to dislike their insurance company but cannot switch because of the financial incentives imbedded in these contracts suffer a welfare loss.

I estimate that the welfare loss resulting from lock-in in Chile reaches around 13% of the yearly premium by the time individuals reach age 60. Around 5% of individuals are locked-in to their insurer at this point, although 60% of individuals experience a lock-in event in their lifetime. The estimated incidence of lock-in depends crucially on the rate at which individuals with preexisting conditions are denied coverage in the spot market, which I estimate to be around 20 %. Small levels of lock-in also imply minor general-equilibrium effects upon transitioning to a community rating scheme.

Even if the presence of lock-in in this particular market is relatively small, the degree of lock-in in long-term arrangements is an empirical question that depends on context-specific levels of differentiation across insurers as well as the evolution of preferences over time. This paper provides a systematic way of empirically evaluating this issue in other markets.

This paper does not deal explicitly with a few important aspects of guaranteed-renewability. First, despite incorporating behavioral biases in estimating demand, I do not study in detail the consequences of consumer mistakes in the evaluation of guaranteed-renewability. However, the problem of lock-in adds an important layer to the design of health insurance markets with behavioral agents. Arguably, the lack of portability of contracts is more problematic when individuals cannot forecast their future preferences or needs.[44] Relatedly, individuals suffer more from lock-in if it is difficult to make a good initial choice before a learning period. Although there is a great deal of consensus that individuals have difficulties in choosing plans, there is limited empirical evidence on whether they learn over time.[45]

Finally, when evaluating the desirability of long-term contracts that generate lock-in, it is important to incorporate other margins of response in the supply that might be not contractible at the start.[46] In health insurance markets, insurance companies generally revise the terms of their agreements with providers (Shepard, 2015). In fact, in April 2015, two of the companies analyzed in this paper went through negotiations with a group of providers that resulted in major changes in their networks. Changes in provider networks have important consequences for individuals who face underwriting. It is an open question to study the dynamic relationship between insurance companies and providers when the demand is subject to long-term contracts with lock-in.

[44] As Pauly and Herring, 1999 warn, the lock-in should "induce more care in the initial choice [...]".

[45] Ketcham et al., 2012 and Jason Abaluck and Gruber, 2013 evaluate learning in Medicare Part D

[46] See Farrell and Shapiro, 1989 for a theoretical analysis of long-term contracts with switching costs and unenforceable quality.

Table 1.1: Share of claims by provider and type in each company

| Firm | Provider | Share Claims cancer | other | Ratio | Firm | Provider | Share Claims cancer | other | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| | | (c) | (o) | (c/o) | | | (c) | (o) | (c/o) |
| A | P1 | 0.42 | 0.07 | 6.0 | B | P4 | 0.33 | 0.01 | 40.0 |
| A | P2 | 0.18 | 0.01 | 17.4 | B | P1 | 0.25 | 0.11 | 2.2 |
| A | P3 | 0.10 | 0.00 | 28.3 | B | P6 | 0.19 | 0.00 | 48.9 |
| A | P4 | 0.09 | 0.00 | 43.9 | B | P5 | 0.04 | 0.15 | 0.3 |
| A | P5 | 0.06 | 0.16 | 0.4 | | | | | |
| **A** | **Cum.** | **0.85** | **0.24** | | B | **Cum.** | **0.81** | **0.27** | |
| C | P7 | 0.29 | 0.04 | 6.9 | D | P4 | 0.54 | 0.01 | 43.6 |
| C | P5 | 0.18 | 0.21 | 0.8 | D | P12 | 0.21 | 0.07 | 2.8 |
| C | P8 | 0.15 | 0.11 | 1.4 | D | P13 | 0.04 | 0.00 | 1541.7 |
| C | P9 | 0.09 | 0.00 | 23.7 | D | P2 | 0.03 | 0.01 | 3.7 |
| C | P10 | 0.07 | 0.04 | 2 .0 | | | | | |
| C | P11 | 0.06 | 0.17 | 0.3 | | | | | |
| **C** | **Cum.** | **0.84** | **0.58** | | D | **Cum.** | **0.82** | **0.10** | |
| E | P7 | 0.55 | 0.14 | 3.8 | F | P4 | 0.56 | 0.01 | 91.1 |
| E | P9 | 0.09 | 0.01 | 15.2 | F | P7 | 0.05 | 0.03 | 1.7 |
| E | P10 | 0.09 | 0.05 | 1.7 | F | P14 | 0.05 | 0.06 | 0.9 |
| E | P8 | 0.05 | 0.16 | 0.3 | F | P6 | 0.05 | 0.00 | 37.3 |
| E | P5 | 0.04 | 0.08 | 0.5 | F | P13 | 0.04 | 0.00 | 288.1 |
| | | | | | F | P15 | 0.03 | 0.01 | 4.7 |
| | | | | | F | P16 | 0.03 | 0.01 | 2.1 |
| **E** | **Cum.** | **0.82** | **0.44** | | F | **Cum.** | **0.81** | **0.12** | |

*Note:* This table shows, for each company, the share of claims related to cancer and to other all other (non-chronic) health conditions, for all claims in 2011. For instance, 42% of cancer-related claims of individuals enrolled in company A where treated by provider $P_1$. That provider treated 7% of the "other" claims for enrollees in the same company.

Table 1.2: Net flow depending on health status

| Firm | $(h_{it}^{preex}=0$ | $(h_{it}^{preex}=1)$ | Difference |
|---|---|---|---|
| A | $-2.6\%$ | $-1.6\%$ | $1.0\%$ |
| B | $-8.5\%$ | $6.1\%$ | $-14.7\%^{***}$ |
| C | $4.8\%$ | $0.9\%$ | $3.8\%^{***}$ |
| D | $-2.2\%$ | $-5.1\%$ | $2.9\%^{***}$ |
| E | $8.6\%$ | $-0.4\%$ | $9.0\%^{***}$ |
| N obs. | 13482 | 3461 | |

*Note:* Table shows the net flow ($entry - exit$) to each company among switchers, as a share of total switchers, for individuals with preexisting conditions and without preexisting conditions, for a sample of enrollees in January 2009 and followed until December 2012
***: Difference is significant at the 95% confidence level.

Table 1.3: Front-loading evidence

| | (a) | | (b) | |
|---|---|---|---|---|
| | Parameter Estimate | Marginal Effect | Parameter Estimate | Marginal Effect |
| $T_{it}$ | $0.078^{***}$ | $0.05^{***}$ | | |
| | (0.007) | (0.005) | | |
| $T_{it} \times (age \leq 35)$ | | | $0.077^{***}$ | 0.05 |
| | | | (0.009) | (0.006) |
| $T_{it} \times (35 < age \leq 45)$ | | | $0.066^{***}$ | 0.043 |
| | | | (0.013) | (0.008) |
| $T_{it} \times (45 < age \leq 60)$ | | | $0.091^{***}$ | 0.058 |
| | | | (0.016) | (0.01) |
| N obs | 1,185,346 | | 1,185,346 | |
| N groups | 45,212 | | 45,212 | |

*Note:* This graph shows GLM estimates of equation (1.4) to show the increasing relationship between tenure in a plan $T_{it}$ and the ratio between total claims and premium, $r_{it} = h_{it}/P_{it}$, on a 4-year monthly panel of enrollees by January 2009. Panel (a) pools all age groups. Panel (b) shows the results by interacting tenure with 3 age groups.
Standard errors in parentheses, clustered at the individual level.
* p<0.10, ** p<0.05, *** p<0.01

Table 1.4: Prevalence of preexisting conditions

| | Prevalence self reported in SPS [%] | Patients with related procedure in ISAPRES claims dataset [%] |
|---|---|---|
| Diabetes | 4 | 8 |
| Depression and Chronic Psyc. Disorder | 5 | 7 |
| Arthritis | 3 | 3 |
| Hypertension and cardiovascular diseases | 10 | 7 |
| Cancer | 1 | 1 |
| Chronic Renal Insufficiency | 1 | 0.1 |

*Note:* This table shows the prevalence of the 6 major preexisting conditions. It compares the prevalence found in the ISAPRES dataset of this paper, using the procedure claims associated with each condition to the self-reported prevalence in the Social Protection Survey of 2009.

Table 1.5: Cox proportional hazard model estimates

| | (1) | (2) | (3) |
|---|---|---|---|
| $\mathbb{1}(h_{it}^{preex} > 0)$ | 0.743*** | 0.776*** | 0.850*** |
| | (0.039) | (0.041) | (0.047) |
| $age$ | | 1.017 | 1.025 |
| | | (0.024) | (0.025) |
| $age^2$ | | 1.000 | 1.000 |
| | | (0.000) | (0.000) |
| $gender$ | | 1.277*** | 1.263*** |
| | | (0.065) | (0.065) |
| $\mathbb{1}(h_{it}^{preex} > 0) \times log(h_{it}^{preex})$ | | | 1.057* |
| | | | (0.034) |
| $\mathbb{1}(h_{it} > 0)$ | | | 0.948 |
| | | | (0.048) |
| $\mathbb{1}(h_{it} > 0) \times log(h_{it})$ | | | 0.872*** |
| | | | (0.018) |
| $N$ | 165409 | 165409 | 155668 |

Exponentiated coefficients; Standard errors in parentheses
* p¡0.10, ** p¡0.05, *** p¡0.01

*Note:* This table shows the estimates of a proportional cox hazard model for the event of switching company, as a function of preexisting conditions and other demographics. Exponentiated coefficients.
Standard errors in parentheses, clustered at the individual level. * p<0.10, ** p<0.05, *** p<0.01

Table 1.6: Descriptive statistics of estimation sample

| cohort = | all | 0 | 0 | 2009 | 2009 | 2009 | 2010 | 2010 | 2011 |
|---|---|---|---|---|---|---|---|---|---|
| year = | all | 2010 | 2011 | 2009 | 2010 | 2011 | 2010 | 2011 | 2011 |
| gender | 0.62 | 0.62 | 0.65 | 0.61 | 0.62 | 0.64 | 0.61 | 0.63 | 0.61 |
| age | 34.05 | 37.50 | 38.11 | 33.10 | 33.89 | 34.66 | 33.42 | 33.87 | 33.03 |
| Nr. Dep. | 0.44 | 0.85 | 0.84 | 0.33 | 0.37 | 0.41 | 0.43 | 0.42 | 0.39 |
| Insurer = A | 0.18 | 0.19 | 0.19 | 0.17 | 0.18 | 0.18 | 0.18 | 0.19 | 0.19 |
| Insurer = B | 0.26 | 0.22 | 0.21 | 0.26 | 0.25 | 0.24 | 0.27 | 0.26 | 0.29 |
| Insurer = C | 0.05 | 0.06 | 0.06 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.04 |
| Insurer = D | 0.22 | 0.25 | 0.26 | 0.21 | 0.22 | 0.24 | 0.21 | 0.23 | 0.21 |
| Insurer = E | 0.29 | 0.28 | 0.28 | 0.31 | 0.30 | 0.29 | 0.30 | 0.29 | 0.27 |
| Santiago | 0.57 | 0.62 | 0.65 | 0.56 | 0.58 | 0.60 | 0.53 | 0.55 | 0.55 |
| Has Preex Cond. | 0.15 | 0.20 | 0.21 | 0.13 | 0.15 | 0.17 | 0.13 | 0.15 | 0.13 |
| Pick GR plan | 0.39 | 0.77 | 0.82 | 0.00 | 0.79 | 0.80 | 0.00 | 0.82 | 0.00 |
| Switch Within | 0.06 | 0.14 | 0.11 | 0.00 | 0.13 | 0.11 | 0.00 | 0.13 | 0.00 |
| N | 313,462 | 20,333 | 15,577 | 56,613 | 45,301 | 34,616 | 51,580 | 36,292 | 53,150 |

*Notes:* Sample means for key variables in the estimation dataset

Table 1.7: Descriptive statistics of plans

| cohort = | 0 | 0 | 2009 | 2009 | 2009 | 2010 | 2010 | 2011 |
|---|---|---|---|---|---|---|---|---|
| year = | 2010 | 2011 | 2009 | 2010 | 2011 | 2010 | 2011 | 2011 |
| Premium (Ths.USD) | 1.33 | 1.42 | 1.17 | 1.29 | 1.40 | 1.20 | 1.32 | 1.29 |
| Coverage in-network | 0.83 | 0.82 | 0.85 | 0.84 | 0.83 | 0.84 | 0.83 | 0.83 |
| Coverage out-network | 0.62 | 0.61 | 0.65 | 0.65 | 0.64 | 0.64 | 0.63 | 0.63 |
| Unrestricted Network | 0.37 | 0.36 | 0.41 | 0.41 | 0.39 | 0.37 | 0.39 | 0.34 |
| Distance | 0.84 | 0.84 | 0.00 | 0.82 | 0.82 | 0.00 | 0.82 | 0.00 |
| N | 121,998 | 93,462 | 283,065 | 271,806 | 207,696 | 257,900 | 217,752 | 265,750 |

*Note*: Sample means for key variables in the estimation dataset

Table 1.8: Spot prices as a function of plan's and individuals' characteristics

| | (1) | (2) | (3) |
|---|---|---|---|
| $c_p$ | 0.402*** | 2.047*** | 2.037*** |
| | (0.082) | (0.074) | (0.074) |
| Free Network | -0.052* | 0.178*** | 0.178*** |
| | (0.028) | (0.023) | (0.023) |
| $log(wage_t)$ | | 0.227*** | 0.222*** |
| | | (0.009) | (0.009) |
| $h_{it-1} = 2$ | | | 0.007*** |
| | | | (0.003) |
| $h_{it-1} = 3$ | | | 0.030*** |
| | | | (0.003) |
| $h_{it-1} = 4$ | | | 0.059*** |
| | | | (0.004) |
| $h_{it-1} = 5$ | | | 0.088*** |
| | | | (0.005) |
| Year fixed effect | Yes | Yes | Yes |
| ISAPRE fixed effect | Yes | Yes | Yes |
| age x gender fixed effects | No | Yes | Yes |
| $N$ | 320190 | 320190 | 320190 |
| $R^2$ | 0.216 | 0.735 | 0.738 |

*Notes* : OLS estimates of equation 1.12, that quantifies the correlation between log of price and plan and individual characteristics. The sample of plans correspond to "spot" plans.

Table 1.9: Parameter estimates

| | Spec. 1 | | Spec. 2 | | Spec. 3 | |
|---|---|---|---|---|---|---|
| | point estimate | s.e. | point estimate | s.e. | point estimate | s.e. |
| $\beta_P$ | -0.18 | (0.016) | -0.26 | (0.027) | -0.22 | (0.026) |
| $\beta_c$ | 0.57 | (0.075) | 0.82 | (0.119) | 0.79 | (0.116) |
| $\mu_A$ | -0.37 | (0.078) | -0.23 | (0.069) | -0.22 | (0.066) |
| $\mu_B$ | 0.05 | (0.044) | 0.05 | (0.064) | 0.05 | (0.056) |
| $\mu_C$ | -1.00 | (0.094) | -1.15 | (0.258) | -1.16 | (0.513) |
| $\mu_D$ | 0.47 | (0.039) | 0.49 | (0.041) | 0.49 | (0.041) |
| $log(\sigma_A)$ | -0.28 | (0.146) | -0.77 | (0.283) | -0.79 | (0.272) |
| $log(\sigma_B)$ | -2.27 | (0.774) | -2.25 | (2.564) | -2.25 | (2.643) |
| $log(\sigma_C)$ | -0.99 | (0.62) | -0.75 | (0.783) | -0.71 | (1.54) |
| $\rho_0^w$ | | | 0.43 | (0.025) | 0.42 | (0.024) |
| $\rho_0^a$ | | | 0.76 | (0.077) | 0.80 | (0.074) |
| $\rho_s$ | | | 0.83 | (0.083) | 0.82 | (0.08) |
| $\kappa_1$ | 0.73 | (0.015) | 0.56 | (0.067) | 0.58 | (0.042) |
| $\kappa_2$ | 1.00 | (0.001) | 0.88 | (0.037) | 0.86 | (0.031) |
| Axstgo | -0.12 | (0.062) | -0.14 | (0.06) | -0.15 | (0.06) |
| Axhealth | -0.13 | (0.044) | -0.14 | (0.067) | -0.15 | (0.067) |
| Bxstgo | -0.19 | (0.052) | -0.19 | (0.057) | -0.20 | (0.056) |
| Bxhealth | -0.11 | (0.037) | -0.08 | (0.058) | -0.08 | (0.058) |
| Cxstgo | -0.06 | (0.076) | 0.08 | (0.099) | 0.07 | (0.058) |
| Cxhealth | -0.12 | (0.063) | -0.22 | (0.109) | -0.21 | (0.085) |
| Dxstgo | -1.26 | (0.053) | -1.38 | (0.058) | -1.38 | (0.057) |
| Dxhealth | 0.10 | (0.038) | 0.30 | (0.055) | 0.30 | (0.054) |
| $\rho_{age}^a$ | | | -0.16 | (0.355) | -0.18 | (0.388) |
| $\rho_{savings}$ | | | | | 0.22 | (0.047) |
| Log L | | -15637.70 | | -14554.40 | | -14490.90 |
| N | | 84507.00 | | 84507.00 | | 84507.00 |

*Notes:* Table shows the parameter estimates of the structural model for three different specifications

Table 1.10: Predicted market shares as a function of time-varying observables

| | | Healthy | Risky |
|---|---|---|---|
| Santiago | A | 21.5 | 19.5 |
| | B | 23.5 | 22.5 |
| | C | 7.9 | 6.5 |
| | D | 8.8 | 12.3 |
| | E | 38.3 | 39.2 |
| Other Regions | A | 15.9 | 13.4 |
| | B | 18.8 | 16.4 |
| | C | 3.9 | 3.0 |
| | D | 37.9 | 44.7 |
| | E | 23.5 | 22.4 |

*Notes :*This table shows the predicted market shares for healthy v/s risky and Santiago v/s Other regions based on the structural estimates

Table 1.11: Health status transition from one year to the next, females at age 25

| $s_t/s_{t+1}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.41 | 0.29 | 0.13 | 0.05 | 0.04 | 0.00 | 0.02 | 0.03 | 0.01 | 0.01 |
| 2 | 0.18 | 0.33 | 0.21 | 0.09 | 0.09 | 0.00 | 0.02 | 0.04 | 0.02 | 0.02 |
| 3 | 0.08 | 0.24 | 0.24 | 0.15 | 0.16 | 0.00 | 0.02 | 0.04 | 0.04 | 0.04 |
| 4 | 0.05 | 0.15 | 0.19 | 0.19 | 0.23 | 0.00 | 0.01 | 0.03 | 0.05 | 0.08 |
| 5 | 0.06 | 0.17 | 0.22 | 0.22 | 0.19 | 0.00 | 0.01 | 0.03 | 0.05 | 0.06 |
| 6 | 0.23 | 0.32 | 0.10 | 0.06 | 0.00 | 0.00 | 0.13 | 0.10 | 0.06 | 0.00 |
| 7 | 0.14 | 0.28 | 0.18 | 0.07 | 0.06 | 0.00 | 0.09 | 0.07 | 0.07 | 0.04 |
| 8 | 0.11 | 0.20 | 0.19 | 0.11 | 0.09 | 0.00 | 0.04 | 0.11 | 0.09 | 0.06 |
| 9 | 0.03 | 0.09 | 0.15 | 0.12 | 0.14 | 0.00 | 0.03 | 0.11 | 0.15 | 0.18 |
| 10 | 0.04 | 0.09 | 0.14 | 0.18 | 0.16 | 0.00 | 0.01 | 0.05 | 0.14 | 0.19 |

*Notes:* Table show the shares of women that are in state $s_{t+1}$ at age 26 among those that were in state $s_t$ at age 25.

Table 1.12: Health status transition from one year to the next, males at age 25

| $s_t/s_{t+1}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.65 | 0.20 | 0.07 | 0.03 | 0.03 | 0.00 | 0.01 | 0.01 | 0.01 | 0.00 |
| 2 | 0.33 | 0.31 | 0.16 | 0.08 | 0.06 | 0.00 | 0.02 | 0.02 | 0.01 | 0.01 |
| 3 | 0.18 | 0.27 | 0.22 | 0.14 | 0.12 | 0.00 | 0.01 | 0.03 | 0.02 | 0.02 |
| 4 | 0.13 | 0.19 | 0.21 | 0.21 | 0.16 | 0.00 | 0.01 | 0.03 | 0.03 | 0.03 |
| 5 | 0.17 | 0.20 | 0.20 | 0.16 | 0.19 | 0.00 | 0.01 | 0.01 | 0.03 | 0.04 |
| 6 | 0.48 | 0.21 | 0.08 | 0.00 | 0.06 | 0.02 | 0.06 | 0.05 | 0.05 | 0.00 |
| 7 | 0.35 | 0.24 | 0.11 | 0.04 | 0.05 | 0.01 | 0.11 | 0.05 | 0.02 | 0.01 |
| 8 | 0.16 | 0.24 | 0.17 | 0.10 | 0.07 | 0.00 | 0.04 | 0.09 | 0.09 | 0.04 |
| 9 | 0.07 | 0.14 | 0.15 | 0.12 | 0.11 | 0.00 | 0.03 | 0.09 | 0.17 | 0.11 |
| 10 | 0.07 | 0.09 | 0.14 | 0.13 | 0.15 | 0.00 | 0.03 | 0.06 | 0.13 | 0.22 |

*Notes:* Table show the shares of men that are in state $s_{t+1}$ at age 26 among those that were in state $s_t$ at age 25.

Table 1.13: Health status transition from one year to the next, females at age 55

| $s_t/s_{t+1}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.42 | 0.24 | 0.13 | 0.05 | 0.02 | 0.00 | 0.04 | 0.05 | 0.03 | 0.02 |
| 2 | 0.17 | 0.29 | 0.21 | 0.09 | 0.05 | 0.00 | 0.04 | 0.07 | 0.05 | 0.03 |
| 3 | 0.09 | 0.22 | 0.23 | 0.15 | 0.07 | 0.00 | 0.03 | 0.07 | 0.09 | 0.05 |
| 4 | 0.05 | 0.10 | 0.21 | 0.23 | 0.14 | 0.00 | 0.01 | 0.05 | 0.12 | 0.10 |
| 5 | 0.05 | 0.11 | 0.16 | 0.20 | 0.20 | 0.00 | 0.01 | 0.04 | 0.10 | 0.15 |
| 6 | 0.28 | 0.05 | 0.13 | 0.03 | 0.03 | 0.08 | 0.23 | 0.18 | 0.03 | 0.00 |
| 7 | 0.14 | 0.17 | 0.10 | 0.05 | 0.04 | 0.01 | 0.17 | 0.17 | 0.10 | 0.05 |
| 8 | 0.05 | 0.10 | 0.13 | 0.09 | 0.04 | 0.01 | 0.08 | 0.22 | 0.21 | 0.08 |
| 9 | 0.02 | 0.05 | 0.10 | 0.10 | 0.06 | 0.00 | 0.03 | 0.13 | 0.31 | 0.20 |
| 10 | 0.02 | 0.05 | 0.05 | 0.08 | 0.09 | 0.00 | 0.02 | 0.07 | 0.21 | 0.41 |

*Notes:* Table show the shares of women that are in state $s_{t+1}$ at age 56 among those that were in state $s_t$ at age 25.

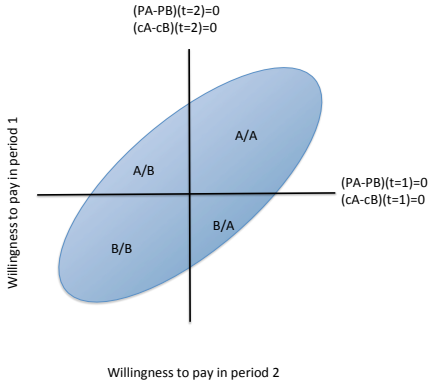
Table 1.14: Health status transition from one year to the next, males at age 55

| $s_t/s_{t+1}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.56 | 0.19 | 0.08 | 0.03 | 0.03 | 0.00 | 0.03 | 0.03 | 0.02 | 0.03 |
| 2 | 0.22 | 0.28 | 0.18 | 0.09 | 0.06 | 0.00 | 0.04 | 0.06 | 0.04 | 0.03 |
| 3 | 0.10 | 0.19 | 0.22 | 0.16 | 0.09 | 0.00 | 0.02 | 0.06 | 0.08 | 0.07 |
| 4 | 0.06 | 0.10 | 0.15 | 0.22 | 0.17 | 0.00 | 0.01 | 0.04 | 0.12 | 0.13 |
| 5 | 0.05 | 0.07 | 0.11 | 0.17 | 0.26 | 0.00 | 0.01 | 0.03 | 0.10 | 0.20 |
| 6 | 0.19 | 0.20 | 0.07 | 0.02 | 0.01 | 0.07 | 0.16 | 0.18 | 0.07 | 0.02 |
| 7 | 0.16 | 0.16 | 0.07 | 0.04 | 0.02 | 0.02 | 0.20 | 0.17 | 0.10 | 0.07 |
| 8 | 0.06 | 0.10 | 0.11 | 0.07 | 0.04 | 0.01 | 0.10 | 0.21 | 0.18 | 0.13 |
| 9 | 0.02 | 0.04 | 0.08 | 0.10 | 0.07 | 0.00 | 0.02 | 0.11 | 0.31 | 0.25 |
| 10 | 0.02 | 0.02 | 0.04 | 0.08 | 0.13 | 0.00 | 0.01 | 0.05 | 0.17 | 0.48 |

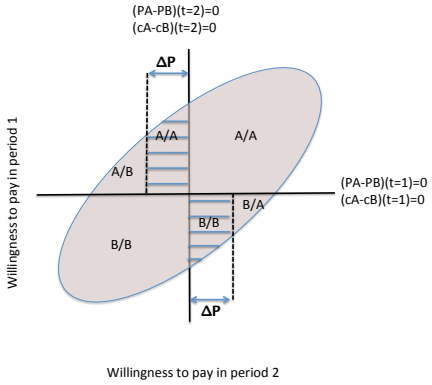
*Notes:* Table show the shares of women that are in state $s_{t+1}$ at age 56 among those that were in state $s_t$ at age 55.

Figure 1.1: Allocation with evolving preference heterogeneity and guaranteed-renewability

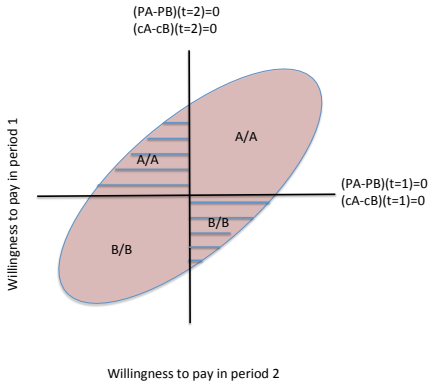
(a) Allocation of healthy individuals



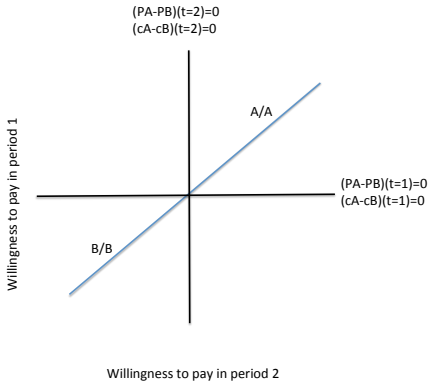

(b) Allocation of risky individuals




(c) Allocation of risky individuals

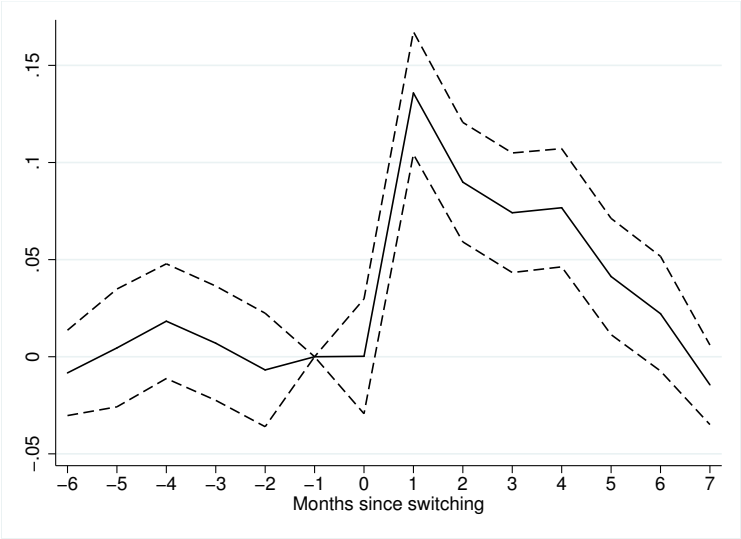



(d) Stable preference heterogeneity




*Notes*: The figures show the allocation across firms in two time periods, of individuals that have heterogeneous and time-varying preferences for two companies A and B. Panel (a) shows that the allocation of healthy individuals is efficient since in both periods the price they pay in each company is the same. Panel (b) and (c) show that some risky individuals inefficiently stay with their company because they are reclassified in the spot market, either through higher premiums or coverage denial. Panel (d) shows that there are no inefficiencies when preferences are stable over time.

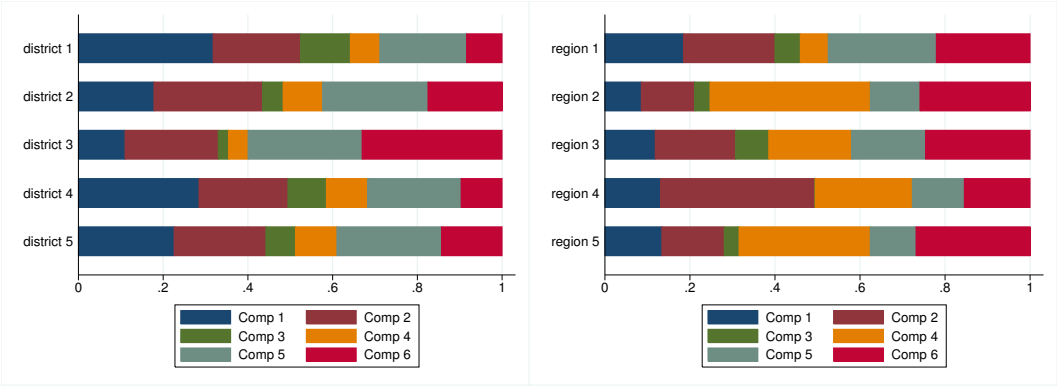
Figure 1.2: Probability of seeing a new provider




*Notes:* Figure plots the estimated coefficients from an event-study regression of the form given in equation 1.3. The dependent variable is a dummy indicator for seeing a new health service provider and time zero is the month of switching company. The bands around the point estimates are 95% cluster-robust confidence intervals (clustered at the individual level. The probability of seen a new health service provider after switching company is about 13 percentage points above the baseline the month after switching ISAPRE.

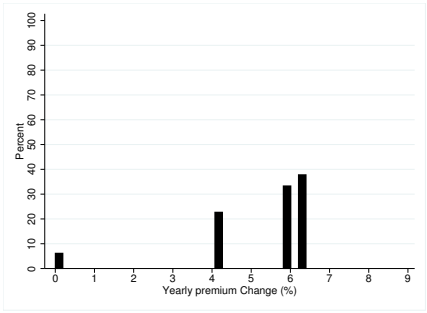
Figure 1.3: Market shares by geographic location

(a) Market share by region (b) Market share by district with region 1

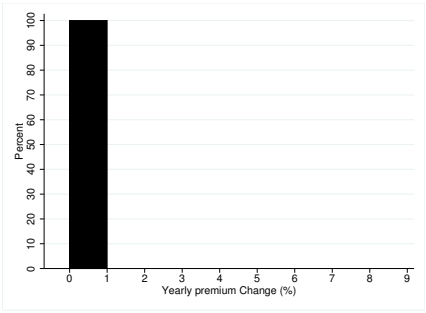



*Notes:* Panel (a) shows the market share of each ISAPRE in the 10 biggest regions of Chile. Panel (b) shows the market share of each ISAPRE in the 10 biggest district of the Santiago region

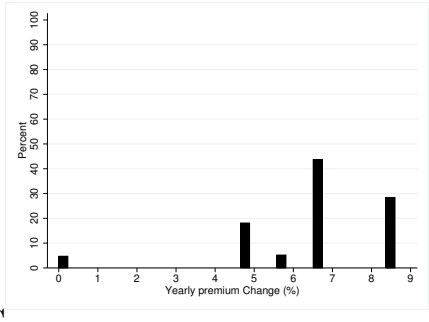
Figure 1.4: Premium change by insurance company and year



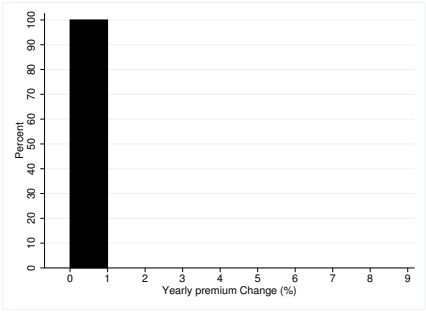

Company A



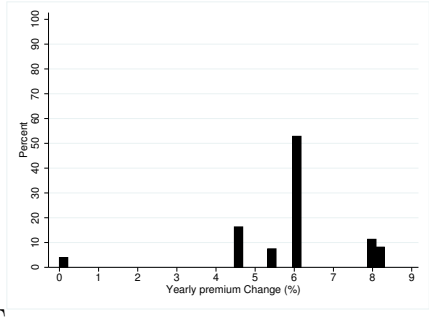

Company B



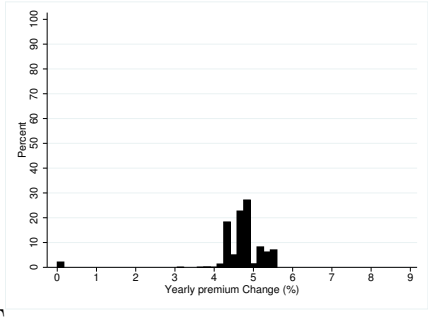

Company C




Company D



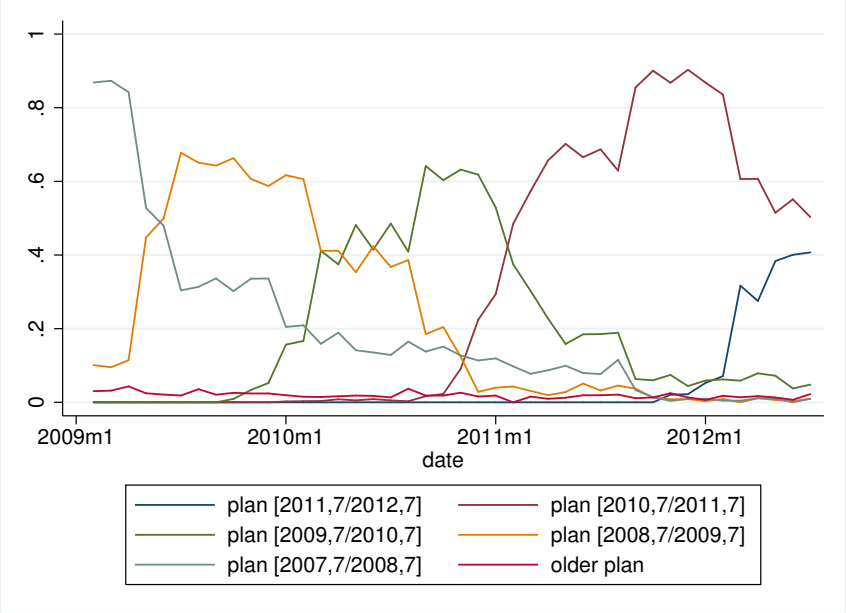

Company E




Company F

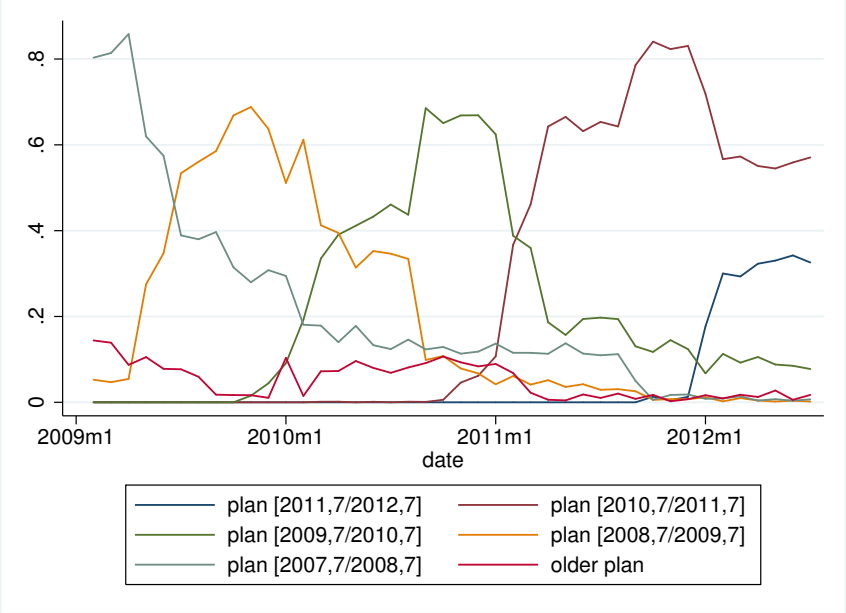
*Notes:* This figure shows the histogram of yearly price increases from 2010 to 2011 for each of the six ISAPREs in this study. It shows the practical workings of the "1.3" rule" described in the text that limits the variance of premium increases of contracts.

Figure 1.5: Cohort of destination plan among switchers across Isapres, by month



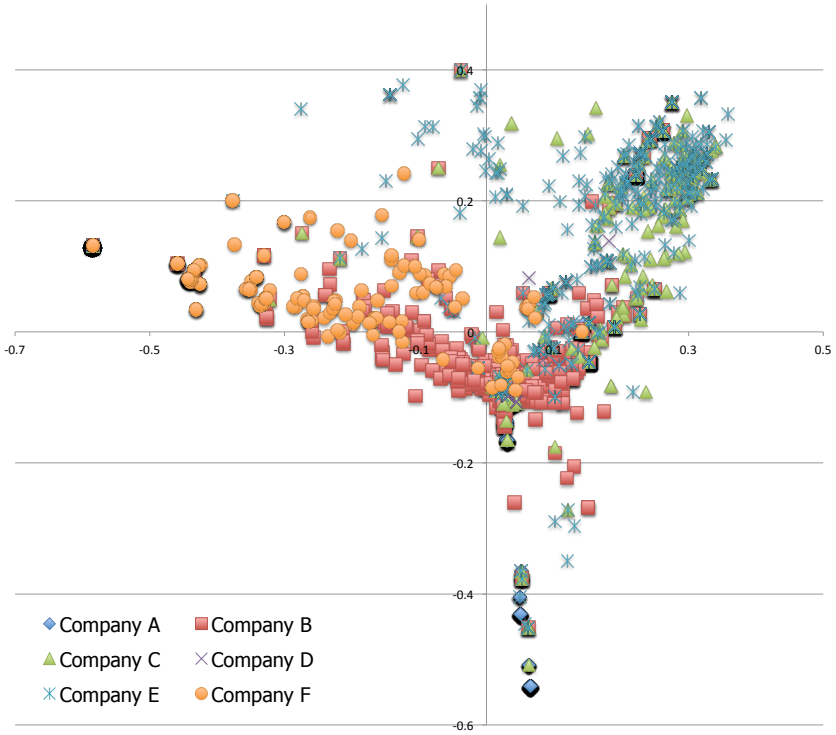

*Note:* Figure shows the share of switchers across Isapres by cohort of the destination plan at each point in time.

Figure 1.6: Cohort of destination plan among switchers within Isapres, by month



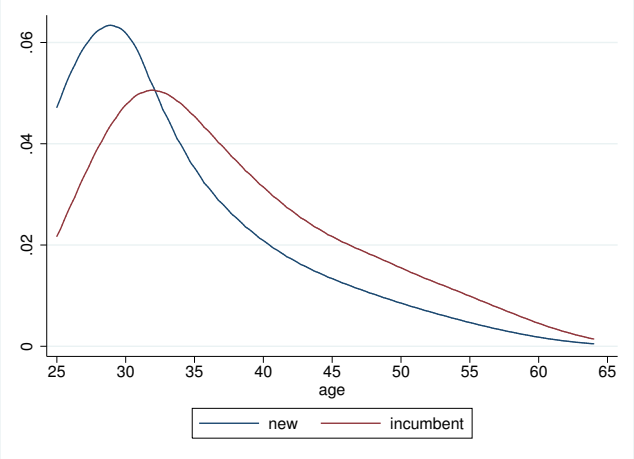

*Note:* Figure shows the share of switchers within Isapres by cohort of the destination plan at each point in time.

Figure 1.7: Provider distance across plans



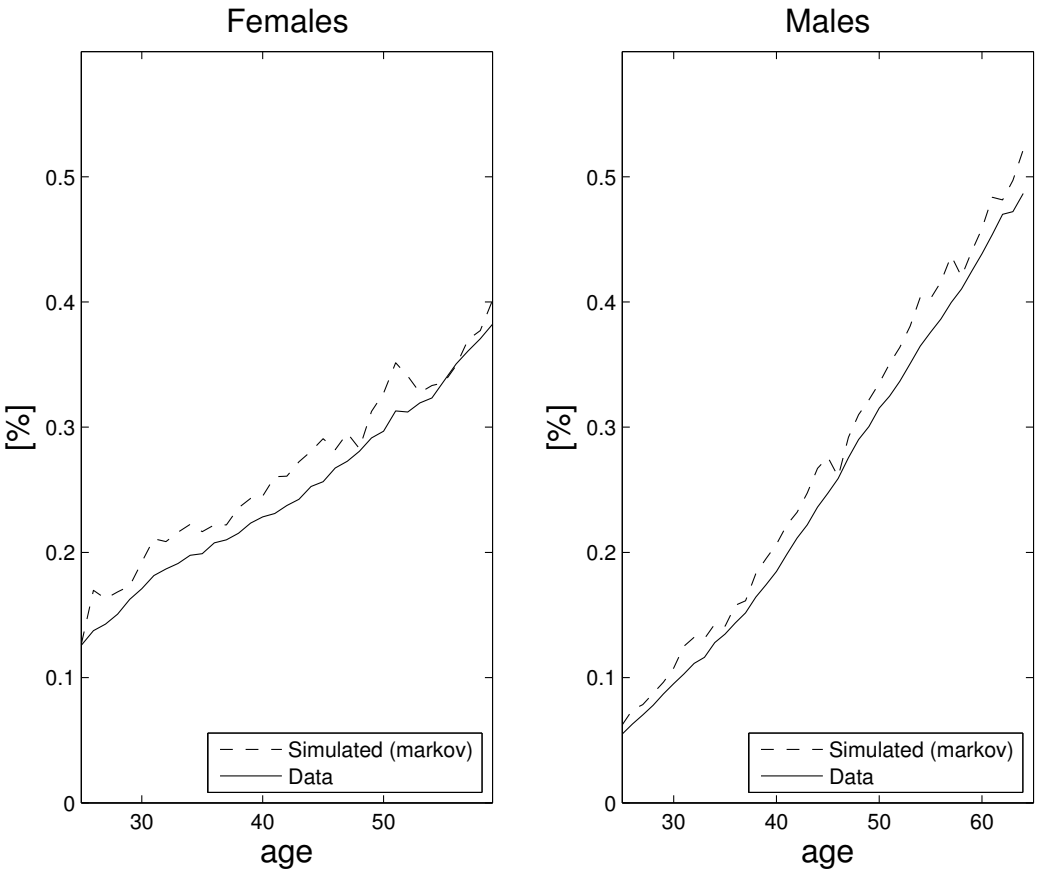

*Note:* Each dot is a restricted network plan in a euclidean plane to represent their distance $d \in [0, 1]$ in terms of the provider network. $d = 1$ if two plans do not share any providers. $d = 0$ for two plans with the same network. Colors represent different ISAPRES. 10 % subsample of plans

Figure 1.8: Age distribution of new and incumbent clients



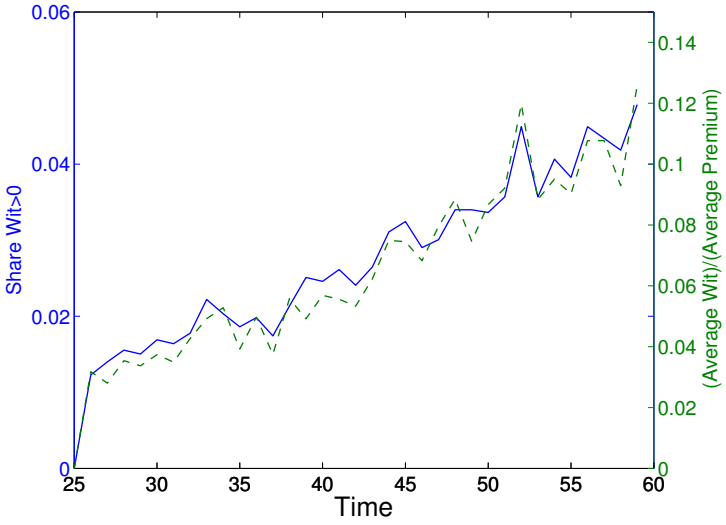

*Notes:* Kernel density estimate of age distribution among "cohort 0" and new enrollees

Figure 1.9: Predicted and actual prevalence of preexisting conditions



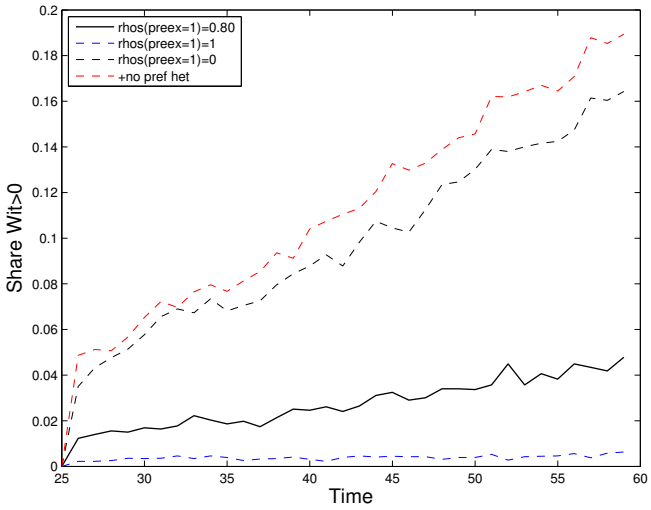

*Notes:* This graph shows the real and simulated probability of having a preexisting condition. The left panel corresponds to females and the right panel to males.

Figure 1.10: Share of individuals with $w_i^t > 0$ and average $w_{it}$



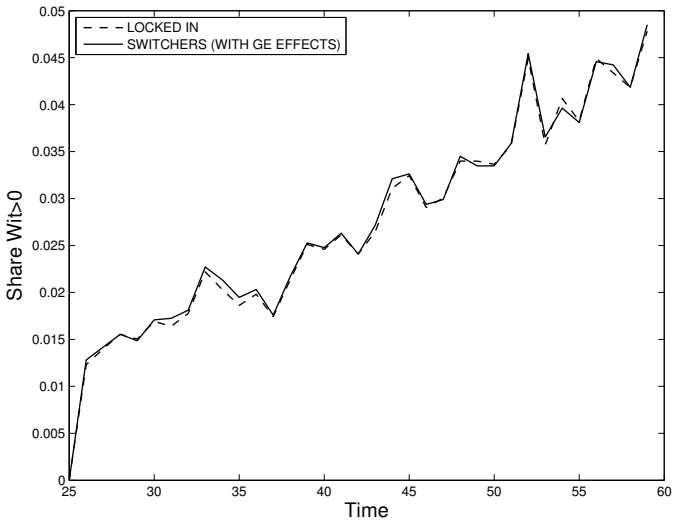

*Notes:* The full line (left Y axis) shows the share of individuals with $w_{it}$, as defined by equation 1.13 using the simulation method described in the text, representing the simulated share of individuals that would have picked a different company if preexisting conditions and risk-rating were banned. The dashed line (right Y axis) shows the average $w_{it}$

Figure 1.11: Share of locked-in individuals under different parameters




*Notes:* This figure shows the sensitivity of the lock-in result to the parameter estimates. The main results are represented by the full black line, that reproduces the result shown in figure 9. The dashed blue line represents the result assuming no coverage denial, and only assuming premium risk-rating. The black dashed line represents the results in the case of full coverage denial, so that all individuals with preexisting conditions are denied coverage in the spot market. Finally, the red dashed line shows the result with full coverage denial and assuming that there is no stable preference heterogeneity

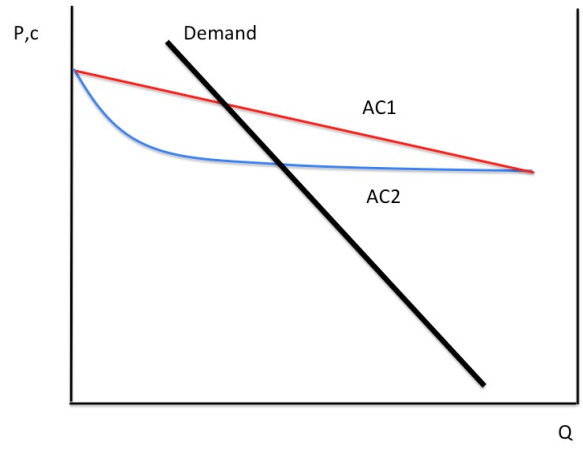
Figure 1.12: Simulated difference in switching rates between current policy and counterfactual policy



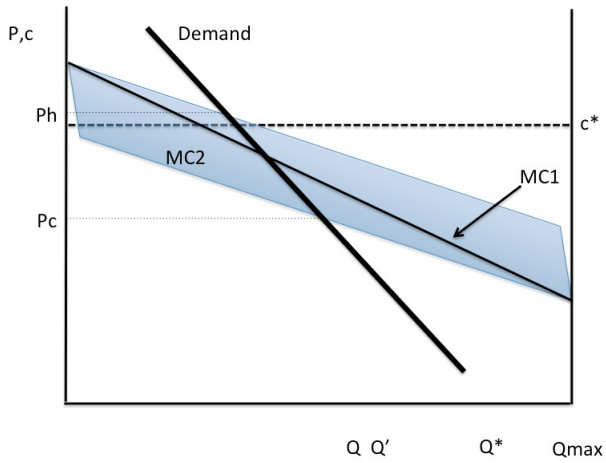

*Notes*: Figure shows the simulated switching rates of the counterfactual policy that bans preexisting conditions and underwritting relative to the simulated switching rates under current policy. Full line corresponds to switching rates across insurance companies and dashed line to switching rates within company.

Figure 1.13: Equilibrium effects of banning preexisting conditions with preference heterogeneity

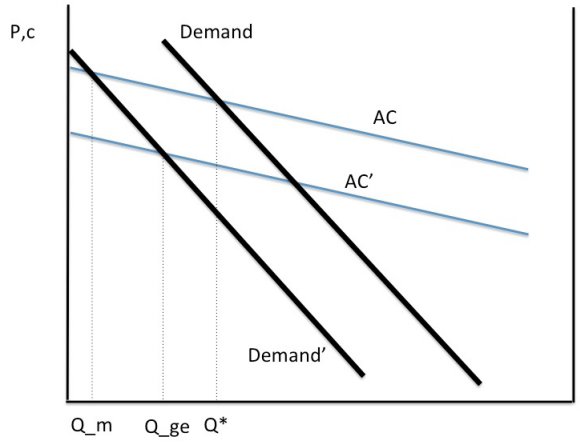
(a) Average Cost Curves



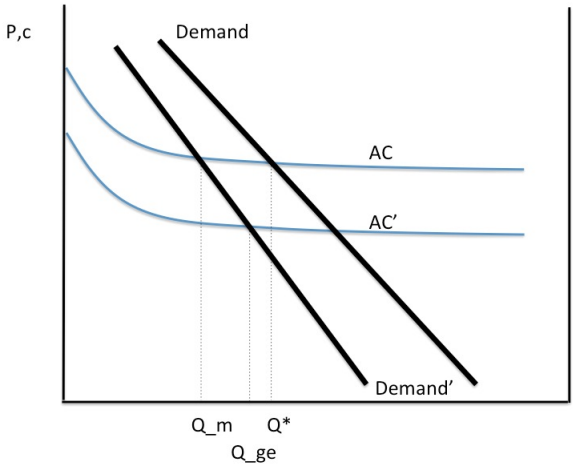

(b) Margina Cost




(c) Introducing Preexisting Conditions, No Preference Heterogeneity




(d) ntroducing Preexisting Conditions, With Preference Heterogeneity




*Notes*: Panel (a) shows average cost curves for the case of no preference heterogeneity within risk ($AC_1$) and for the case of preference heterogeneity within risk ($AC_2$). Panel (b) shows the respective marginal cost curve ($MC_1$) and marginal cost correspondence ($MC_2$). Panel (c) shows the mechanical effect and general-equilibrium effect of introducing preexisting conditions in the case of no preference heterogeneity. Panel (d) performs the same exercise in the case of preference heterogeneity.

I

# Chapter 2

# Peer Effects in the Emergency Department

## 2.1 Introduction

Workplaces commonly feature workers interacting to jointly produce output. An important question for the optimal organization of these workplaces is how peer interaction influences the productivity of individual workers. Theory is ambiguous on this question: on the one hand, if workers produce output in teams, a moral hazard problem arises in which individuals have an incentive to shirk; on the other hand, workers may motivate one another to provide higher effort (or, on the contrary, drag others down).[1]

Recent findings suggest that productivity spillovers operate in a variety of settings. Mas and Moretti (2009) find substantial productivity responses of grocery store cashiers to the introduction of a highly productive peer. Similarly, Falk and Ichino (2006) find that university students tasked with stuffing envelopes are more productive when working in the same room with a productive peer. Fruit pickers are less productive when they work in fields with their friends (Bandiera, Barankay, and Rasul, 2010).[2] These studies feature low-skilled workplaces, in which the social incentives of workers are thought to be stronger than in high-skilled workplaces. To this point, Guryan, Kroft, and Notowidigdo (2009) find no evidence of peer effects in random groupings of professional golfers.

In this paper, we add to the literature by documenting substantial productivity spillovers in a high-skilled, high-stakes occupation – physicians working in the emergency department (ED). The nature of clinical shift work in the ED lends itself to answering the question of how peers influence each other's productivity. Each patient is assigned to one and only one physician, who is primarily responsible for directing the care of that patient. In this sense,

[1] For a theoretical discussion, see Kandel and Lazear (1992).

[2] Other studies examine longer-run effects of coworkers on their peers' productivity, see for instance Jackson and Bruegmann, 2009. These studies use identification strategies based on more permanent changes in a worker's peer group, and identify parameters associated with long-term learning and human capital spillovers, rather than the transitory effects studied here.

production in the ED is physician-specific, but the load of work (demand) is shared across all physicians. The ED is a setting in which problems of free-riding may arise if incentives are not placed on individual productivity, but it is also a setting in which workers may be under various forms of peer pressure to keep up with demand.

We use data from two hospitals, one midsize hospital in the US and one larger private hospital in Chile. Our first finding is that productivity across physicians in each hospital is highly dispersed. This finding is in line with the vast literature in health economics that documents large productivity differences at both the individual and aggregate levels (Chandra and Staiger, 2007; Skinner and Staiger, 2009; Baicker and Chandra, 2004). Next, using variation in one's coworkers within a given shift, we find that working with a peer who is 10 percent more productive increases worker productivity by 1.5 percent. There is no spillover onto quality of care, consistent with physicians having some slack in their working up patients. This is plausible in the ED, where physicians typically have some discretion in the pace at which they alternate between their patients. To put this estimate in context, we calculate that replacing a physician from the 25th percentile of the productivity distribution with one from the 75th percentile for a twelve-hour shift would allow each coworker to care for one more patient in her shift. Physicians in our sample see an average of 15-20 patients per shift, so we view this as a large effect, in terms of keeping waiting times down and freeing up labor resources to handle higher-intensity cases.

This finding holds across the two hospitals despite their differences in organizational structure – notably that one hospital pays a bonus based on patients seen, whereas the other pays a flat hourly wage. Spillovers are observed only within physicians likely to see similar cases. Our estimated spillover is not driven by mechanical complementarities between workers, such as those arising from patient selection or from resource constraints. Finally, we find that in the US hospital in our sample, more productive workers are influenced more by their coworkers, and that most of the spillover is generated by working with a coworker in the lowest tercile of the productivity distribution. In the Chilean hospital, however, we find little evidence of heterogeneity in the spillover effect. This heterogeneity (or lack thereof) has implications for the optimal sorting of workers to shifts. If the goal were to maximize output, diversification of workers within shifts may actually be a bad idea, but if the emergency department cares about flows and preventing congestion, then they may still want to diversify shifts, so as not to have their least productive workers congesting the ED and leaving it unprepared in the event of an influx of high-acuity patients.

This paper makes several contributions. First, our setting is one in which worker productivity has substantive externalities. Emergency department overcrowding in the United States has garnered national attention in recent years. The Institute of Medicine issued a report in 2007 describing hospital-based emergency care as "at the breaking point" (Institute of Medicine, 2007). Recent evidence suggests that overcrowding of EDs is associated with reduced health care quality and patient safety (Fee et al., 2007; Hoot and Aronsky, 2008). Emergency physicians face increasing demands on their time within the hospital, and their ability to maintain high productivity throughout a shift is an important determinant of ED congestion.

We also contribute methodologically to the empirical literature on peer effects. Our basic empirical approach is similar to that of Mas and Moretti (2009), which uses within-shift changes in a worker's peer group to identify productivity responses to the introduction of a highly productive peer. However, our setting has a few key advantages that allow us to rule out more mechanical peer effects. The first concern is that the tasks assigned to a worker depend on the productivity of his peers. In our case, this amounts to physicians being assigned different types of cases when working with more or less productive peers. To address this concern, we devise a test of patient-physician assignment. We find that patients in each hospital are sorted to physicians primarily based on physicians' relative caseloads, similar to a queueing system. Most importantly, patient observable characteristics do not predict to which physician a patient is assigned at the time that patient arrives at the emergency department.

One potential confounder of previous estimates of peer effects in the workplace is that high- and low-productivity workers could differentially use shared resources in such a way that working with a low-productivity peer limits a worker's access to this resource, decreasing her productivity. We directly address whether resource constraints contribute to productivity spillovers by examining whether contemporaneous procedural utilization by slower peers accounts for the estimated spillover. For example, if less productive physicians tend to utilize more CT scans and x-rays, other physicians working at the same time as this physician will be slowed down simply because they will have to wait to access the imaging resources. We find that contemporaneous resource utilization of other physicians significantly slows down physicians. Nonetheless, our spillover estimates are robust to flexibly controlling for this resource utilization.

The bulk of our evidence suggests that peer effects play an important role in the emergency department, despite the workforce being highly trained and, in one hospital, compensated partly based on productivity.

## 2.2 Medical literature and context

In addition to adding to our economic understanding of workplace productivity, documenting the extent of peer effects and how they operate in the ED has significant implications for policy, patient care, and costs in the US healthcare system. Hospital costs in the US make up 30% of healthcare expenditures and physicians another 20% (Hartman et al., 2013); increasing efficiency of the hospital, most importantly the efficiency of the workforce within the hospital, promises to have large cost-savings and large external benefits to patients. In the ED, which is the point of entry to the healthcare system for many, efficiency is crucial to patient care. Overcrowding is one of the largest policy concerns facing emergency care (Institute of Medicine, 2007). Crowding has become more and more common over the past 20 years, as demand for ED services has risen, while the number of EDs operating has declined without any substantive increase in ED scale.

Since the 1986 enactment of the Emergency Medical Treatment and Active Labor Act

(EMTALA), EDs have been required by law to perform a medical screening on any patients arriving at the ED to determine need for care. Health care providers, including hospitals and physicians, point to EMTALA as one of the main reasons for the increase in demand for medical services and the decreasing financial viability and closure of many EDs from the mid 1990s to 2006 (Delia and Cantor, 2009).

Other long-term changes in the nature of ED caseloads have contributed to the need for efficient practices. Caseloads are quite mixed in most hospitals, and range from providing basic, time-insensitive care to those who cannot afford primary care and use the ED as their de facto primary care provider, to providing timely care to patients with acute, traumatic conditions, e.g., heart failure, major trauma from car accidents or gunshots, or severe stroke. Workloads for emergency physicians can be quite demanding and tend to fluctuate from one day to the next. The skills to work under stressful and unpredictable environments, and to manage multiple patients simultaneously are necessary for successful delivery of emergency care. Nevertheless, we find that physicians vary substantially in their efficiency; within each hospital in our sample, the most efficient physicians are 40-50 percent more efficient in patient care than the least efficient physicians. This illuminates the importance not only of the quantity of staffing that allows EDs to run efficiently, but also the mixture and abilities of the staff. On the intensive margin, physicians may be inefficient, but may be able to increase their effort when under substantial pressure to do so. This is not enough to keep an ED running smoothly, however, as large fluctuations in demand require staff to be in command of the patient load at all times.

Unsurprisingly, clinical staffing of EDs has been cited as a key factor in the process of keeping up with patient demand, and is likely more important than constraints on physical space within the ED (Institute of Medicine, 2007). Bed and labor constraints in the inpatient setting of the hospital (e.g., the intensive care unit) also contribute to crowding of the ED. To provide these various types of care efficiently, EDs need staff, especially physicians, who are capable of keeping up with ever-fluctuating demand for services.

## 2.3 Data

### Description

We use emergency room discharge data from two hospitals in our analysis. The first hospital (subsequently Hospital A) is a mid-size, non-profit urban hospital with about 40,000 emergency visits per year. We observe all emergency department discharges for Hospital A from November 2011 until early January 2013, regardless of whether the patient was discharged to home or admitted to the hospital. The emergency department has 25 beds, and the larger hospital in which the ED is nested has over 400 beds. Within these discharge records, we observe patient arrival time, patient complaint, patient gender, patient age, mode of arrival, disposition of discharge, and time of discharge. We also have administrative billing data which details primary diagnosis, charges, and CPT procedural coding used for billing.

Hospital A's patient mix is quite poor, and relies on public insurance. Medicaid is the primary insurer for roughly a third of the patients, while Medicare accounts for over a fifth of cases. Roughly 14% of cases report self-pay, i.e., no insurance, and the remaining share of cases are insured privately (see Table 2.1).

For 16 hours each day, Hospital A has two physicians on duty in the ED. During the early morning hours (1am to 9am), one physician staffs the ED. Figure 2.1 illustrates the typical shifts physicians work at Hospital A. The eleven emergency physicians at this hospital are not employees of the hospital, but all are partners of the same physician group. Physicians are compensated on a competitive hourly wage, and profit-sharing by the group is proportional to the share of total clinical hours worked. Compensation for clinical work thus boils down to hourly compensation with minimal additional incentives on quality or quantity of care.[3] The physician group also employs four physician assistants as midlevels who independently care for many of the least complicated cases that arrive at the hospital. The hospital employs the remaining labor, including nurses and technicians. When a patient arrives at Hospital A, details about the patient's case – time of arrival, severity, primary complaint, and other characteristics – are written on a public board in the center of the ED, and the patient waits for a physician to sign on to her case. Once a physician signs on to see a patient, the time a physician has signed on and the identity of the physician are documented on the central board.

Physicians working at Hospital A have all worked in the physician group for at least four years by the start of our sample. The oldest physicians have worked with the group for over 25 years and have practiced emergency medicine for their entire careers, while the younger physicians graduated from US medical schools roughly 10 years ago. Half of the physicians are female.

The second hospital (henceforth Hospital B) for which we have obtained discharge data is a large private urban hospital in Chile with roughly 80,000 emergency room visits per year. Patients at this hospital are mostly privately insured and come from the upper tail of the income distribution. Physicians from four specialties staff the emergency room at any given time.

Hospital B is amongst the two largest, highest-quality private hospitals in Chile. It was accredited by the Joint Commission International Accreditation in 2007 and 2010, and belongs to the network of partners of Johns Hopkins. The entire hospital has about 3000 employees, including 700 doctors. We observe 92 physicians staffing the emergency department over the course of fifteen months. Hospital B has a 24-hour emergency room with

[3] A formalization of the pay for physician $i$ working $h_{iy}$ hours in year $y$ is :

$$W_{iy} = h_{iy} w_y + \frac{h_{iy}}{\sum_{j=1}^{11} h_{jy}} \Pi_y$$

where $\Pi_y$ is the group's profits in year $y$, which depend on physician productivity, but also on a host of other revenue and cost determinants, such as patient mix, the billing department's efficiency in collecting revenues.

44 beds. A typical day shift includes a surgeon (who acts as the head of the shift), four internists, four pediatricians and two traumatologists. A typical night shift includes one surgeon, two pediatricians, two internists, and one traumatologist. Figure 2.2 provides a graphical depiction of the shifts at Hospital B.

In Chile, becoming a physician qualified to work in an ED entails five years of undergraduate coursework, two years of internship, and finally three years of specialization. Physicians mostly enter the ED after completing their specialization. The mean age in our sample is 36 years, and the average tenure is 20 months. Almost 90 percent of the physicians that see patients in the ED work at Hospital B only as ED physicians and are not part of the staff. In our sample the average physician works 7 shifts a month. Physicians who are not part of the staff complement their work at Hospital B with ED duties and/or private consultation at other hospitals. Staff physicians, on the other hand, have private offices in the Hospital and are stakeholders. Promotion to staff is rare and competitive. Only one or two physicians per year are promoted to a staff-level position.

The labor market for ED physicians in Chile is highly competitive and salaries are high. ER doctors at Hospital B receive a flat monthly salary of roughly US$6,000, plus a performance bonus at the end of the year of up to one month's salary.[4] Performance is measured in terms of number of patients seen, complaints received, and an evaluation by the head of the ED.

The high competition for ED physicians creates substantial turnover of the physicians working at Hospital B. Out of the 92 doctors in our dataset, 30 entered the hospital during the 15 months of our sample (October 2011-December 2012). Table 2.2 summarizes the main characteristics of Hospital B's emergency department physicians.

Despite differences in patient characteristics, patient flows at either hospital are quite similar. However, Hospital B has a median length of stay of each patient that is much lower than in Hospital A. Much of this difference is likely explained by differences in the types of care and the characteristics of patients in the two settings. In Hospital A, median throughput is just over 2 hours, and in Hospital B, median throughput is about 70 minutes. A median throughput of 2 hours is normal for US hospitals. Hospital B is a private, for-profit hospital with direct motivation to limit patients' waiting times. This likely influences Hospital B's median throughput times, whereas Hospital A is a non-profit with the majority of their revenues coming from public insurance. There is quite a bit of variation across specialties in Hospital B, which is unsurprising given the differences in patient pools treated by each specialty. These descriptives are shown in Figures 2.3 and 2.4.

## Sources of variation

In order to identify the effect of coworker ability on own productivity, a few conditions need to hold.

[4]Mean monthly income for salaried workers in Chile in 2011 was USD 710.

1. The assignment of shifts and coworkers must be uncorrelated with omitted factors that directly influence productivity.

2. Additionally, if one is interested in a form of peer effects that operate through social incentives, then assignment of patients to physicians within coworker pairs/groups needs to be uncorrelated with other factors that influence productivity. For example, unobservably more difficult patients may be sorted to the best physician on duty, inducing a mechanical peer effect.

From a policy perspective, this latter concern is less notable, because if one is interested in the optimal design of shifts, the spillover parameter of interest would incorporate both the spillover operating through peer pressure and the spillover operating through shifts in the nature of work for each individual. Nonetheless, to address the second concern, we condition on patient observables and on physician caseloads in all of our empirical analyses. We also present evidence that the assignment mechanisms in either hospital do not match patients to physicians based on any observable information except for the physicians' caseloads upon the arrival of the patient - so it is typically the least busy physician who is responsible for the next patient who walks in the door, regardless of that patient's characteristics or chief complaint. We examine the plausibility of these assumptions in this section.

**Physician scheduling**

Importantly, physicians in both settings are assigned to shifts well in advance, and with little room for switching shifts. We observe physicians working with a variety of peer groups throughout our samples. Physicians schedules are set at least two months in advance of any given shift at each hospital.

At Hospital A, physicians are expected to be flexible in their scheduling, and to work different days of the week, and different shifts of each day.[5] This contributes to the fact that we observe each physician pair working together quite regularly throughout the sample.

Table 2.3 shows the number of patients that each physician cares for when paired with any of the possible set of coworkers. There are substantial numbers of patients taken care of in most of the pairs, and by each physician within each pair. Physicians are ranked in this chart based on their estimated efficiency. The bottom right corner of the table indicates the number of patients cared for by pairs of very inefficient physicians, for example. One can also compare mirror elements of this table across the main diagonal to see whether a pair's caseload loads more heavily on one physician than on another. Indeed, from inspection of these elements, one can see that when there is imbalance in the efficiency of physicians on duty, the more efficient physician, who takes care of cases more quickly on average, provides care for more patients than the less efficient physician. When physicians are relatively close

[5]One caveat is that not all physicians work night shifts at Hospital A. Night shifts contribute little to the analysis, however, because half of the night shift is spent in single coverage. In the analysis, we control for hour-by-day of week effects, so that the spillover is estimated within an hour of the day, and only on cases in which peers were presently working in the ED.

in terms of estimated efficiency, they see similar caseloads when working together. This is one way in which the spillover effects may be thought to operate - more work is incurred by the faster of the physicians in any given pairing. To adjust for the fact that faster physicians see more cases than their slower counterparts, we include non-parametric controls for a physician's caseload when seeing a patient.[6] [7]

Physicians' caseloads are also largely a function of the number of patients available to see. We abstract from number of patients seen as the measure of output from hereon because of this concern.

It is highly atypical for either hospital to change the shift scheduling, or to call in extra (potentially more productive) physicians, when unexpectedly high demand periods occur. In this sense, there is little concern that variation in peer groups is being driven by demand.

**Patient arrivals**

There does not seem to be systematic assignment of patients to physicians based on physician productivity in either hospital. Patient characteristics, including age, sex, and complaint, are not major factors in determining assignment of patients to physicians. Instead, how much work a physician currently has and how much work his coworkers have (in terms of number of patients that have previously arrived) plays the most substantial role in determining patient-physician matching. This can be seen prominently in Figure 2.5, which plots the p-values from F-tests of joint significance of the labeled variables in within-pair regressions of physician identity on patient characteristics and the characteristics of each of the two physicians currently working at the ED, in particular their current caseloads. This test is easiest to implement in Hospital A, where physicians work in pairs and where we observe patient's characteristcs. We estimate models for cases within each physician pair of the form:

$$1[physician_c = i] = f(\alpha + X_c\beta + \gamma RelativeCensus_i + \epsilon_c)$$

where the identity of the physician assigned to the case is regressed on patient characteristics and the physician pair's relative census, defined as the ratio of the number of patients under care of physician $i$ to the total number of patients under care by physician $i$ and physician $j$, the other physician in the pair. We report estimates from linear probability models, although results look similar for logit regressions as well.

Under the null hypothesis that patients are equally likely to be assigned to either physician, conditional on that physician's current workload, the stacked p-values of the F-test of joint significance from each pair should resemble a uniform distribution. Visually, the only factors that appear to matter for patient assignment are the "Census" factors, which are

[6] Physicians are seeing at most $\mathbf{X}$ other patients during a given case, so we include dummies for seeing 1,2,..$\mathbf{X}$ other patients in the regressions.

[7] If physicians always held constant the number of cases under their names, the skew of the distribution of patients across physicians would be mechanical. Physicians have some say over the number of beds they are using at any given time, so slower physicians could theoretically see just as many patients, but occupy more bed-hours over the course of a shift.

comprised of two continuous variables, one for each physician's caseload when the reference patient arrives at the ED at Hospital A. The assignment mechanism seems to operate as expected – whomever is less busy signs on to the next patient, with little influence of the characteristics of those patients on the assignment.

Preliminary results for Hospital B, in which we estimate multinomial logit regressions within each team to test the assignment mechanism. We find that the main driver of assignment in Hospital B is also relative census, defined as the share of current patients under the care of each physician at the time the reference patient enters the ED. The main explanatory variable included in this regression, aside from relative caseload, is an interaction of the physician's estimated fixed effect with the triage severity rating of the patient. This tests whether more severe patients are sorted to physicians differentially by physician productivity. The p-values from the F-tests in Hospital B across teams of physicians are plotted in the second half of Figure 2.5.

The productivity spillovers of peers that we document are thus unlikely to be driven by selection of coworkers or of patients.

## 2.4 Econometric framework

Our primary measure of productivity is throughput, defined as the time between a patient arriving in the emergency department's waiting room and that patient's discharge from the emergency department. We focus on throughput for several reasons. First, it is a readily observable measure of productivity in most emergency room discharge data, and it is used widely in the literature on the operation of EDs.[8] Second, physician throughput contributes to waiting times, patient satisfaction, and the responsiveness of the ED to inflows of patients. Keeping up with demand is one of the largest challenges facing EDs (Institute of Medicine, 2007), and throughput is one of the most important factors in doing so. We acknowledge the fact that quality measures are particularly important as well in the context of medical care, and we address concerns that decreased throughput, while having positive externalities on the set of other patients, may have negative effects on the patient whose care is sped up.

For the bulk of our results, we use an empirical strategy common in the literature on workplace peer effects (Mas and Moretti, 2009; Bandiera, Barankay, and Rasul, 2010). Because the productivity data in our samples are high frequency, we can precisely estimate an individual worker's average productivity, net of patient, ED, and coworker characteristics by estimating a fixed effects regression. We then map these estimates of an individual's productivity to each individual when she serves in the role of a coworker. In a second regression, we use the estimated fixed effects mapped to the coworker as righthand-side variables in explaining variations in a physician's productivity.

[8]Beginning in 2014, CMS will begin using throughput in its outpatient payment updates for the public reporting system (AHRQ 2011)

To formalize this framework, assume that throughput $y_c$ of a given case $c$ is a function of case characteristics $X_c$, a physician's fixed level of productivity $\theta_i$, time fixed-effects $\gamma_t$ and the identities of other physicians currently working in the ED:

$$y_c = \alpha + X_c\psi + \theta_{i(c)} + \gamma_{t(c)} + P_{i(c)}(\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k) + \epsilon_c$$

In this form, $c$ indexes the case, which has a number of characteristics, including a physician $i$ assigned to case $c$, $i(c)$. $P_{i(c)}()$ is the effect of the fixed productivity of a physician's current peer group on her own productivity (relative to her fixed productivity). $P$ is written as if it may vary across physicians. In the first step of the estimation, we replace $P_{i(c)}()$ with a set of dummies, where each dummy is unique to the set of identities in the peer group. In Hospital A, this amounts to including an indicator for which physician is coworking with physician $i(c)$, while in Hospital B, the coworker group comprised of workers $W_1$ and $W_2$ is associated with its own dummy, and the coworker group consisting of workers $W_1$, $W_2$ and $W_3$, for example, is associated with a separate, mutually exclusive dummy variable. The point of this first regression is to retrieve estimates of $\theta_i$, the physician fixed effects purged of the influence of her peers.

In the second step of this regression analysis, we replace $P_{i(c)}()$ with $\beta\frac{1}{N_{cow}}\sum\hat{\theta}_{-i(c)}$, the average fixed effect of physician $i$'s coworkers who are on duty while physician $i$ is taking care of patient $c$.[9] The estimating equation is:

$$y_c = \alpha + X_c\psi + \theta_{i(c)} + \gamma_{t(c)} + \beta\frac{\sum\hat{\theta}_{-i(c)}}{N_{cow}} + \epsilon_c \tag{2.1}$$

This linear-in-means assumption is a functional form that is standard in the literature on peer effects. In Hospital A, the estimated fixed effect of one's coworker is the average fixed effect. In Hospital B, more information is lost by collapsing coworker characteristics down to their mean efficiency, but we opt for this functional form for the sake of comparison to other studies.

## 2.5 Results

We begin by presenting results of the baseline regression specification for Hospital A, using log(throughput) as the outcome variable in both steps of the analysis. Columns (1)-(3) of Table 2.4 present evidence of productivity spillovers with increasingly large sets of control variables. Each of these columns use variation in a physician's peer group arising within an hour of the day and day of the week, and within a month. This amounts to comparing a physician's average throughput when working with different coworkers on, for example,

[9]We have experimented with Bayesian shrinkage techniques when including the estimated fixed effects as regressors, and doing so does not affect our results. Bayesian shrinkage hardly alters our estimated fixed effects, because our within-physician sample sizes are quite large and our estimated physician fixed effects are precise.

different Tuesdays, in the same hour of the day. We include increasingly large sets of controls, and in column (3), we include all patient observables (complaint, age bin, gender, mode of arrival, disposition of discharge) and time-varying physician and ED observables (physician caseload, patients in and arriving to the ED). Including the physician caseload is particularly important as slow coworkers could slow down the physician because they might generate a higher caseload to her.

The point estimate of the productivity spillover is remarkably stable across these specifications, and is precisely estimated. We estimate that having a peer who is 10 percent more efficient over the course of the sample induces a physician to provide contemporaneous care that is 1.8% faster. The magnitude is similar, and slightly larger, than previous studies.[10]

Column (4) of Table 2.4 explores a different source of identification – within-day variation in one's coworker group. In this specification, we include a dummy for each calendar date in the sample, so that the identifying variation in peer groups is arising through changes in one's peers within a given day. The result for the productivity spillover is similar, at 1.43%. The estimate is quite robust to a number of alternative specifications not presented here, including leaving out any patients who are admitted, exchanging the main independent variable with a dummy for whether your coworker has an above median fixed effect.

Remarkably, the baseline estimates of spillovers in Hospital B are quite similar, despite the differences in organization and operation from Hospital A. In Column (1) of Table 2.5 we present the results of estimating 2.1 for the entirety of Hospital B, using a similar specification to the one used for Hospital A. The same specification is estimated by each specialty within Hospital B in columns (1), (3) and (5) of Table 2.6. In each table, the peer group on each case is the set of physicians active during the case *within the specialty* other than the physician assigned to the case. We construct the peer group this way primarily because the design of Hospital B is such that physicians primarily interact with other physicians in the same specialty.[11]

The estimated magnitude of these spillovers is in line with previous studies of workplace spillovers (Mas and Moretti, 2009; Falk and Ichino, 2006). In the emergency department, efficiency spillovers of this magnitude could have potentially large impacts on the operation of the workplace.

The dataset for Hospital B allows us to control for the number of procedures. In principle, slower coworkers might also be those prescribing more procedures that would slow down the case of the rest of physicians. We address this concern by incorporating the number of procedures of the cases seen by other physicians of the same speciality that haven't been discharged at the time of check-in. We also incorporate the number of procedures of the case an additional control. The results are in Column (2) and (3) of Table 2.5 and Columns (2), (4) and (6) of Table 2.6. The point estimate of the peer effect decreases to 0.167, but it is still economically and statistically significant. Breaking down by specialty reveals that

[10]Mas & Moretti find that having a peer group that is 10% more productive increases own productivity by 1.6 percent. Falk & Ichino estimate the effect at 1.4 percent.

[11]In later specifications, we consider the spillovers occurring across specialties within Hospital B.

the point estimate decreases for all specialties and it not longer statistically significant for internists.

The prevalence of ED overcrowding is in no small part a function of individual productivity. The estimated distributions of efficiency measures in each hospital (and within specialties in Hospital B) have wide ranges; the fastest physicians are on average 40-50 percent faster than the slowest physicians, with no evidence of lower quality care.[12] One simple interpretation of our spillover estimates is that physicians who are on the bottom end of this distribution are not only slowing down the flow of work in the ED through their own patient care. These physicians are also generating slowdowns for the patients under the supervision of other on-duty physicians. Of course another interpretation is that the fastest of the physicians generate larger efficiency gains than would be implied by their estimated efficiency measures, through mechanisms such as monitoring, peer pressure, or motivation. In the next section, we explore the possible mechanisms underlying the spillovers we have documented.

How do the spillovers we estimate arise? These spillovers could arise due to peer pressure as in Mas and Moretti's study of cashiers, where less productive workers incur a disutility from being further below the prevailing effort norm, especially when peers can generate this disutility through monitoring. Alternatively, motivation to work harder in the presence of more productive peers may explain why physicians work faster when their peers are faster. Physicians could be influenced by their coworkers through knowledge spillovers. Finally, physicians could be complements in production, so that the marginal effort of one physician is more valuable when working with certain other physicians.

We present a number of tests to differentiate between or rule out the many stories that could lead to a physician working more efficiently when around more efficient peers in this section.

## Congestion of shared resources

A mechanical way in which one worker's productivity could influence the productivity of others is through utilization of shared resources. If less productive peers perform tests and procedures that are more resource-intensive, and if these resources are shared and scarce, then surrounding physicians would be unable to access these resources as quickly and may be slowed down as a result. This is one form of productivity spillovers that relies only on the production technology. Ex ante, this may an important mechanism through which the spillovers we estimate arise.

To test for this channel, we include as a regressor the number of procedural orders by other physicians at the time a physician begins with the current patient. The orders we observe in the Chilean data are primarily for imaging – x-rays and CT scans; these are truly a scarce resource, as the emergency department has only a few machines that perform these

[12] In later tables, we show that quality measures such as readmissions seem not to vary with physician speed.

tasks. If the spillover we estimate is generated by congestion effects that are stronger when a physician works with slower, resource-congesting peers, then conditional on the current utilization of resources by a physician's peers, the spillover parameter should attenuate to 0. This is not the case. We find a modest degree of attenuation of the spillover parameters across specialties in Hospital B when we additionally include the procedural orders of other physicians of the same specialty in the regressions in Table 2.6. The spillover estimate is most attenuated for internists; the estimate drops from .152 to .089. For pediatricians, the estimate drops from .176 to .135, and for traumatologists, the spillover parameter actually increases slightly from .243 to .265. As expected, other physicians' procedural orders are estimated to slow down patient care. The coefficients on others' procedural orders are precisely estimated; an additional 5 procedural orders on other patients when the current patient arrives slows down her care by .65 to 2.4 percent. This channel may explain some of the productivity spillover among internists in the Chilean hospital, but little to none in other specialties.

## Differential responses

The spillover of a coworker's productivity onto one's own productivity may have a few types of heterogeneity. First, some physicians may be more or less likely to respond to their peers. A more efficient physicians may be unaffected by her peers if the spillover is being driven by peer pressure from faster physicians onto slower physicians. Alternatively, if a less efficient physician is unreceptive to peer pressure, then her effort will not vary with whether her peers are high or low productivity. Another type of heterogeneity lies in the spillover generated by different types of workers.

One model of social interaction that may help clarify why there could be differential responses is one of inequity aversion (Fehr and Schmidt, 1999) – in which agents have a disutility associated with putting in more or less effort than their peers. If there is an asymmetry in how much disutility is generated by putting in some effort amount greater than one's peer compared to putting in the same amount less effort than one's peer, then one should see differential responses of physicians to different types of peers. In particular, if physicians have greater distaste for being the higher effort peer than for being the lower effort peer, less productive peers will tend to bring down a physician's effort more than more productive peers will bring up a physician's effort and thus efficiency. This is only one model for thinking about these differential effects, and we only try to clarify some ways in which differential responses to peers may be driven by behavioral mechanisms other than peer pressure.

Here we present the results of estimating Equation 2.1 with interaction terms to capture these forms of heterogeneity. In Table 2.9, we find that faster physicians are more susceptible to influence from their peers productivity than slower physicians, i.e., the interaction term in Column (1) on whether the reference physician is faster than median with her coworker's productivity has a positive and significant coefficient. Other forms of heterogeneity, as estimated in the subsequent columns show little evidence for differential effectiveness. However, in Table 2.10, we estimate the efficiency of fast, median, and slow physicians when working

with physicians of who are themselves fast, median, or slow. The spillovers become quite evident by inspection of this table, and it becomes clear that most if not all of the response to one's peers are being generated by working with the slowest peers in Hospital A.[13]

We estimate similar models in Hospital B, the results of which are presented in Tables 2.11 and 2.12. The results of these interaction models in Hospital B do not seem to support any large differences in the incidence of spillovers across physicians. All types of workers respond positively to their peer group's productivity. These results do not support one of the primary findings of Mas & Moretti (2009), that spillovers are mostly onto the less productive workers and are generated primarily by the most productive workers. As such, the mechanism for the spillover in the ED is unlikely to be the result of peer pressure through monitoring of slower physicians by faster physicians.

Considering the evidence from both hospitals, it does not appear that more productive peers are generating the spillover. If anything, in Hospital A, the effect is driven primarily by the slower physicians.

## Knowledge spillovers and expertise

Another potential mechanism that could generate spillovers across both the high and low productivity physicians is a model with comparative advantage. If some physicians specialize in certain types of patients, they may lend their expertise to other physicians who encounter patients of those types. In Hospital A, this is more likely to happen, as physicians are responsible for caring for all types of patients, whereas in Hospital B, physicians explicitly specialize and only accept cases that fall under their realm of expertise.

A rough test of whether spillovers in efficiency are driven by physicians sharing expertise on a particular case is to estimate patient-type-specific throughput for each physician and to estimate spillovers for each patient-type. These tests are underpowered in our sample, but we present the results in Table 2.13 for completeness. In short, physicians do not seem to be specialists; patient-type-specific fixed effects for physicians are all highly correlated with one another. Physicians seem to be either fast or slow on all types of cases. Formalizing this section is a goal of future research.

## Contagious enthusiasm or malaise

Spillovers may operate because the composition of entire groups of workers may create an atmosphere of hard work or malaise. If this were the mechanism through which spillovers operated, then one would expect the average efficiency of the entire set of physicians working in the hospital to generate spillovers. The structure of the workplace in Hospital B is such that physicians in separate specialties see one another and interact in the hallways of the ED, but their workloads are distinct. This provides a compelling research design

[13]We cannot separately identify whether physicians are slowed down by the slow or sped up by others, rather we can only identify the relative effects.

to test for the presence of spillovers operating through contagious enthusiasm or malaise. Spillovers should operate across specialties. In Table 2.14, we estimate models in which we allow for the spillover onto physicians in one specialty to be a linear combination of the average productivity of coworkers in the same specialty and the average productivity of coworkers in all other specialties. We find that spillovers across specialties are insignificant and quite small relative to the spillovers operating within specialty. In the case of internists cross-specialty spillovers are estimated to be negative. Spillovers are arising mostly through whom physicians are most directly interacting with, those with whom they share a common workload and common experience.

## Spillovers across the shift

The evidence provided so far does not point to any one particular mechanism generating the observed spillovers. Another approach to discerning how these effects are arising is through direct investigation of changes in coworker. In the following section, we present preliminary evidence using an event-study design to gain more insight.

Is the slowdown from working with a slower peer immediate, or is it something that only manifests over the course of the shift? If the effect sets in gradually, then one might think that the effect is coming about through a slowdown of the emergency room resulting from the new slower physician's procedural orders. If instead the effect is immediate and persistent, the spillover is more likely generated through some form of social interaction, rather than shared resources (e.g., CT scanners, x-rays, nurses) being slowly backlogged by orders from the new slower physician. Additionally, if backlogging of resources is generating the slowdown, it should persist into the future after the slow physician has been replaced by a faster physician.[14]

To this end, we present evidence of the timing of the spillover effect, as evidenced by a set of event studies. We focus on Hospital A because shifts there overlap imperfectly, allowing us to examine throughput responses to discrete changes in the efficiency of your coworker. We estimate models of the form:

$$y_c = \alpha + X_c\psi + \theta_{i(c)} + \gamma_{t(c)} + \sum_{s=-5}^{5} \beta_s \Delta\hat{\theta}_{-i,c+s} + \epsilon_c \tag{2.2}$$

That is, we calculate changes in a physician's peer group's fixed productivity over the course of her shift, and we regress her throughput on a case on the usual set of controls, and on lags and leads of changes in her coworker productivity. The parameters of interest in this equation are the set of $\beta$s, which index the patient of interest, relative to the case on which

[14]Our first piece of indirect evidence addressing the timing of the effect is from Column (4) of Table 2.4, which shows that variation in a physician's peers within a calendar day is still associated with a spillover. Since some of the shift overlaps are not that long in Hospital A, one would not expect the physician working the second morning shift to be affected much by the physician working the evening shift if indeed the effect is cumulative. However, the spillovers are estimated to be just slightly smaller within day.

a physician has a new coworker ($s = 0$). In Figure 2.6 we present results from this exercise in two ways. In the first panel of Figure 2.6, we plot the basic estimates of 2.2, which demonstrate an immediate, persistent effect of new coworkers on a physician's productivity throughout the rest of the shift.[15] In the second panel, we examine how this effect varies when looking only at the entry of new slower physicians to the shift. The visual evidence is striking. There appears to be a large anticipatory effect of having a new slower coworker. Physicians provide speediest care to the patients they care for just before the entry of a new slower coworker. Once the slower coworker begins a shift, the reference physician slows down care for future patients considerably. This is inconsistent with explanations of the spillover operating through overuse of shared resources. Instead, this reaction may indicate that slowdowns are partly generated by transitions and reorganization of work upon the entry of new physicians. However, the event study is not simply examining the impact of switching coworkers; instead it provides an estimate of the impact of switching to *slower* coworkers, as the plotted estimates are regression coefficients on changes in estimated coworker fixed effects from the first-step regression.

## 2.6 Conclusions

Unlike most of the previous evidence in the literature, this paper documents peer effects in a high-skilled and high-stakes profession. In particular, we show important peer-effects among physicians in the emergency department. If management were to replace a physician from the 25th percentile of the estimated productivity distribution with one at the 75th percentile for a 12-hour shift, other physicians would be each able to see an additional patient. Our findings are surprisingly similar across two different institutional settings with different payment incentives.

Our data allows us to discard several alternative explanations that would mechanically generate spurious peer-effects. We devise a test for non-random patient-physician assignment, to rule out the possibility that productive physicians are assigned more complex cases when working with low-productivity peers. We find that cases are distributed across physicians as a function of their caseload, and that patient's characteristics do not predict to which physician they are assigned. On the other hand, we do not find evidence that slow physicians decrease the productivity of their peers through an increase in utilization of scarce resources. The data analyzed in this paper provides fruitful avenues which we plan to explore in future research. An important feature of emergency department care is the hierarchical interaction between physicians, who largely order and interpret tests, and nurses, who are primarily responsible for carrying out a physician's orders. The hospitals in our data both feature quasi-random assignment of physicians and nurses to cases, making it possible for us to assess empirically the importance of these hierarchical relationships in the workplace.

[15]Evidence not shown here suggests that physician throughput does not vary much across the shift, and physicians may in fact speed up slightly through the shift, so the effect shown here is not one due to tiring of physicians across the shift.

Given high rates of turnover of physicians in Hospital B, we can also assess the degree to which learning to work together matters for speed, spending, and outcomes, net of overall on-the-job learning.

Table 2.1: Payer mix at Hospital A

| Insurance Type | |
|---|---|
| Medicaid | 33.1 |
| Medicare | 21.0 |
| Private | 30.2 |
| Uninsured | 15.8 |
| Total | 100.0 |

*Notes:* Share cases by payer, at Hospital A

Table 2.2: Physician characteristics at Hospital B, October 2012

| | N | mean | standard deviation | min | max |
|---|---|---|---|---|---|
| Gender (1 if male) | 93 | 0.61 | 0.49 | | |
| Age (in years) | 92 | 37.4 | 6.02 | 28 | 54 |
| Tenure (in months) | 93 | 20.6 | 26.8 | 0 | 125 |

*Notes:* This table shows descriptive statistics of physician characteristics for Hospital B. Tenure is as of October 2011

Table 2.3: Frequency of interaction of pairs, ranked by physician efficiency

| FE/FE | -0.186 | -0.175 | 0 | 0.006 | 0.028 | 0.058 | 0.069 | 0.222 | 0.236 | 0.283 | 0.323 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -0.186 | 0 | 337 | 425 | 428 | 435 | 403 | 399 | 267 | 292 | 115 | 357 |
| -0.175 | 341 | 0 | 153 | 249 | 268 | 246 | 178 | 192 | 232 | 24 | 269 |
| 0 | 415 | 192 | 0 | 315 | 211 | 191 | 427 | 299 | 364 | 61 | 392 |
| 0.006 | 344 | 251 | 247 | 0 | 342 | 122 | 259 | 172 | 177 | 61 | 261 |
| 0.028 | 445 | 284 | 204 | 390 | 0 | 123 | 156 | 193 | 221 | 17 | 373 |
| 0.058 | 326 | 212 | 152 | 124 | 118 | 0 | 110 | 96 | 127 | 15 | 237 |
| 0.069 | 515 | 224 | 496 | 337 | 186 | 147 | 0 | 287 | 177 | 44 | 282 |
| 0.222 | 198 | 138 | 201 | 165 | 146 | 105 | 182 | 0 | 116 | 30 | 176 |
| 0.236 | 234 | 203 | 347 | 169 | 179 | 138 | 151 | 146 | 0 | 54 | 309 |
| 0.283 | 68 | 13 | 47 | 52 | 14 | 13 | 32 | 29 | 38 | 0 | 20 |
| 0.323 | 287 | 206 | 318 | 258 | 298 | 226 | 200 | 207 | 285 | 16 | 0 |

*Notes:* This table shows the number of patients that each physician cares for when paired with any of the possible set of coworkers. Physicians are ranked in this chart based on their fixed effect in throughput regressions. Larger FE indicate slower physicians

Table 2.4: Estimated spillovers in Hospital A

| | (1) $log(throughput)$ | (2) $log(throughput)$ | (3) $log(throughput)$ | (4) $log(throughput)$ |
|---|---|---|---|---|
| $\theta_{coworker}$ | 0.170*** | 0.189*** | 0.186*** | 0.143*** |
| | (0.038) | (0.036) | (0.025) | (0.034) |
| Hour-by-day-of-week effects | No | Yes | Yes | No |
| Month effects | No | Yes | Yes | No |
| Calendar date effects | No | No | No | Yes |
| Hour effects | No | No | No | Yes |
| Other controls | No | No | Yes | Yes |
| $R^2$ | 0.058 | 0.083 | 0.419 | 0.429 |
| N | 23350 | 23350 | 23307 | 23307 |

*Notes:* Standard errors clustered at the physician level. Other controls include dummies for mode of arrival, age of patient, sex of patient, and patient complaint. Controls also include: number of visits by the patient in the sample period, the total number of patients under the care of a given patient's physician when a patient arrives, the total number of patients that have yet to be discharged by the ED when a patient arrives, and the total number arrivals in the 7 30-minute intervals before and the 6 after the arrival of a patient. Standard errors in parentheses, * p<0.10, ** p<0.05, ***p<0.01)

Table 2.5: Estimated spillovers in Hospital B

| | (1) $log(throughput)$ | (2) $log(throughput)$ | (3) $log(throughput)$ |
|---|---|---|---|
| Average Perm. Prod of Coworkers | 0.204*** (0.0364) | 0.170*** (0.0412) | 0.167*** (0.0407) |
| Number of Coworkers | 0.0144*** (0.00429) | 0.00701** (0.00339) | -0.000233 (0.00330) |
| proc_other | | | 0.00148*** (0.000247) |
| procedures | No | Yes | Yes |
| Observations | 67468 | 67468 | 67468 |
| $R^2$ | 0.153 | 0.466 | 0.467 |

*Notes:* All specifications include day-of-week-hour and month dummies severity, time elapsed during the shift and total number of patients under care of physican when the patient arrives. *proc_other* is the number of procedures of other cases not discharged at the time of check-in of the focal case, seen by other physicians of the same specialty. Standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01

Table 2.6: Estimated spillovers in Hospital B, by specialty

| | (1) Internists | (2) Internists | (3) Pediatricians | (4) Pediatricians | (5) Traumatologists | (6) Traumatologists |
|---|---|---|---|---|---|---|
| Average Perm. Prod of Coworkers | 0.152** | 0.0888 | 0.176*** | 0.135** | 0.243*** | 0.265*** |
| | (0.0672) | (0.0615) | (0.0428) | (0.0604) | (0.0646) | (0.0616) |
| Number of Coworkers | 0.00692 | -0.00261 | 0.0199*** | 0.00132 | 0.0234 | 0.0134 |
| | (0.00454) | (0.00408) | (0.00610) | (0.00404) | (0.0139) | (0.0130) |
| *proc_other* | | 0.00137*** | | 0.00167*** | | 0.00474*** |
| | | (0.000272) | | (0.000347) | | (0.00157) |
| procedures | No | Yes | No | Yes | No | Yes |
| Observations | 29966 | 29966 | 28572 | 28572 | 8930 | 8930 |
| $R^2$ | 0.069 | 0.374 | 0.093 | 0.523 | 0.093 | 0.283 |

*Notes:* All specifications include day-of-week-hour and month dummies severity, time elapsed during the shift and total number of patients under care of physican when the patient arrives. *proc_other* is the number of procedures of other cases not discharged at the time of check-in of the focal case, seen by other physicians of the same specialty. Standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01

Table 2.7: Quality of care, Hospital A

| | Readmits within 14 days same complaint | | Readmits for UTI within 14 days | |
|---|---|---|---|---|
| $log(throughput)$ | 0.0000617 | | 0.000104 | |
| | (0.00187) | | (0.000643) | |
| $\theta_i$ | | 0.00275 | | 0.000564 |
| | | (0.00769) | | (0.00215) |
| Hour-by-day-of-week effects | Yes | Yes | Yes | Yes |
| Month effects | Yes | Yes | Yes | Yes |
| Other controls | Yes | Yes | Yes | Yes |
| Mean of dep. var | 0.03 | 0.03 | 0.00 | 0.00 |
| Adjusted $R^2$ | 0.036 | 0.036 | 0.007 | 0.007 |
| N cases | 23307 | 23307 | 23307 | 23307 |

*Notes:* Controls same as basline specifications. Worker fixed effects included in columns (1) and (3) Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01

Table 2.8: Quality of care, Hospital B

| | Readmitted within 14 days | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| $\theta_i$ | 0.0177* | 0.0169* | 0.0133 |
| | (0.00894) | (0.00922) | (0.00904) |
| Specialty Dummies | Yes | Yes | Yes |
| Time Dummies | No | Yes | Yes |
| Severity Dummies | No | Yes | Yes |
| Coworker group effects | No | No | Yes |
| N | 81809 | 81809 | 68206 |
| Adjusted $R^2$ | 0.00372 | 0.00623 | 0.0378 |

*Notes:* Controls same as basline specifications. Worker fixed effects included in columns (1) and (3) Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01

Table 2.9: Heterogeneity in spillover by worker and coworker type, Hospital A

| | (1) Diff. responsiveness b/se | (2) Diff. influence b/se | (3) Interaction b/se |
|---|---|---|---|
| $\theta_{coworker}$ | 0.113*** | 0.211*** | 0.196*** |
| | (0.024) | (0.059) | (0.027) |
| $\theta_{cow} \times I\{\theta_{phys} fast\}$ | 0.129*** | | |
| | (0.033) | | |
| $\theta_{cow} \times I\{\theta_{cow} fast\}$ | | -0.096 | |
| | | (0.080) | |
| $I\{theta_{cow} fast\}$ | | -0.002 | |
| | | (0.010) | |
| $\theta_{cow} \times \theta_{phys}$ | | | -0.285* |
| | | | (0.135) |
| Hour-by-day-of-week effects | Yes | Yes | Yes |
| Month effects | Yes | Yes | Yes |
| Other controls | Yes | Yes | Yes |
| Adj. R-squared | 0.419 | 0.419 | 0.419 |
| N cases | 23307 | 23307 | 23307 |

*Notes:* Standard errors clustered at the physician level. All specifications include physician fixed effects. Other controls same as final column in previous table * p<0.10, ** p<0.05, *** p<0.01

Table 2.10: Throughput by pair type, Hospital A

| | (1) Fast/slow b/se | (2) Fast/med/slow b/se |
|---|---|---|
| fastfast2type | -0.268*** | |
| | (0.058) | |
| fastslow2type | -0.214*** | |
| | (0.066) | |
| slowfast2type | -0.045*** | |
| | (0.012) | |
| fastfast3type | | -0.372*** |
| | | (0.039) |
| fastmed3type | | -0.393*** |
| | | (0.050) |
| fastslow3type | | -0.328*** |
| | | (0.043) |
| medfast3type | | -0.249*** |
| | | (0.024) |
| medmed3type | | -0.260*** |
| | | (0.026) |
| medslow3type | | -0.188*** |
| | | (0.027) |
| slowfast3type | | -0.033* |
| | | (0.017) |
| slowmed3type | | -0.022 |
| | | (0.024) |
| Hour-by-day-of-week effects | Yes | Yes |
| Month effects | Yes | Yes |
| Other controls | Yes | Yes |
| Adj. $R^2$ | 0.401 | 0.415 |
| N cases | 23307 | 23307 |

*Notes:* Standard errors clustered at the physician level. * p<0.10, ** p<0.05, *** p<0.01

Table 2.11: Heterogeneity in spillover by worker and coworker type, Hospital B

| | (1) Diff. responsiveness | (2) Interaction |
|---|---|---|
| | Internists | |
| $\theta_{coworker}$ | 0.158*** | 0.115 |
| | (0.051) | (0.102) |
| $\theta_{cow} \times I\{\theta_{phys} fast\}$ | -0.046 | |
| | (0.095) | |
| $\theta_{cow} \times \theta_{phys}$ | | 0.218 |
| | | (0.488) |
| Adj. R-squared | 0.294 | 0.294 |
| N cases | 29052 | 29052 |
| | Pediatricians | |
| $\theta_{coworker}$ | 0.210** | 0.284*** |
| | (0.077) | (0.057) |
| $\theta_{cow} \times I\{\theta_{phys} fast\}$ | 0.042 | |
| | (0.114) | |
| $\theta_{cow} \times \theta_{phys}$ | | 1.271*** |
| | | (0.366) |
| Adj. R-squared | 0.373 | 0.373 |
| N cases | 27544 | 27544 |
| | Traumatologists | |
| $\theta_{coworker}$ | 0.330*** | 0.359*** |
| | (0.062) | (0.088) |
| $\theta_{cow} \times I\{\theta_{phys} fast\}$ | -0.068 | |
| | (0.137) | |
| $\theta_{cow} \times \theta_{phys}$ | | -0.451 |
| | | (0.735) |
| Adj. R-squared | 0.233 | 0.233 |
| N cases | 8468 | 8468 |

*Notes:* Includes Hour-by-day-of-week effects, Month effects, Physician fixed effects and all controls from previous table included. Standard errors in parenthesis, * p<0.10, ** p<0.05, *** p<0.01

Table 2.12: Heterogeneity in spillover by worker type, Hospital B

| | (1) Internists | (2) Pediatricians | (3) Traumatologists |
|---|---|---|---|
| $\bar{\theta}_{-i} \times I(fastworker)$ | 0.143 | 0.197 | 0.204 |
| | (0.121) | (0.131) | (0.126) |
| $\bar{\theta}_{-i} \times I(medianworker)$ | 0.139 | 0.210** | 0.399*** |
| | (0.098) | (0.082) | (0.046) |
| $\bar{\theta}_{-i} \times I(slowworker)$ | 0.151*** | 0.336*** | 0.249 |
| | (0.042) | (0.069) | (0.152) |
| Hour-by-day-of-week effects | Yes | Yes | Yes |
| Month effects | Yes | Yes | Yes |
| Other controls | Yes | Yes | Yes |
| Adj. $R^2$ | 0.294 | 0.373 | 0.233 |
| N cases | 29052 | 27544 | 8468 |

*Notes:* Standard errors clustered at the physician level. Physician fixed effects and all controls from previous table included. * p<0.10, ** p<0.05, *** p<0.01

Table 2.13: Spillovers among patient types, Hospital A

| | (1) Abd Pain b/se | (2) Chest Pain b/se | (3) Ped Illness b/se | (4) Resp Prob b/se | (5) Injured Extremity b/se | (6) Mult Complaints b/se |
|---|---|---|---|---|---|---|
| $\theta_{coworker}$ | 0.263*** | 0.009 | 0.232** | 0.183** | 0.082 | 0.186** |
| | (0.055) | (0.097) | (0.096) | (0.081) | (0.135) | (0.063) |
| Hour-by-day-of-week effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Month effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Other controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Adj. R-squared | 0.307 | 0.094 | 0.310 | 0.402 | 0.303 | 0.265 |
| N cases | 2836 | 1407 | 736 | 1602 | 1592 | 2409 |

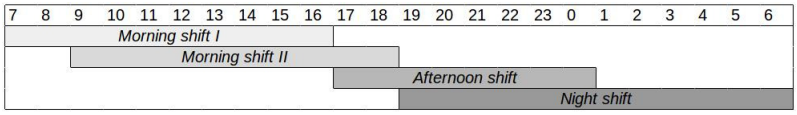
*Notes:* This table shows the result of estimating different patient-type-specific throughput for each physician and to estimate spillovers for each patient-type. Standard errors in parentheses, * p<0.10, ** $p$¡0.05, *** p<0.01

Table 2.14: Spillovers across specialties, Hospital B

| | (1) Internists b/se | (2) Pediatricians b/se | (3) Traumatologists b/se |
|---|---|---|---|
| $\bar{\theta}_{samespecialty}$ | 0.106** | 0.238*** | 0.293*** |
| | (0.047) | (0.074) | (0.061) |
| $\bar{\theta}_{otherspecialties}$ | -0.049 | 0.060 | 0.071 |
| | (0.046) | (0.070) | (0.097) |
| Hour-by-day-of-week effects | Yes | Yes | Yes |
| Month effects | Yes | Yes | Yes |
| Other controls | Yes | Yes | Yes |
| Adj. R-squared | 0.296 | 0.372 | 0.233 |
| N cases | 28424 | 27413 | 8457 |

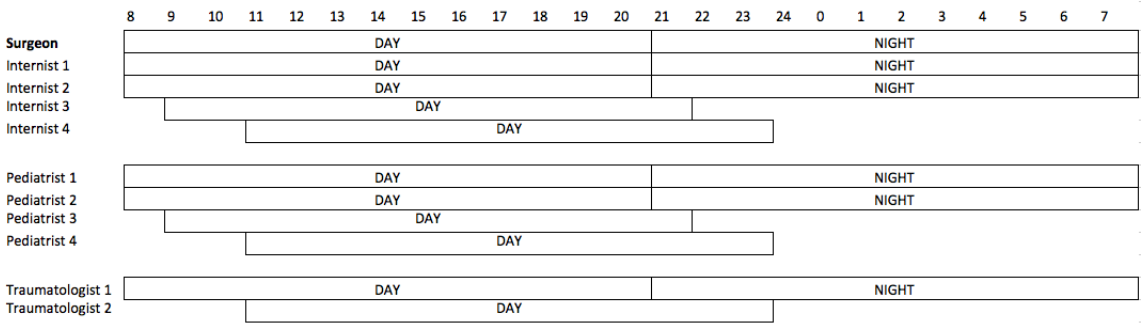
*Notes:* Standard errors clustered at the physician level. * p<0.10, ** p<0.05, *** p<0.01

Figure 2.1: Shifts in Hospital A




*Notes:* Shift ours in Hospital A

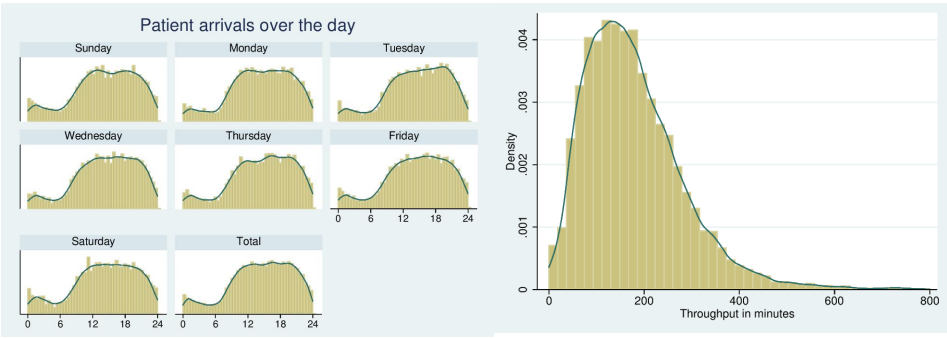
Figure 2.2: Shifts in Hospital B




*Notes:* Shift ours in Hospital B

Figure 2.3: Throughput and arrivals in Hospital A

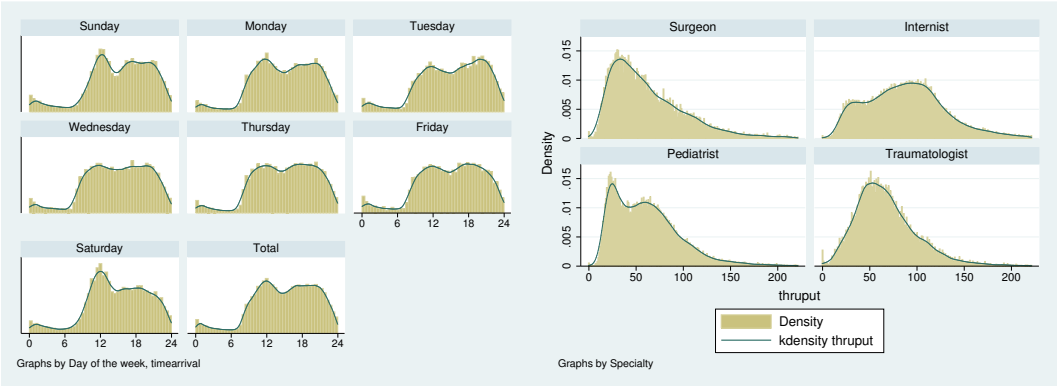
(a) Arrivals by time of the day (b) Throughput in minutes




*Notes:* Arrivals by hour of the day and throughput (in minutes) for Hospital A

Figure 2.4: Throughput and arrivals in Hospital B

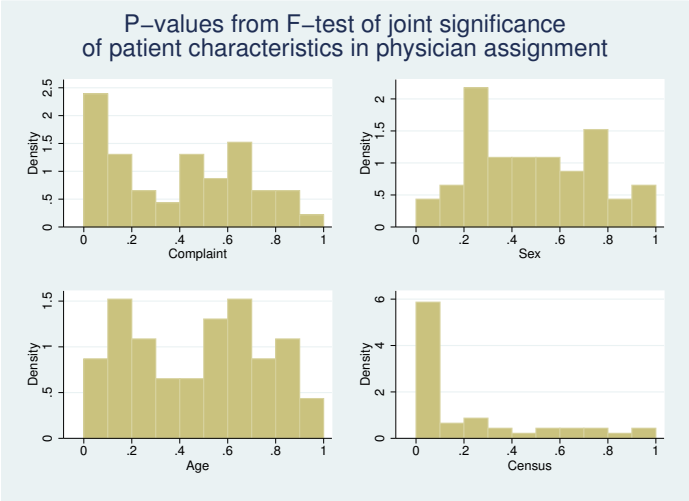
(a) Arrivals by time of the day (b) Throughput in minutes, by specialty



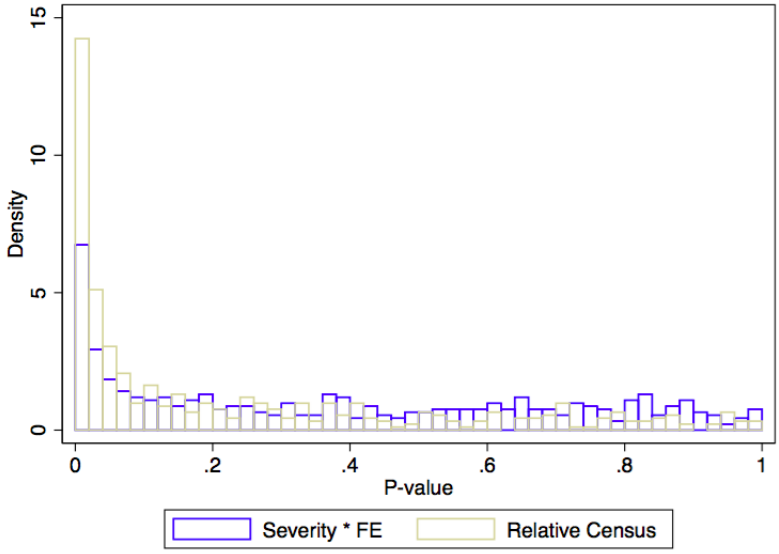

*Notes:* Arrivals by hour of the day and throughput (in minutes) by specialty for Hospital B

Figure 2.5: Explanatory variables of assignment

(a) Hospital A

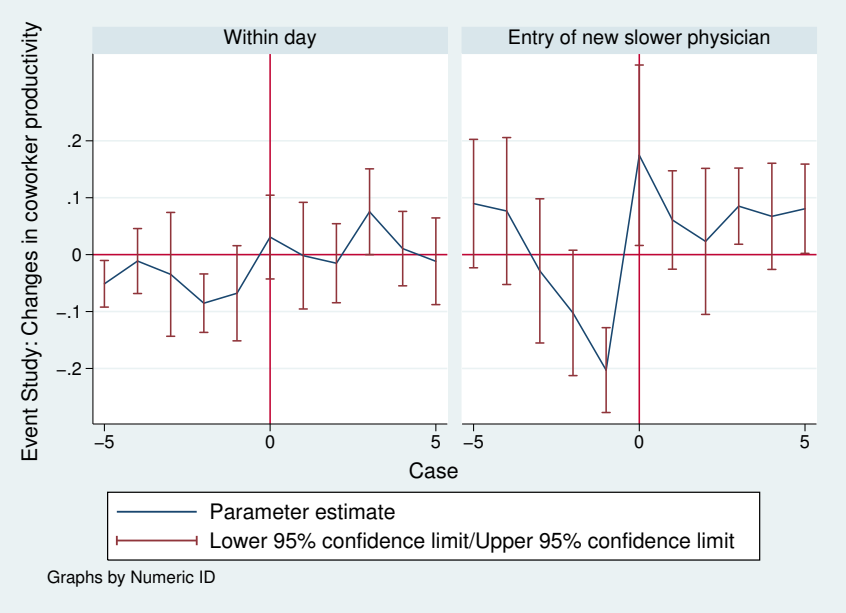



(b) Hospital B




*Notes:* Panel (a) shows stacked p-values for different variables in linear probability models for the identity of the physician assigned to the case in Hospital A. Physician pairs relative census is defined as the ratio of the number of patients under care of physician $i$ to the total number of patients under care by physician $i$ and physician $j$. Panel (b) shows a similar test for Hospital B, where we estimate multinomial logit regressions. Aside from relative caseload, is an interaction of the physicians estimated fixed effect with the triage severity rating of the patient

Figure 2.6: Event study evidence of spillovers




*Notes:* The first panel plots the basic estimates of equation 2.2, which demonstrate an immediate, persistent effect of new coworkers on a physicians productivity throughout the rest of the shift. In the second panel, we examine how this effect varies when looking only at the entry of new slower physicians to the shift

# Chapter 3

# Quality regulation and competition: Evidence from a pharmaceutical policy reform in Chile

*with José Ignacio Cuesta and Morten Sæthre*

## 3.1 Introduction

Developing economies are some of the fastest-growing markets for pharmaceutical companies nowadays. However, not much is known about their workings. These markets are characterized by relatively weak intellectual property protection and quality regulations, with potentially far-reaching consequences for the competitive environment faced by innovators and health outcomes of patients.

The most prevalent means that regulatory agencies in developed economies use to ensure the effectiveness of orally-administered generic drugs is the requirement of bioequivalence (BE). In short, a generic drug is bioequivalent to its originator counterpart – the reference drug– when its rate and extent of absorption does not show a significant difference from the rate and extent of absorption of the reference drug when administered under the same conditions (David et al., 2013).[1] BE became the primary requirement for the approval of generic drugs in the U.S. after the passage of the Waxman-Hatch Act in 1984, with the goal of simplifying the application process and fostering the entry of generics. Before 1984, the approval process for generic drugs was costly and lengthy, as it required the submission of preclinical (animal) and clinical (human) data to establish safety and effectiveness (NDA) or enough scientific literature to support the safety and efficacy of a generic drug (paper NDA) – which was generally not made available by the originator company. Currently, many

[1]Bioequivalence does not apply to topical medications, vaccines, or any other type of drugs that are not orally administered.

OECD countries either allow, encourage or require substitution of innovators for cheaper bioequivalent products (OECD, 2000).

Although BE requirements were originally implemented in the developed world to foster the entry of generics, they have been recently adopted by low and middle income countries as the primary tool for testing the effectiveness of the drugs allowed in their markets.

Quality regulations are in principle desirable in markets with asymmetric information. This is notably the case of the pharmaceutical market, since consumers generally do not have the means to determine the quality of the products at the time of purchase. However, the effectiveness of BE regulations will ultimately depend on the strategic responses of firms to them. In particular, BE requirements will only be successful if laboratories decide to invest in often expensive *in vivo* tests. However, a laboratory producing a socially desirable drug may decide not to undertake the tests if its private costs exceed its private benefits, which may decrease the availability of desirable generic options and, moreover, decrease competition.

A separate competitive effect might arise if consumers and/or doctors are at least somewhat informed about potential differences in treatment quality between generic products which are not bioequivalent and the originator product. In this case, firms might find it optimal to have different levels of quality, to reduce substitution between the firm's product and the products of other firms, which could allow more firms to be active in the market. If we view bioequivalence as imposing a minimum required quality, profits could be reduced at all quality levels, which could potentially induce exit. If exit happens due to such mechanisms, it can potentially lead to less product variety and higher prices after the regulation.

The main goal of this paper is to empirically investigate the consequences of BE requirements on market outcomes and product availability. In particular, this paper is an early exploration of the effect of imposing a BE regulation for the case of Chile, a country that recently adopted BE requirements for a list of 172 molecules, leading to the BE approval of 642 generic drugs between March 2009 and March 2015.

We combine data from the national drug registry and IMS to study how the introduction of BE requirements lead to the introduction of BE drugs in the market, and what are the ultimate consequences of the entry of BE drugs on overall market prices, and on the prices and market shares of reference, branded, and generic drugs.

First, we show strong evidence supporting the notion that BE requirements generated a substantial increase in the BE approval in this market. We document that drugs which are registered for the first time are more likely to have BE approval in the period following the announcement of the requirement — i.e., approaching the regulatory set deadline for approved BE status — as well as after the deadline has expired (even though drugs without approval to a large extent still obtain registration for marketing in all cases). We also document that drugs registered before the announcement also obtain BE in a similar pattern as first-time registrations.

Then, we turn to the data on sales and prices available in IMS for the period January 2011 to March 2015, to study how the entry of BE in the market affected prices, market shares, and competition. We study outcomes at the molecule-month level and focus primarily on (the log of) average price per gram, (the log of) sales in grams, and the number and relative

concentration of firms selling the molecule. We also disaggregate these outcomes by different types of drugs –reference, branded and unbranded– within each molecule.

Our baseline specification controls for molecule fixed effects and month fixed effects, and relies primarily on exploiting the difference in the timing of BE requirements across molecules as instruments for the introduction of BE drugs.

We do not find significant price changes for molecules facing BE entry, neither at the aggregate level, nor when looking at the reference, branded, and generic markets separately. Our results also show that the relative market share of these different types of drugs are not significantly affected by the entry of BE drugs. Finally, market concentration at the laboratory level also remains unaffected. These results are in stark contrast with the fears that BE requirements would lead to less availability of generic drugs and an increase in prices, as well as with the claims that BE drugs would increase competition and decrease the price of innovator drugs.

Our lack of effects is not uncommon to the literature analyzing the interaction between branded and generic drugs when the innovator goes off-patent, which shows disagreement on whether generic entry has an impact in affecting prices of branded drugs (see Grabowski et al., 2006 and Knittel and Huckfeldt, 2012 for two reviews of this literature).

The inconclusiveness of the empirical evidence is consistent with the ambiguity of the theoretical predictions. Although branded drugs face higher degree of competition after patent expiration, they also target a more inelastic part of the demand curve (Frank and Salkever, 1992). Moreover, empirical evidence for the U.S. shows that the potential effects of increased competition following generic entry are partially offset by changes in marketing and advertisement (Caves et al., 1991; Lichtenberg and Duflos, 2009; Lakdawalla, Philipson, and Wang, 2006; Knittel and Huckfeldt, 2012). On the other hand, BE requirements increase the production cost of generics, which may be passed on to consumers in the form of higher prices.

This paper is closely related to Balmaceda, Espinoza, and Diaz (2015), which is, to our knowledge, the only empirical evaluation of the BE requirements in Chile. The authors use a a differences-in-differences strategy to estimate a reduced-form effect of the BE requirements on drug prices, and find heterogeneous effects across different drugs, even within the same molecule. In this paper we complement their evidence in several ways: First, our main dependent variable is the number of BE drugs with the corresponding molecule in the market, which is instrumented by the deadlines of the BE requirements. Also, we look at overall market effects as well as the effect on different subsegments within the same molecule. This strategy provides interpretable coefficient in terms of market competition. Also, instead or relying on parallel-trend assumptions, our identifying assumption is that the timing and deadlines of the BE requirements are not correlated with price trends.

## 3.2 Pharmaceutical market and quality regulation in Chile

### Institutional framework

Compared to OECD standards, chileans spend a relatively low share (0.9 percent) of their GDP on pharmaceuticals (OECD, 2013). However, pharmaceutical spending accounts for more than half of all out-of-pocket health expenditures in the country (Cid and Prieto, 2012).

Overall, survey evidence shows that 33.4% of individuals paid their prescription drugs fully out-of-pocket (Ministerio de Salud, 2013). The level of financial coverage for prescription drugs depends mostly on whether the individual opts to enroll in the public insurance system (FONASA) or to buy a health insurance plan in the private sector, and on the specific disease to be treated.[2] FONASA enrollees who opt to receive health care within the network of public providers face copayment rates that depend on socio-economic variables, although outpatient claims are free of charge, including prescription drugs.[3] FONASA enrollees who instead opt for receiving care in a private hospitals pay procedure-specific prices negotiated between FONASA and each provider.[4]

Insurance plans in the private system do not generally include coverage for prescription drugs.

Three large pharmacy chains account for more 90% of the chilean pharmaceutical retail market (Ministerio de Salud, 2013). Unlike in the US, direct advertisement of prescription drugs is prohibited in Chile, although there is ample evidence of a strong role of drug representatives from laboratories marketing their products with doctors. Also, laboratories and pharmacies have been found to provide (illegal) incentives to the retail sales agents. These chains are vertically integrated with laboratories, and a fraction of their sales correspond to own-brand drugs.[5].

### Bioequivalence regulation in Chile

Bioequivalence is a standard request for drug commercialization in most high income countries (Balmaceda, Espinoza, and Diaz, 2015). BE is established in order to demonstrate therapeutic equivalence between the generic (test) drug product and corresponding reference drug. In particular, two products are considered bioequivalent when the rate and extent of absorption of the test drug do not show a significant difference from the rate and extent of absorption of the reference drug when administered at the same molar dose of the therapeutic ingredient under similar experimental conditions (David et al., 2013). According

[2]For a more detailed description of the health insurance market in Chile, see Duarte (2012).

[3]For a set of 80 prioritized diseases, the total level of copayment is capped

[4]With the exception of the pharmacological treatment of a list of 8 'high-cost' diseases, for which there is full coverage.

[5]see `http://tinyurl.com/jjggenz`, `http://tinyurl.com/jxlu698` and `http://tinyurl.com/hq5xq5e`

to the FDA, therapeutically equivalent drugs can be substituted with the full expectation that the substituted product will produce the same safety effect and safety profile as the originally prescribed (reference) product (Food and Drug Administration, 2016).

In Chile, BE requirements were put in place because of the low perceived quality of generic drugs. The stated goal of the BE regulation was to increase generic quality, increase competition, and reduce prices. Before the BE requirements, quality standards required following guidelines of International Pharmacopeia books, which do not ensure therapeutic efficiency.

Generally, BE is determined through *in vivo* clinical studies for one specific presentation of a given drug, and then *in vitro* studies are performed for other presentations of the same drug. BE approvals of imported drugs are generally validated in Chile if they were done in places considered by the Public Health Institute (*Instituto de Salud Pública*, ISP) to have high quality certifications, like Canada, the U.S. and Europe, among others. BE is given *ad eternum* for a given formula and production technology. A change in any of these dimensions will imply that the manufacturer will require a new proof of bioequivalence.

In 2005, the Chilean Ministry of Health (MINSAL) published a list of molecules that were subject to BE. However, it wasn't until 2009 that the ISP established the technical norms for the realization of these studies (Balmaceda, Espinoza, and Diaz, 2015).

BE requirements have rolled out step-wise since then: requirements for different groups of molecules have been established at different points in time through a set of law decrees. After the passage of each decree, all new drugs containing the corresponding molecules need to show proof of BE before been awarded a sales permit by the ISP. Each decree also specifies the deadline for BE testing among those drugs that are already registered with the ISP. However, the slow uptake and capacity constraints of laboratories forced the regulator to extend the original deadlines in all decrees. For instance, the first list of 12 molecules was published in June 2009. The deadline for proof of BE among registered drugs was set to July 2009. However, the deadline was extended twice, and was finally set for December 2013. In section 3.3, we describe these for all molecules in more detail.

Still, enforcement of the deadlines is weak, since it is determined by limited enforcement capabilities of the ISP. According to industry experts, the low auditing capabilities of ISP make deadlines not being fully binding. However, BE requirements do become binding every time a drug has to renew its registry with the ISP, which happens every 5 years. Therefore, in practical terms, the BE requirement binds, for the most part, at the time when the registry expires.

## 3.3 Data and descriptive statistics

### Data sources

We use three main sources of data. First, we the universe of registered drugs in Chile for all molecules under BE requirements.[6] The data used in this version of the paper corresponds to all registered drugs by May 2015. For each registered drug, the dataset provides information on manufacturer (laboratory), date when drug was first licensed in Chile, date of last license approval and date of next license approval.[7] It also includes information on the drug dosage, its presentation (e.g., tablet, capsule, and injectable, among others), and its marketing status (prescription, over-the-counter or discontinued). We combine this data with the list of drugs that obtained BE approval up to May 2015. This list contains 642 drugs together with the date at which they obtained BE approval, the corresponding reference product, and which treatment the drug is intended for.

Finally, we use data from IMS Chile, which contains detailed information on monthly prices and sales for drugs in Chile for the period between January 2011 to March 2015.[8] IMS Chile collects data from wholesale transactions between distributors and pharmacies.[9] The dataset reports variables aggregated across small groups of counties.[10] In terms of products, the dataset provides information for each product available from each laboratory in each of its available dosage and presentations for all reference and branded drugs. For unbranded drugs, however, the dataset aggregates the information at the dosage and presentation level and thus does not identifies laboratories.

A key aspect in constructing the dataset we use in this paper was to be able to match products in the IMS dataset to license and BE information in ISP datasets. We were able to match 95% of reference and branded drugs in the IMS data to a license registry in ISP.[11] This allows us to provide a precise picture of the evolution of the incidence of BE for each molecule in our sample.

---

[6] These data are available at `http://registrosanitario.ispch.gob.cl/`.

[7] The renewal data corresponds to exactly 5 years after the date of last license approval, or 5 years after the date first licensed if the drug was first licensed less than 5 years ago.

[8] We were granted confidential access to this data as part of a research agreement with *Servicio Nacional del Consumidor* (SERNAC).

[9] Drug prices are corrected in two dimensions. First, we adjust them for inflation and measured in 2013 Chilean pesos. The adjustment uses a measure of health CPI available at the National Institute of Statistics (*Instituto Nacional de Estadística*, INE). Second, we adjust for the fact that different presentations have different amounts of the drug. This adjustment is done by calculating prices per gram of the product.

[10] This version of the paper restricts the sample to a limited set of counties where the quality of the data is the highest. Concretely, these counties are Huechuraba, Colina, Quilicura and Renca, all of which are located in the capital city, Santiago. These counties represent 8% of the population in the region, and have an average poverty rate of 10.85%, close to the region's average of 11.5%. Future versions of the paper will include additional counties in the analysis.

[11] Since different package sizes of the same drug do not require different license registry, the same registry may correspond to different products in the IMS data

## Defining product types

Throughout the rest of the paper, define different *types* of drugs in the data. Concretely, we distinguish between reference, branded and unbranded products. Reference products are the innovator product within each molecule. Branded products (often referred as 'branded generics') are non-innovators that adopt a fantasy name, and are often packaged in ways visually as attractive as innovator drugs. For the most part, bioequivalent drugs correspond to unbranded generics that get (or enter the market with) bioequivalence approval. Finally, unbranded drugs (or 'unbranded generics') are generic drugs that are sold under the name of the molecule.

## Descriptive statistics on drug licensing

Before we assess the effect of the regulation on market outcomes, we document how the reform affected the number of bioequivalent and non-bioequivalent product specifications in the market. As mentioned, the BE requirements were announced at different times for different molecules, with different deadlines applying. Furthermore, the deadlines were often extended, particularly for the molecules which were planned to be put under regulation at earlier dates.

These differences in timing of the reform are documented in Table 3.1. The table shows that molecules can be classified into 7 different groups based on the timing of their first BE requirement and subsequent deadline extensions. For instance, group 1 corresponds to the set of molecules that have the first BE requirement announced in January 28 2011, and deadline in February 16 of the same year. All drugs in this group had their deadline extended twice, to July 6, 2012 and then to December 31, 2013, through decrees announced in June 1, 2012 and June 7, 2013, respectively. Groups 2-7 of molecules all had different combinations of initial announcements and subsequent extensions.

In Table 3.2, we report the number of unique molecules, defined by ATC (Anatomical Therapeutic Chemical) codes, and number of unique specifications which have been registered for marketing for each of these 7 groups. In addition, we report the number of drugs with BE approval as of March 2015, as well as the number of drugs which have voluntarily suspended their marketing license at that time. For instance, group 1 is comprised by 28 different molecules and 749 unique drugs. By March 2015, 189 had BE approval and 120 had suspended their marketing license.

In the following, we take the announcement of the reform as being given by the first announcement for the molecule of the drug, as shown in Table 3.1, while the deadline is taken to be the last deadline reported.

In Figure 3.1, we show the number of specifications which register for the first time—either obtaining or not obtaining BE approval—in time windows around the date of announcement and the date of the deadline. The time scale is in days relative to these events (i.e., normalized to zero at the time of the event). The drugs included in these graphs is the

set of drugs registering for a first-time marketing license at most one year before announcement and later.

The bars show the actual number of registering specifications in each group (bioequivalent and non-bioequivalent), such that the total number of specifications registering in a given time window is given by the sum of the height of the two overlapping bars. In both panels, it is apparent that the majority of drugs which are registered did not get BE approval within the time period of our data. From the upper panel, we see that the proportion of first-time registering drugs becoming bioequivalent increases sharply after the announcement. The marked drop in the total number of registered drugs towards the end of the period after announcement is due to reaching the end of our sample window for many of the drugs, the announcement of the regulation happens much later than for other drugs. From the bottom panel, we see that the proportion of bioequivalent drugs increases as we approach the deadline, and also after the deadline has passed – though this is not obvious just from inspection of the graph.

In Figure 3.2, we show the timing of BE approval against the time of entry, relative to the announcement of the reform (in the left-hand panel) and the deadline (in the right-hand panel), respectively. This figure only covers drugs which obtained approval of bioequivalence at some point before March 2015. Each dot represents a BE approval, and their placement in the x-axis, relative to the dashed line, corresponds to the time difference between their registration with the ISP and the (first) announcement of BE requirement for their molecule. Dots to the the right of the dashed line correspond to drugs that register for a marketing license for the first time after the BE requirement was put in place. Approvals along the 45 degree line correspond to drugs that register at the same time as they provide documentation of bioequivalence to the ISP. Entries above the 45 degree line correspond to approvals which happens after the drug is registered with the ISP. The vertical distance between the 45 degree line and each dot is therefore the time-lag between entry and approval.

In the left-hand panel, we see that there is initially a substantial number of drugs obtaining approval with a noticeable time-lag compared to their entry date, while for drugs entering more than 1.5 years after announcement, virtually all obtain approval simultaneously with their marketing license. In the right-hand panel, which compares the entry date and approval date of each drug with the date of the deadline, there's no obvious pattern in the number of drugs or length of period between entry and approval as one approaches the deadline. On the other hand, after the deadline has passed all drugs which at some point in our sample obtain bioequivalence approval get this at the same time as they enter.

Analogously to the previous plot, Figure 3.3 shows number of BE approvals within time windows around announcement and deadline, for drugs registered at least 1 year earlier than the announcement – i.e., the drugs which first registered for a marketing license earlier than the drugs covered in Figures 3.2. From the upper panel, we see that the announcement does not immediately lead to existing specifications obtaining BE approval, though the number increases sharply about 1.5 years after the announcement – similar to the pattern found for first-time registering drugs. The bottom panel suggests that approaching and passing the deadline is more strongly related to a higher number of BE approvals for existing

specifications.

### Descriptive statistics on market outcomes

By matching the IMS dataset to information ISP sources, we constructed a balanced monthly panel dataset for the period between January, 2011 and March, 2015. The resulting dataset covers 181 molecules, including all molecules are subject to BE requirements within our sample period.[12] The dataset contains 3,018 unique products, defined as a unique combination of product name, dosage, and presentation. These products are provided by 79 different laboratories.

Importantly, not all products in the panel are sold every period. For instance, 59.3% of the products register positive sales across the sample and 83.9% are considered to be *active*.[13] Monthly prices are recorded for all products in periods with positive sales, although 9.4 % of prices are missing.[14]

Table 3.3 displays basic descriptives statistics for the IMS data. On average, reference products are priced at around twice the mean product in the market, while the price of branded products is close to the mean price and unbranded products are remarkably cheaper. The highest market share is captured by branded drugs, with an average market share of 56 %, followed by generics with a market share of 25 % and reference drugs with 19%. The remaining 2% corresponds to the market share of bioequivalents. However, the market share of BE has increased substantially with the introduction of BE requirements, from only 0.2% in 2011, to 6.7% in 2015 – an increase of more than 30 times (not shown in the table).

The average market has almost 14 products and 5 laboratories actively participating in it. As expected, the number of products and laboratories is remarkably larger in the branded segment than in the reference and the unbranded segment.[15]

## 3.4 The effects of quality regulation on market outcomes

In this section, we study the effects of bioequivalence on a variety of market outcomes. Following Duggan, Garthwaite, and Goyal (2016), we treat each molecule as a different market. We estimate the effects of the entry of bioequivalent products to markets on a range of outcomes, including prices and market shares, among others.

[12] We also include the set of molecules that Balmaceda, Espinoza, and Diaz (2015) use as a control group, although our main specification does not consider these in the estimation

[13] We consider a product as being active in a given period if it register sales in any period before and in any period after it.

[14] Although we observe prices for all products with positive sales, we are not always able to measure the amount of grams in a product, which implies that we cannot measure price per gram in those cases.

[15] This comes partly from the data limitations in terms of identifying unbranded products.

For this empirical application, we exploit the panel structure of the data to control for permanent differences across markets and for common shocks to all markets using fixed effects. Moreover, building on our results in section 3.3 we utilize the differential timing of the reform rollout in order to propose an instrumental variables strategy with which to deal with the potential endogeneity of market-level incidence of bioequivalence, our treatment variable.

Our first outcome of interest is the evolution of market prices. We expect the introduction of BE drugs in a given market to affect prices in several ways. First, there is a direct (or 'mechanical') effect of introducing a new drug in the average market price, by changing the composition of the market. On the other hand, the introduction of BE drugs potentially increases competition with branded and generics, with consequences in prices and market shares of competitors.

In order to shed light on the effects along those different margins, the following section presents a decomposition of the evolution of aggregate market prices of a given molecule, into components that reflect price increases holding shares constant, and changes in market shares, including the impact of entry and exit of drugs.[16]

## A decomposition of prices

Let the market price $P_{mt}$ be defined as a weighted average of the prices of all products in the market, where weights $w_{it}$ are given by the market share of the products in the market:

$$P_{mt} \equiv \sum_i w_{it} P_{it}$$

Each product $i$ in the market can be classified in one of four possible product types, denoted by $k \in \mathcal{K} = \{\text{reference, branded, bioequivalent, unbranded}\}$. We can then rewrite $P_{mt}$ as:

$$\begin{aligned} P_{mt} &= \sum_k \sum_{i \in k} w_{it} P_{it} \\ &= \sum_k w_{kt} \sum_{i \in k} \tilde{w}_{it} P_{it} \\ &\equiv \sum_k w_{kt} \tilde{P}_{kt} \end{aligned}$$

where $w_{kt}$ is simply the market share of segment $k$, defined as $w_{kt} \equiv \frac{\sum_{ik} sales_{it}}{\sum isales_{it}}$ and $\tilde{w}_{it} = \frac{w_{it}}{w_{kt}}$, such that $\tilde{P}_{kt}$ is simply a weighted average of product prices in segment $k$.

[16] Similar decompositions have been extensively used in the literature of productivity dynamics using plant-level data. See for instance Foster, Haltiwanger, and Krizan (2011).

Denote by $\Delta P_{mt} \equiv P_{mt+1} - P_{mt}$ the change in market price between $t$ and $t+1$. We can write this change as:

$$\begin{aligned}\Delta P_{mt} &= \left[\sum_{k\in entry} w_{kt+1}\left(\tilde{P}_{kt+1} - \bar{P}_{mt}\right) - \sum_{k\in exit} w_{kt}\left(\tilde{P}_{kt} - \bar{P}_{mt}\right)\right] \\ &+ \sum_{k\in stay} \Delta w_{kt}\left(\tilde{P}_{kt} - \bar{P}_{mt}\right) + \sum_{k\in stay} w_{kt}\Delta\tilde{P}_{kt} + \sum_{k\in stay}\Delta w_{kt}\Delta\tilde{P}_{kt} \\ &\equiv \Delta P^{EX}_{mt} + \Delta P^{RW}_{mt} + \Delta P^{PC}_{mt} + \Delta P^{CS}_{mt}\end{aligned}$$

where $\bar{P}_{mt}$ indicates the average of the market price over $t$ and $t+1$. This decomposition separates the change in the average market price in four additive components. The first component, $\Delta P^{EX}_{mt}$, captures price changes at the market level produced by the entry and exit of a given type in the market. For instance, this component will be different than zero when BE drugs enter the market, in which case $\Delta P^{EX}_{mt}$ will be equal to average price of BE at the time of entry (relative to the average market price), times the share of sales in the market corresponding to BE drugs. The second component, $\Delta P^{RW}_{mt}$, measures the contribution of changes in the market shares of different types on the change in the average market price. This component is positive when relatively expensive drugs increase their market share. The third component, $\Delta P^{PC}_{mt}$, measures the contribution of the actual price changes of each type to the change in the average market price. Finally, the last component $\Delta P^{CS}_{mt}$ captures the correlation between the changes in prices and the changes in market shares. For instance, this term will be negative whenever types that have a price increase betwen periods are also the types that loose market share between periods.

Adding up period-by-period price changes between the initial period ($t = 0$), and an arbitrary period $t$, the market price at period $t$ can be written as a function of the initial market price $P_{m0}$ and all future (decomposed) price changes, as:

$$P_{mt} = P_{m0} + \sum_{s=0}^{t-1}\Delta P^{EX}_{ms} + \sum_{s=0}^{t-1}\Delta P^{RW}_{ms} + \sum_{s=0}^{t-1}\Delta P^{PC}_{ms} + \sum_{s=0}^{t-1}\Delta P^{CS}_{ms} \tag{3.1}$$

To understand the price evolution at the micro-level within each type, we further decompose the change in average segment prices $\Delta\tilde{P}_{kt}$ in a similar way into four components, as follows:

$$\begin{aligned}\Delta\tilde{P}_{kt} &= \left[\sum_{i\in k,entry} \tilde{w}_{it+1}\left(P_{it+1} - \bar{P}_{kt}\right) - \sum_{i\in k,exit} \tilde{w}_{it}\left(P_{it} - \bar{P}_{kt}\right)\right] \\ &+ \sum_{i\in k,stay} \Delta\tilde{w}_{it}\left(P_{it} - \bar{P}_{kt}\right) + \sum_{i\in k,stay} \tilde{w}_{it}\Delta P_{it} + \sum_{i\in k,stay}\Delta\tilde{w}_{it}\Delta P_{it} \\ &\equiv \Delta\tilde{P}^{EX}_{kt} + \Delta\tilde{P}^{RW}_{kt} + \Delta\tilde{P}^{PC}_{kt} + +\Delta\tilde{P}^{CS}_{kt}\end{aligned}$$

Using this decomposition, we can rewrite the average price for type $k$ at any period $t$ as:

$$\tilde{P}_{kt} = \tilde{P}_{k0} + \sum_{s=0}^{t-1} \Delta\tilde{P}_{ks}^{EX} + \sum_{s=0}^{t-1} \Delta\tilde{P}_{ks}^{RW} + \sum_{s=0}^{t-1} \Delta\tilde{P}_{ks}^{PC} + \sum_{s=0}^{t-1} \Delta\tilde{P}_{ks}^{CS} \tag{3.2}$$

where each term has the same interpretation as in equation 3.1, but in terms of the contribution of each component to average segment prices.

Below, we illustrate how this decomposition operates in practice and the insights it provides using a case utilize a case study. Next, we introduce our empirical strategy with which we attempt to formally measure the role of the entry of bioequivalents in explaining the different components of price evolutions for all markets.

**An example decomposition: the case of Metformin**

In this section, we illustrate the workings of the proposed decompositions for the case of Metformin, the molecule with the highest revenue in our sample.[17] In January 2011, 75% of total revenue in this market corresponded to reference drugs, while branded and generics accounted for 19% and 6% of the revenue respectively. Figure 3.4 shows the evolution of $P_m$ as well as $P_k$ for each of the four type for drugs containing this molecule.

In this particular market, bioequivalent drugs entered at the end of 2012. The entry of bioequivalents changes, by definition, the composition of this market, with potential immediate consequences on the average price. However, as seen in Figure 3.4, bioequivalents enter at a price that is similar to the average price in the market, and therefore the average price does not change substantially at the entry of the BE type. Still, the figure also highlights that BE products potentially played a direct role in decreasing the average price in the months following their first entry, as their average price falls below the price average price. Finally, the figure highlights substantial heterogeneity both in the level and evolution of prices across different groups. Although the average price in this market increases by only 3.4% during the period, branded drugs had a 80% increase, while generics had a 25% decrease.

Figure 3.5 shows the result of the decomposition of equation 3.1 for Metformin. The line denoted by $P$ reproduces the overall average price of Figure 3.4, and corresponds to the average market price $P_{mt}$.

The decomposition provides insights to explain the 3% price increase in the average price of this molecule. First, $P_{PC}$ is almost always positive and upward sloping, meaning that incumbent types generally increase their prices in the period. On the other hand the component $P_{RW}$ is large and negative starting in mid 2013, which means that relatively expensive types (reference and/or branded) loose substantial market share in the period. $P_{EX}$ is only different than zero starting in September 2012, when BE drugs enter this market for the first time. Although $P_{EX} > 0$, its magnitude is very small, reflecting that the price of BE drugs at the time of their introduction is not much larger than the average market

[17] Metformin is a drug mainly used to treat patients with insuline resistance and type-2 diabetes

price, and that their market share at their introduction is not large enough to substantially affect the average price.

Figure 3.6 shows the result of the decomposition in equation 3.2 for the four types of drugs in the the Metformin market. The series denoted by $P$ in each panel represent $P_k$, the average price of each type, and reproduce those shown in Figure 3.4. These figures allow to shed light on how the decomposition suggested in equation 3.2 helps explaining the patterns behind the average price increase in each type. For instance, Panel (a) shows that the price increase among reference drugs is explained mostly by a price increase among incumbent drugs and also by an increasing market share of the most expensive drugs in the market. On the other hand, Panel (b) shows that the main reason behind the strong price increase of branded drugs in this market is the increased market share of the most expensive drugs, as well as an increase in the price of each drug. Moreover, changes in market shares are negatively correlated with price changes over the period, which contributes to a cumulative price decrease of more than 20 percent. As expected, Panel (c) highlights that while new BE drugs enter this market, most of the changes in prices within the BE type are due to changes in composition rather than changes in the prices of incumbent drugs. The opposite is true starting at the beginning of 2014, when the number of BE drugs in this market stabilizes to 14 drugs. Finally, Panel (d) shows that most of the changes within the unbranded segment are caused by changes in the prices of incumbent drugs rather than changes in the composition of this type.

## Empirical strategy

This section describes the econometric approach we use to study the effects of BE regulation on prices and other market-level outcomes.

For a given market-level outcome $y_{mt}$ in market $m$ at month $t$, the main specification we study is:

$$y_{mt} = \beta BE_{mt} + X'_{mt}\gamma + \alpha_m + \delta_t + \varepsilon_{mt} \tag{3.3}$$

where $BE_{mt}$ is a measure of bioequivalence incidence. The coefficient of interest is $\beta$, which measures the effect of bioequivalence on the outcomes of interest. Additionally, we include a vector of market level control variables, $X_{mt}$, and two sets of fixed effects: $\alpha_m$ are market fixed effects that control for permanent differences across markets that may be correlated with $BE_{mt}$, and $\delta_t$ are time fixed effects that control for shocks common to all markets in a given period of time. The identifying assumption in this specification would be that there are no market-level trends that could be driving the entry of BE products across markets.

We are mostly worried that bioequivalence incidence could be correlated with unobservable shocks to market level outcomes. This would be the case, for instance, if laboratories decide to perform BE tests based on future market unobservables that increase prices, biasing the OLS estimate of $\beta$ on equation 3.3 upwards.

In order to address this potential endogeneity, we propose an instrumental variable approach. Section 3.3 provided abundant evidence pointing towards the timing of the policy

rollout being a key driver of entry of BE products in the market. Importantly, the timing of the policy differs across markets. We argue that such feature provides exogenous variation across markets and time that induced increases in market-level BE incidence, and therefore, that can be used to construct instrumental variables.

In practice, we define a vector of instrumental variables $Z_{mt}$ that includes dummies indicating the periods after the first and last decree and corresponding deadline for each market, with corresponding first stage equation given by:

$$BE_{mt} = Z'_{mt}\eta + X'_{mt}\gamma + \alpha_m + \delta_t + \varepsilon_{mt} \tag{3.4}$$

The main exclusion restriction behind this strategy is that decrees and deadlines for a given molecule were not set as a function of price shocks not captured by molecule fixed effects and price effects. A violation to this assumption would happen if, for instance, decrees set earlier deadlines for molecules that were expected to have earlier price increases. Although we cannot directly test this hypothesis, the fact that decrees were set and modified mostly because of capacity constraints of laboratories testing BE makes it unlikely that deadlines were timed as a function of expected price increased of specific molecules.

The Table 3.4 shows the result of the first-stage for different subsets of instruments available from the decrees. Consistent with the previous sections, the corresponding $F$ statistic shows that the instruments are strongly relevant.

## Effects on prices

We start by discussing the results of estimating equation 3.3 on average prices and the different components of equation 3.1. Specifically, we estimate equation 3.3 for five different dependent variables: average market price $P_{mt}$, and the four components of the changes in average market price from the decomposition in equation 3.1, $\sum_{s=0}^{t-1}\Delta P_{ms}^{EX}$, $\sum_{s=0}^{t-1}\Delta P_{ms}^{RW}$, $\sum_{s=0}^{t-1}\Delta P_{ms}^{PC}$ and $\sum_{s=0}^{t-1}\Delta P_{ms}^{CS}$.

The results in column (1) show the OLS results. They indicate that one additional bioequivalent product in the market is associated with a 1.7 percent increase in the market price $P_{mt}$. This increase is mostly the result of a 2.7 increase of the average price of incumbent markets, combined with a 1.5 decrease in the effect on market shares, as shown by the coefficent in Panel B associated with $\sum_{s=0}^{t-1}\Delta P_{ms}^{PC}$ and $\sum_{s=0}^{t-1}\Delta P_{ms}^{RW}$, respectively. On the other hand, the effect on the net-entry term and in the cross-correlation term are an order of magnitude smaller. These results are consistent with the overall pattern seen for Metformin in Figure 3.5.

Column (2) and (3) we show the coefficients of the instrumental-variable regressions that exploit the time-variation of BE decrees, as discussed in the previous section. In column (2), we use a dummy variable equal to 1 for dates after the last decree associated with the drug, and after the last deadline set by such decree. In column (3), we also include a dummy for dates after the first decree and after the first deadline. The IV results do not show

any significant effect of the number of BE drugs on the average price and on the different components of its evolution.

In Table 3.6 we show the results of estimating equation 3.3 on average prices and the different components of equation 3.2, for each type of competitors to BE drugs: reference, branded, and unbranded. Although the results highlight that the effects of BE are potentially heterogeneous across the different components of the decomposition, the effects are not precisely estimated, and overall indicate that the entry of BE drugs have not had a significant effect overall prices across the different type of drugs, and on the different components of its evolution.

### Effects on market shares, number of products and competition

In Table 3.7 we show results for log of market share of each group defined previously: reference, branded, unbranded and bioequivalent. We only find a significant increase in the share of bioequivalent products. The point estimate shows that one more bioequivalent introduced in a market increases the bioequivalent market share between 1.2 and 1.6 percent. However, we are not able to detect any significant increase in the market share of any other product category, pointing towards the fact that bioequivalent products did not substitute clearly a particular type of drug. This result is also aligned with those found in Table 3.8, where we study the number of products and laboratories present in the market and in each type of group. We find that an extra bioequivalent product corresponds to increase of 10-22 percent in this market. Although the coefficients point towards a slight but significant decrease in the total number of products, we do not find this effect to be particularly concentrated in any other group.

Finally, we study whether there is change in the concentration of laboratories following the introduction of bioequivalent drugs. Table 3.9 shows the effects on an extra bioequivalent drug on the Herfindahl-Hirschman Index (HHI). The first panel calculates the HHI at the laboratory level, and the second at the conglomerate level. We do not find a significant effect of the introduction of bioequivalent drugs on these outcomes.

## 3.5 Discussion

In the last decade, Chile adopted bioequivalence requirements as the primary tool for regulating the quality of generic drugs allowed in the market. Since then, and up to March 2015, more than 600 generic drugs gained bioequivalence approval.

Although there were significant fears that bioequivalence requirements would lead to less availability of generic drugs and to an increase in prices, in this paper we do not find significant evidence to support those claims. Looking at average market prices and also at prices for reference, branded, and generic drugs, we find that the entry of bioequivalent drugs did not have a significant effect on prices. Moreover, our analysis shows that the market

share of these different types of drugs, as well as the market concentration of laboratories, were not been significantly altered by the introduction of bioequivalents.

The lack of strong effects of bioequivalent entry on market prices suggest that cheaper drugs are not disproportionally loosing their marketing license because of the requirements. On the other hand, they suggest also that innovator drugs have not decreased their price in reaction to the competitive pressure from bioequivalent drugs. This latter result is in line with previous evidence in other contexts suggesting empirically small (and theoretically ambiguous) price effects on branded drugs from the entry of generics after patent expiration.

It is important to highlight that this paper is an early attempt to study the impacts of bioequivalence in the Chilean market, and within the time period of this study bioequivalents reached a peak of only 6.5 percent of market share. We expect potentially stronger impacts on prices and other outcomes as more bioequivalent drugs enter. In future research we plan to extend our period of analysis to include several hundred new bioequivalence approvals not covered in this study.

Table 3.1: Timing of reform: Announcements and deadlines

| | 1st decree | | 2nd decree | | 3rd decree | |
|---|---|---|---|---|---|---|
| N | Announced | Deadline | Announced | Deadline | Announced | Deadline |
| 1 | 2011-01-28 | 2011-02-16 | 2012-06-01 | 2012-07-06 | 2013-06-07 | 2013-12-31 |
| 2 | 2011-02-28 | 2011-01-31 | 2012-06-01 | 2012-07-06 | 2013-06-07 | 2013-12-31 |
| 3 | 2012-10-24 | 2013-10-24 | 2013-10-25 | 2014-04-30 | | |
| 4 | 2012-12-24 | 2013-01-31 | 2013-06-07 | 2013-12-31 | | |
| 5 | 2012-12-24 | 2013-12-31 | | | | |
| 6 | 2012-12-24 | 2014-12-31 | 2014-12-26 | 2015-12-31 | | |
| 7 | 2014-02-24 | 2015-12-31 | | | | |

*Notes:* This table displays the dates of announcement and deadlines of bioequivalence requirements for different groups of molecules. Each group includes molecules subject to the same deadlines for bioequivalence approval

Table 3.2: Timing of reform: Number of affected products

| N | # ATC | # drugs | # Bio.eq. | # suspended |
|---|---|---|---|---|
| 1 | 28 | 749 | 189 | 120 |
| 2 | 15 | 619 | 51 | 85 |
| 3 | 12 | 611 | 53 | 209 |
| 4 | 21 | 630 | 101 | 179 |
| 5 | 30 | 1427 | 69 | 384 |
| 6 | 49 | 1146 | 104 | 217 |
| 7 | 17 | 355 | 18 | 35 |

*Notes:* This table quantifies the number of affected products by in each group of molecules affected by the bioequivalent decrees. Groups 1-7 are defined in table 3.1 and correspond to groups of molecules subject to the same deadlines for bioequivalence approval. #ATC corresponds to the number of molecules in each product. # of drugs corresponds to the number of different products. # Bio.eq. is the number of products with bioequivalence approval by March 2015 and # suspended is the number of products with suspended marketing license by March 2015.

Table 3.3: Descriptive statistics for IMS data

| Variable | N | Mean | SD | p10 | p50 | p90 |
|---|---|---|---|---|---|---|
| *Price per gram* | | | | | | |
| All products | 82,756 | 47.37 | 139.78 | 0.30 | 6.84 | 98.51 |
| Reference | 11,045 | 93.06 | 224.13 | 0.69 | 15.18 | 216.19 |
| Branded | 62,083 | 44.89 | 124.70 | 0.52 | 8.40 | 98.08 |
| Unbranded | 9,628 | 10.97 | 79.22 | 0.07 | 0.56 | 9.11 |
| Bioequivalent | 3,006 | 63.20 | 198.34 | 0.20 | 9.03 | 114.60 |
| *Market shares* | | | | | | |
| Reference | 8,808 | 0.19 | 0.27 | 0.00 | 0.05 | 0.59 |
| Branded | 8,808 | 0.56 | 0.36 | 0.00 | 0.64 | 1.00 |
| Unbranded | 8,808 | 0.25 | 0.35 | 0.00 | 0.04 | 1.00 |
| Bioequivalent | 8,808 | 0.02 | 0.10 | 0.00 | 0.00 | 0.03 |
| *Number of products* | | | | | | |
| All products | 9,231 | 13.99 | 13.56 | 2.00 | 10.00 | 32.00 |
| Reference | 9,231 | 1.83 | 2.30 | 0.00 | 1.00 | 5.00 |
| Branded | 9,231 | 10.81 | 11.97 | 0.00 | 7.00 | 27.00 |
| Unbranded | 9,231 | 1.35 | 1.76 | 0.00 | 1.00 | 3.00 |
| Bioequivalent | 9,231 | 0.43 | 1.45 | 0.00 | 0.00 | 1.00 |
| *Number of laboratories* | | | | | | |
| All products | 9,231 | 5.13 | 3.46 | 1.00 | 4.00 | 10.00 |
| Reference | 9,231 | 0.51 | 0.52 | 0.00 | 0.00 | 1.00 |
| Branded | 9,231 | 4.06 | 3.36 | 0.00 | 3.00 | 9.00 |
| Bioequivalent | 9,231 | 0.20 | 0.62 | 0.00 | 0.00 | 1.00 |

*Notes:* This table displays descriptive statistics from the IMS data. Statistics for prices are calculated at the product level, while the remainder are calculated at the market level. Market shares are only observed for markets in which at least one product is sold in the period. Statistics for the number of product and laboratories are computed using only observations for which the product or laboratory is found to be active in the corresponding market.

Table 3.4: First stage regressions

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Dep. var.: Number of bioequivalents in market | | | | |
| Post first decree | -1.898*** | -1.584*** | | | -1.534*** |
| | (0.066) | (0.072) | | | (0.073) |
| Post first deadline | | -0.700*** | | | -1.270*** |
| | | (0.066) | | | (0.070) |
| Post last decree | | | 0.742*** | 0.508*** | 1.067*** |
| | | | (0.065) | (0.070) | (0.067) |
| Post last deadline | | | | 0.773*** | 1.032*** |
| | | | | (0.082) | (0.084) |
| Constant | 0.218* | 0.340*** | 0.075 | 0.075 | 0.454*** |
| | (0.113) | (0.113) | (0.117) | (0.117) | (0.109) |
| Market FE | Y | Y | Y | Y | Y |
| Month FE | Y | Y | Y | Y | Y |
| $N$ | 9,027 | 9,027 | 9,027 | 9,027 | 9,027 |
| $R^2$ | 0.602 | 0.607 | 0.571 | 0.575 | 0.633 |
| F-test | 836.2 | 479.6 | 128.9 | 109.1 | 409.8 |

*Notes:* Each column in this table is a regression of the number bioequivalent products in market $m$ at period $t$ on different sets of indicator variables related to policy events. All regressions are weighted by market revenue at the beginning of the sample period. Reported F-tests test for the joint significance of the coefficients on the respective policy indicators. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3.5: Price effects of bioequivalence: Type-level decomposition

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. var: | OLS | IV1 | IV2 |
| *Panel A: Overall effect* | | | |
| Average market price: $P_{mt}$ | 0.017*** | -0.027 | 0.008 |
| | (0.004) | (0.042) | (0.018) |
| $R^2$ | 0.742 | 0.674 | 0.739 |
| *Panel B: Decomposed effect* | | | |
| Entry and exit: $\sum_{s=0}^{t-1} \Delta P_{ms}^{EX}$ | 0.001 | 0.003 | 0.000 |
| | (0.000) | (0.003) | (0.002) |
| $R^2$ | 0.329 | 0.283 | 0.326 |
| Market shares: $\sum_{s=0}^{t-1} \Delta P_{ms}^{RW}$ | -0.015 | 0.016 | -0.099 |
| | (0.036) | (0.234) | (0.155) |
| $R^2$ | 0.421 | 0.418 | 0.399 |
| Price changes: $\sum_{s=0}^{t-1} \Delta P_{ms}^{PC}$ | 0.027** | 0.129 | 0.058 |
| | (0.012) | (0.149) | (0.043) |
| $R^2$ | 0.478 | 0.403 | 0.472 |
| Correlation in changes: $\sum_{s=0}^{t-1} \Delta P_{ms}^{CS}$ | 0.004 | -0.174 | 0.050 |
| | (0.042) | (0.367) | (0.169) |
| $R^2$ | 0.447 | 0.407 | 0.444 |
| Market FE | Y | Y | Y |
| Month FE | Y | Y | Y |
| $N$ | 5,916 | 5,916 | 5,916 |

*Notes:* Each cell in this column corresponds to the coefficient on the number of bioequivalent products in the market. Each column corresponds to a different estimator. Column 1 displays results from OLS regressions, column 2 and 3 display results from IV regressions. Column 2 uses last decree and deadline indicators as instruments. Column 3 uses first and last decree and deadline as instruments. Each row displays results for a different outcomes, each of which is one of the terms in the price decomposition in equation 3.1. Clustered standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3.6: Price effects of bioequivalence: Drug-level decomposition

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| | | Reference | | | Branded non-bioequivalent | | | Unbranded | |
| Dep.var: | OLS | IV1 | IV2 | OLS | IV1 | IV2 | OLS | IV1 | IV2 |
| *Panel A: Overall effect* | | | | | | | | | |
| Average market price: | 0.003 | -0.000 | 0.015 | 0.020* | 0.033 | 0.022 | 0.015 | -0.022 | -0.040 |
| $P_{mt}$ | (0.004) | (0.016) | (0.011) | (0.010) | (0.027) | (0.015) | (0.011) | (0.039) | (0.065) |
| $R^2$ | 0.693 | 0.692 | 0.674 | 0.783 | 0.778 | 0.783 | 0.716 | 0.689 | 0.657 |
| *Panel B: Decomposed effect* | | | | | | | | | |
| Entry and exit: | -0.000 | 0.000 | 0.000 | -0.003 | 0.000 | 0.001 | -0.001 | 0.002 | -0.003 |
| $\sum_{s=0}^{t-1} \Delta P_{ms}^{EX}$ | (0.000) | (0.002) | (0.001) | (0.004) | (0.008) | (0.006) | (0.001) | (0.002) | (0.002) |
| $R^2$ | 0.740 | 0.739 | 0.736 | 0.545 | 0.532 | 0.525 | 0.324 | 0.276 | 0.319 |
| Market shares: | 0.087 | -0.006 | 0.278 | 0.008 | 0.437 | 0.162 | -0.000 | 0.058 | -0.051 |
| $\sum_{s=0}^{t-1} \Delta P_{ms}^{RW}$ | (0.079) | (0.451) | (0.201) | (0.036) | (0.333) | (0.113) | (0.033) | (0.094) | (0.082) |
| $R^2$ | 0.693 | 0.692 | 0.685 | 0.740 | 0.562 | 0.717 | 0.812 | 0.806 | 0.808 |
| Price changes: | 0.075 | -0.115 | 0.266 | 0.001 | 0.344 | 0.124 | -0.003 | 0.090 | 0.031 |
| $\sum_{s=0}^{t-1} \Delta P_{ms}^{PC}$ | (0.077) | (0.424) | (0.195) | (0.027) | (0.268) | (0.091) | (0.030) | (0.099) | (0.045) |
| $R^2$ | 0.695 | 0.687 | 0.687 | 0.771 | 0.641 | 0.754 | 0.844 | 0.833 | 0.843 |
| Correlation in changes: | -0.158 | 0.121 | -0.528 | 0.014 | -0.749 | -0.265 | 0.020 | -0.172 | -0.018 |
| $\sum_{s=0}^{t-1} \Delta P_{ms}^{CS}$ | (0.156) | (0.871) | (0.391) | (0.056) | (0.590) | (0.207) | (0.058) | (0.182) | (0.089) |
| $R^2$ | 0.696 | 0.692 | 0.688 | 0.756 | 0.600 | 0.735 | 0.837 | 0.823 | 0.837 |
| Market FE | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Month FE | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Observations | 4,826 | 4,826 | 4,826 | 6,820 | 6,820 | 6,820 | 4,818 | 4,818 | 4,818 |

*Notes:* Each cell in this column corresponds to the coefficient on the number of bioequivalent products in the market. Each column corresponds to a different estimator. Column 1 displays results from OLS regressions, column 2 and 3 display results from IV regressions. Column 2 uses last decree and deadline indicators as instruments. Column 3 uses first and last decree and deadline as instruments. Each row displays results for a different outcomes, each of which is one of the terms in the price decomposition in equation 3.1. Clustered standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3.7: Quantity effects of bioequivalence

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. var: | OLS | IV1 | IV2 |
| Reference market share | -0.002** | 0.017 | 0.002 |
| | (0.001) | (0.014) | (0.004) |
| $R^2$ | 0.974 | 0.948 | 0.973 |
| Branded market share | 0.003* | -0.018 | -0.002 |
| | (0.002) | (0.015) | (0.005) |
| $R^2$ | 0.965 | 0.931 | 0.963 |
| Unbranded market share | 0.000 | 0.007 | 0.004 |
| | (0.002) | (0.007) | (0.004) |
| $R^2$ | 0.949 | 0.941 | 0.947 |
| Bioequivalent market share | 0.012*** | 0.019** | 0.016** |
| | (0.004) | (0.009) | (0.008) |
| $R^2$ | 0.627 | 0.585 | 0.617 |
| Market FE | Y | Y | Y |
| Month FE | Y | Y | Y |
| $N$ | 9,027 | 9,027 | 9,027 |

*Notes:* Each cell in this column corresponds to the coefficient on the number of bioequivalent products in the market. Each column corresponds to a different estimator. Column 1 displays results from OLS regressions, column 2 and 3 display results from IV regressions. Column 2 uses last decree and deadline indicators as instruments. Column 3 uses first and last decree and deadline as instruments. Each row displays results for a different outcome, each of which is the market share of a segment of the market. Clustered standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3.8: Effects of bioequivalence on the number of products and laboratories

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Number of products | | | Number of laboratories | | |
| Segment | OLS | IV1 | IV2 | OLS | IV1 | IV2 |
| All segments | 0.009** | -0.109** | -0.065* | 0.007** | -0.061* | -0.042* |
| | (0.005) | (0.054) | (0.036) | (0.004) | (0.033) | (0.025) |
| $R^2$ | 0.989 | 0.972 | 0.983 | 0.985 | 0.972 | 0.978 |
| Reference segment | 0.001 | -0.060 | -0.043 | 0.001 | 0.014 | 0.014 |
| | (0.005) | (0.042) | (0.036) | (0.002) | (0.024) | (0.019) |
| $R^2$ | 0.977 | 0.969 | 0.973 | 0.733 | 0.732 | 0.732 |
| Branded segment | 0.009* | -0.036 | -0.012 | 0.003 | -0.023 | -0.006 |
| | (0.005) | (0.047) | (0.018) | (0.004) | (0.023) | (0.009) |
| $R^2$ | 0.982 | 0.973 | 0.980 | 0.968 | 0.961 | 0.967 |
| Bioequivalent segment | 0.223*** | 0.195*** | 0.215*** | 0.131*** | 0.114*** | 0.127*** |
| | (0.022) | (0.057) | (0.034) | (0.012) | (0.034) | (0.018) |
| $R^2$ | 0.924 | 0.920 | 0.924 | 0.834 | 0.829 | 0.833 |
| Market FE | Y | Y | Y | Y | Y | Y |
| Month FE | Y | Y | Y | Y | Y | Y |
| $N$ | 9,027 | 9,027 | 9,027 | 9,027 | 9,027 | 9,027 |

*Notes:* Each cell in this column corresponds to the coefficient on the number of bioequivalent products in the market. Each column corresponds to a different estimator. Columns 1 and 4 displays results from OLS regressions, columns 2, 3, 5 and 6 display results from IV regressions. Columns 2 and 5 uses last decree and deadline indicators as instruments. Column 3 and 6 uses first and last decree and deadline as instruments. Each row displays results for a different outcome, each of which is the log of the number of products or laboratories in the market. Clustered standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

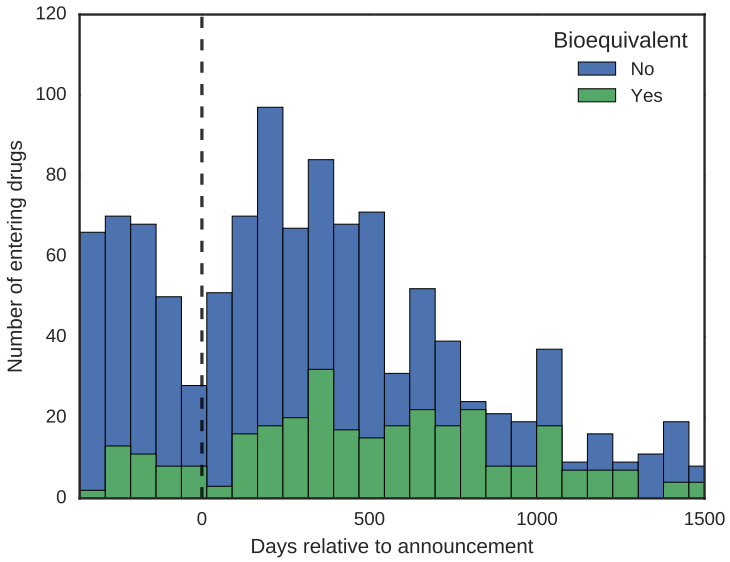
Table 3.9: Effects of bioequivalence on market concentration

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. var: | OLS | IV1 | IV2 |
| Laboratory HHI | -0.003*** | 0.009 | -0.001 |
| | (0.001) | (0.009) | (0.003) |
| $R^2$ | 0.945 | 0.919 | 0.945 |
| CO-Laboratory HHI | -0.003** | 0.009 | -0.001 |
| | (0.001) | (0.009) | (0.003) |
| $R^2$ | 0.934 | 0.911 | 0.934 |
| Market FE | Y | Y | Y |
| Month FE | Y | Y | Y |
| $N$ | 8,808 | 8,808 | 8,808 |

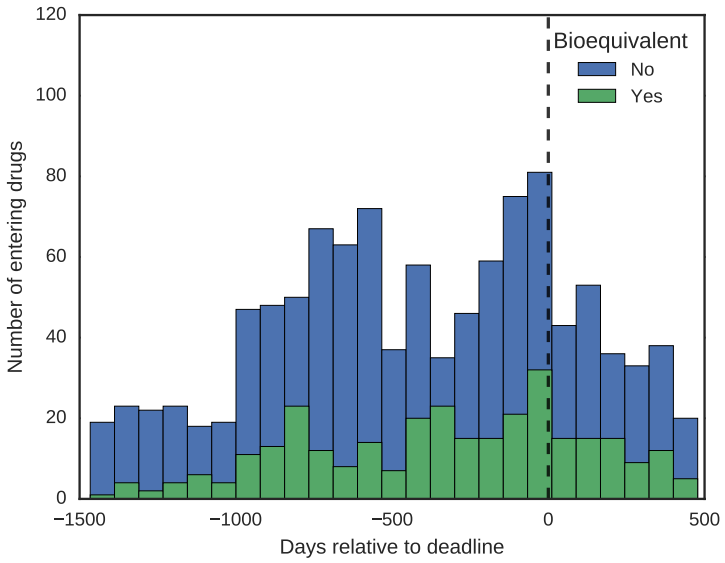
*Notes:* Each cell in this column corresponds to the coefficient on the number of bioequivalent products in the market. Each column corresponds to a different estimator. Column 1 displays results from OLS regressions, column 2 and 3 display results from IV regressions. Column 2 uses last decree and deadline indicators as instruments. Column 3 uses first and last decree and deadline as instruments. Each row displays results for a different measure of HHI, one at laboratory level and one at the level of laboratory ownership. Clustered standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Figure 3.1: Timing of entry for drugs with and without bioequivalence approval around announcement and deadline

(a) Around announcement

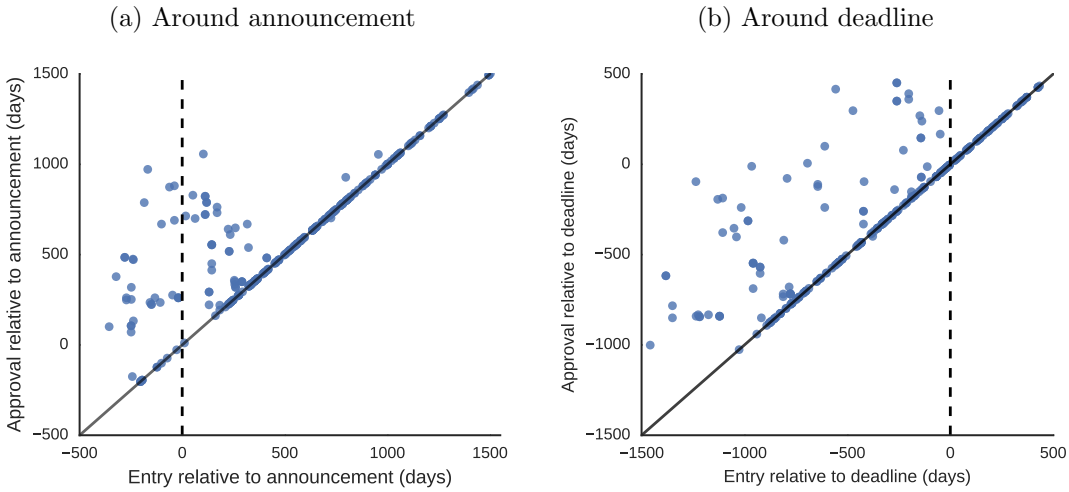



(b) Around deadline




*Notes:* Panel (a) shows the number of new registered drugs with and without bioequivalence requirement, around the date of announcement. Panel (b) shows the number of new registered drugs with and without bioequivalence requirement, around the deadline.

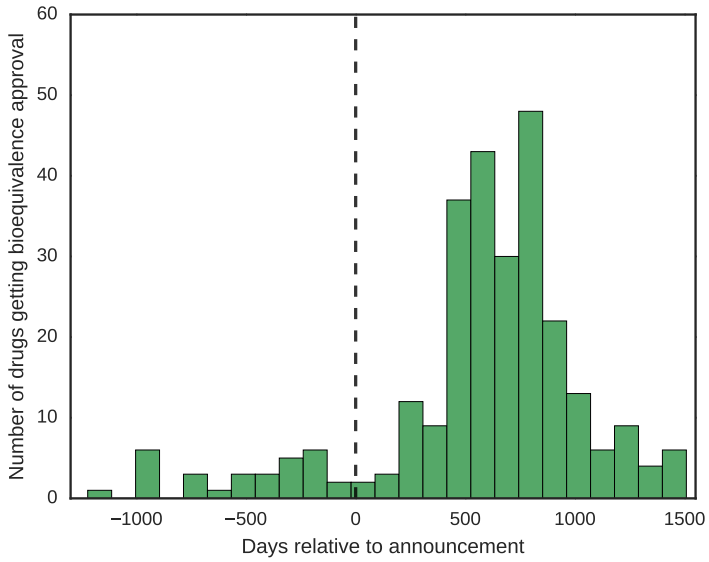
Figure 3.2: Timing of bioequivalence approval relative to entry around time of announcement and deadline



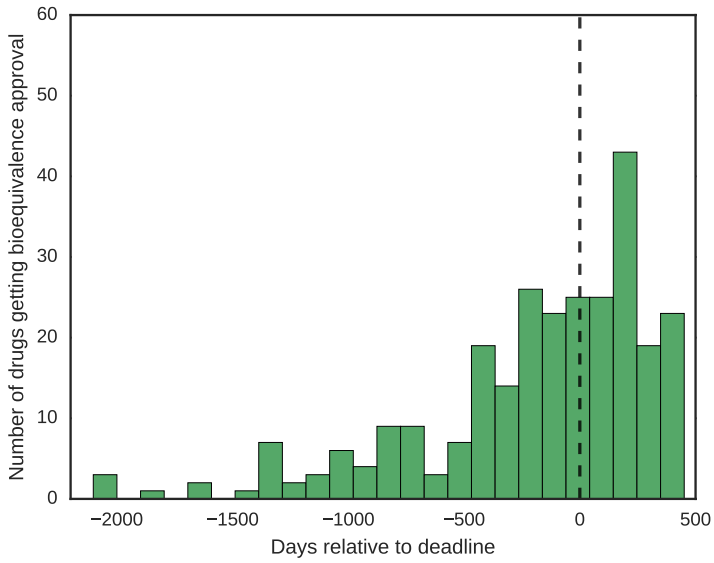

*Notes:* Panel (a) shows the timing of approval against the time of entry, relative to the announcement of the BE requirement. Panel (b) shows the timing of approval against the time of entry, relative to the deadline imposed by the requirement.

Figure 3.3: Timing of bioequivalence approval for drugs with previous registration around announcement and deadline

(a) Around announcement

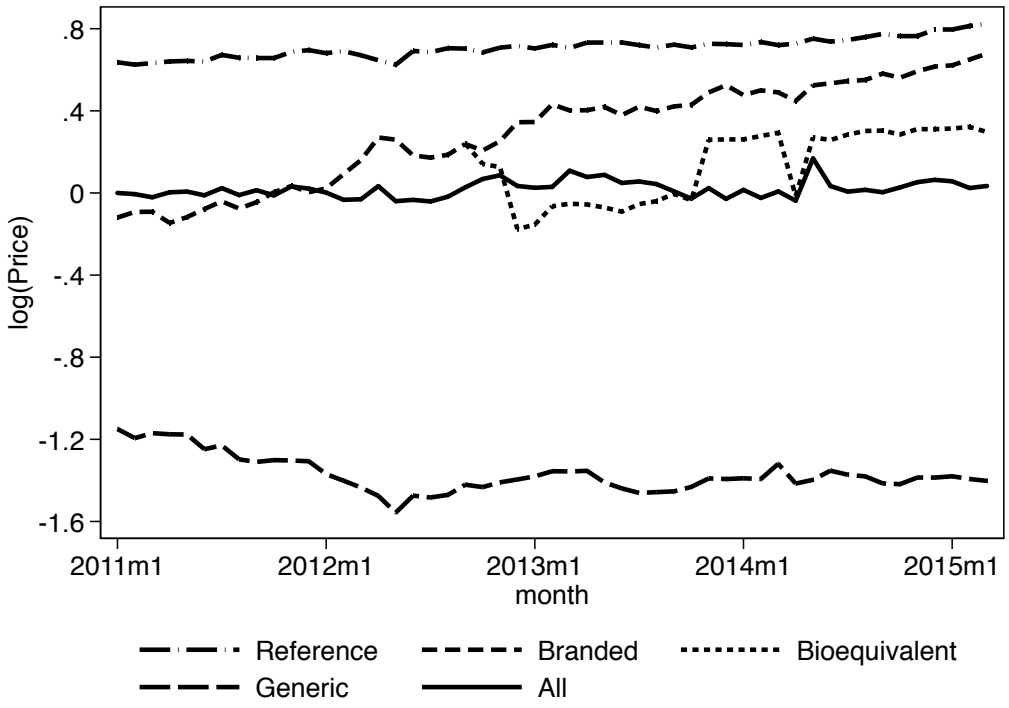



(b) Around deadline



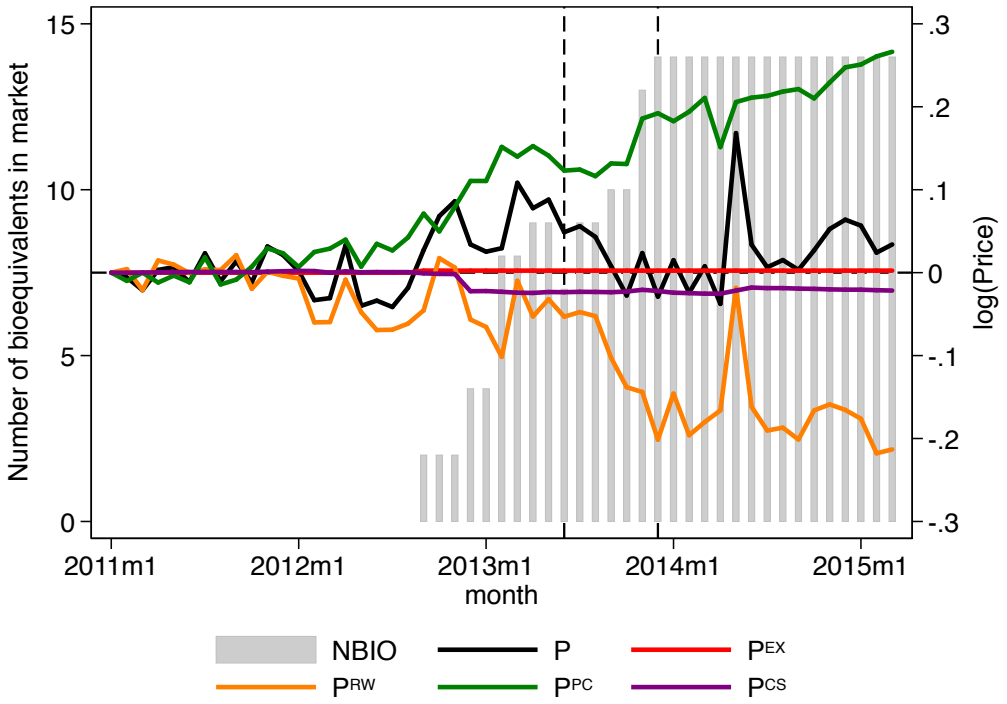

*Notes:* Panel (a) shows the timing of BE approval relative to the announcement for drugs registered at least one year before the announcement. Panel (b) shows the timing of BE approval relative to the deadline for drugs registered at least one year before the announcement

Figure 3.4: Price evolution by group, Metformin




*Notes:* This figure shows the evolution of the average price for Metformin as well as the evolution of the price of each of the four groups : Reference, branded, bioequivalent, and generic. Series are in logs and normalized to be equal to 0 for the average price in January 2011m1

Figure 3.5: Decomposition of average market prices

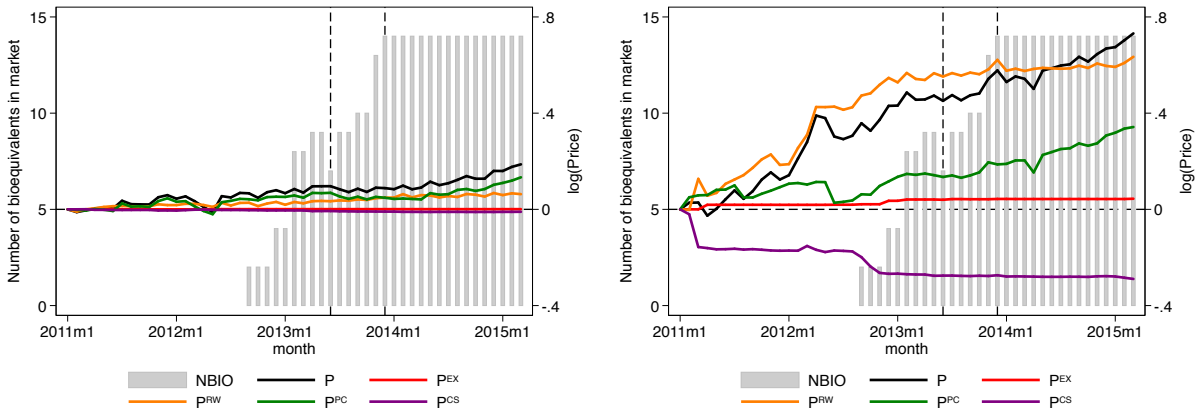



*Notes:* This figure displays an example for the decomposition proposed in equation 3.1. The market utilized for the plot is Metformin, the largest one in our sample in terms of revenue. In particular, lines in the figure plot the time series for each component of the decomposition. The observed average market price is plotted in black for reference. The number of bioequivalent products in this market is also included in the figure.

Figure 3.6: Decomposition of average segment prices

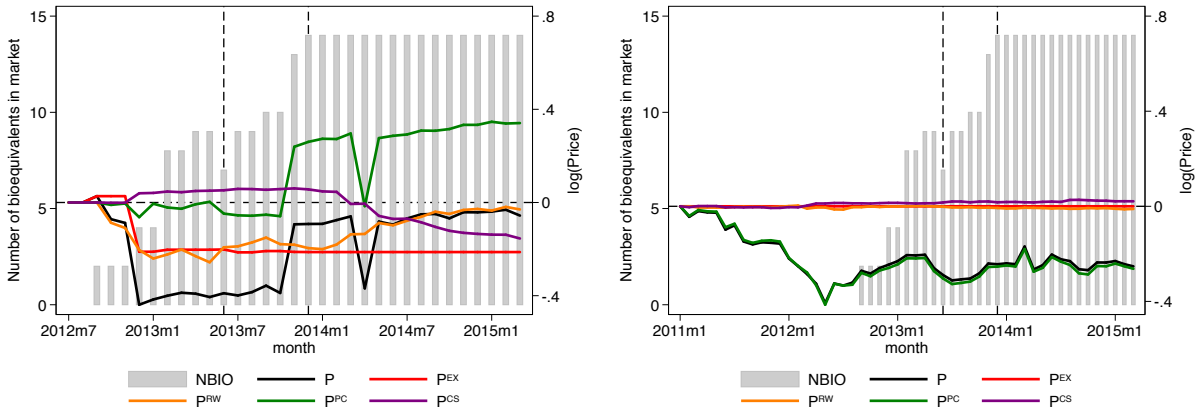
(a) Reference products

(b) Branded non-bioequivalent products




(c) Branded bioequivalent products

(d) Unbranded products




*Notes:* This figure displays an example for the decomposition proposed in equation 3.2. The market utilized for the plot is Metformin, the largest one in our sample in terms of revenue. In particular, lines in the figure plot the time series for each component of the decomposition. The observed average segment price is plotted in black for reference. The number of bioequivalent products in this market is also included in each figure.

# Bibliography

Abaluck, Jackson and Jonathan Gruber (2011). "Choice Inconsistencies Among the Elderly: Evidence from Plan Choice in the Medicaire Part D Program". In: *American Economic Review* 101, pp. 1180–1210.

Abaluck, Jason and Jonathan Gruber (2013). "Evolving Choice Inconsistencies in Choice of Prescription Drug Insurance". NBER Working Paper 19163.

Abaluck, Jason, Jonathan Gruber, and Ashely Swanson (2015). "Prescription Drug Utilization under Medicare Part D: A Dynamic Perspective". NBER Working Paper 20976.

Akerlof, George (1970). "The Market for "Lemons": Quality Uncertainty and the Market Mechanism". In: *Quarterly Journal of Economics* 84.3.

Baicker, Katherine and Amitabh Chandra (2004). "The Productivity of Physician Specialization: Evidence from the Medicare Program". In: *American Economic Review* 94.2, pp. 357–361.

Balmaceda, Carlos, Manuel Espinoza, and Janepsy Diaz (2015). "Impacto de una Politica de Equivalencia Terapeutica en el Precio de Medicamentos en Chile". In: *Value in Health Regional Issues*.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2010). "Social Incentives in the Workplace". In: *Review of Economic Studies* 77.2, pp. 417–458.

Berry, Stephen (1996). "Estimating Discrete Choice Models of Product Differentiation". In: *RAND Journal of Economics* 25, p. 242.

Bitran, E., L. Escobar, and P. Gassibe (2010). "After Chile's Health Reform: Increase in Coverage and Access, Decline in Hospitalization And Death Rates". In: *Health Affairs* 29.12, pp. 2161–2170.

Blough, D.K., C.W. Madden, and M.C. Hornbrook (1999). "Modeling Risk using Generalized Linear Models". In: *Journal of Health Economics*.

Bundorf, Kate, Jonathan Levin, and Neale Mahoney (2012). "Pricing and Welfare in Health Plan Choice". In: *American Economic Review* 102.7, pp. 1–38.

Buntin, Melinda Beeuwkes and Alan Zalavsky (2004). "Too much ado about Two-part Models and Transformation?: Comparing Methods of Modeling Medicare Expenditures". In: *Journal of Health Economics*.

Cabral, Marika (2015). "Claim Timing and Ex Post Adverse Selection". Working Paper.

Caves, Richard E. et al. (1991). "Patent Expiration, Entry, and Competition in the U.S. Pharmaceutical Industry". In: *Brookings Papers on Economic Activity. Microeconomics* 1991, pp. 1–66.

Chandra, Amitabh and Douglas Staiger (2007). "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks". In: *Journal of Political Economy* 115, pp. 103–140.

Ching, Andrew, Tulin Erdem, and Michael Keane (2009). "The Price Consideration Model of Brand Choice". In: *Journal of Applied Econometrics* 24.3, pp. 393–420.

Cid, Camilo and Lorena Prieto (2012). "El Gasto de Bolsillo en Salud: El Caso de Chile, 1997 y 2007". In: *Revista Panamericana de Salud Publica* 31.4, pp. 310–316.

Cochrane, John (1995). "Time-Consistent Health Insurance". In: *Journal of Political Economy* 103.3, pp. 445–473.

Criteria Research (2008). "Dimensiones de Valor para el Usuario de ISAPRES en la elección de Planes de Salud". Study for Superintendencia de Salud.

Crocker, Keith and John Moran (2003). "Contracting With Limited Commitment: Evidence From Employment-Based Insurance". In: *RAND Journal of Economics* 94, pp. 321–344.

Dague, Gaston and Laura Palmucci (2015). "The Welfare Effects of Banning Risk-Rated Pricing in Health Insurance Markets: Evidence From Chile". Working Paper.

Dalton, Christina, Gautam Gowrisankaran, and Robert Town (2015). "Myopia and Complex Dynamic Incentives: Evidence from Medicare Part D". NBER Working Paper 21104.

David, Barbara et al. (2013). "International Guidelines for Bioequivalence of Systemically Available Orally Administered Generic Drug Products : A Survey of Similarities and Differences". In: *The AAPS Journal* 15.4.

Delia, Derek and Joel C Cantor (2009). "Emergency Department Utilization and Capacity". In: *The Synthesis project. Research synthesis report* 17, pp. 1–11.

Duarte, Fabian (2012). "Price Elasticity of Expenditures Across Health Care Services". In: *Journal of Health Economics*.

Duggan, Mark, Craig Garthwaite, and Aparajita Goyal (2016). "The Market Impacts of Pharmaceutical Product Patents in Developing Countries: Evidence from India". In: *American Economic Review* 106.1, pp. 99–135.

Einav, Liran and Amy Finkelstein (2011). "Selection in Insurance Markets: Theory and Empirics in Pictures". In: *Journal of Economic Perspectives* 25.1, pp. 115–138.

Ericson, Keith (2012). "Consumer Inertia and Firm Pricing in the Medicare Part D Prescription Drug Insurance Exchange". NBER Working Paper No. 18359.

Falk, Armin and Andrea Ichino (2006). "Clean evidence on peer effects". In: *Journal of Labor Economics* 24.1, pp. 39–57.

Farrell, Joseph and Paul Klemperer (2007). "Coordination and Lock-In: Competition with Switching Costs and Network Effects". In: vol. 3. Handbook of Industrial Organization. Elsevier. Chap. 31, pp. 1967–2072.

Farrell, Joseph and Carl Shapiro (1989). "Optimal Contracts with Lock-in". In: *American Economic Review* 79.1, pp. 51–68.

Fee, C. et al. (2007). "Effect of Emergency Department Crowding on Time to Antibiotics in Patients Admitted with Community-acquired Pneumonia". In: *Annals of emergency medicine* 50.5, pp. 501–509.

Fehr, Ernst and Klaus M Schmidt (1999). "A Theory of Fairness, Competition, and Cooperation". In: *The Quarterly Journal of Economics* 114.3, pp. 817–868.

Finkelstein, Amy, Kathleen McGarry, and Amir Sufi (2005). "Dynamic Inefficiencies in Insurance Markets: Evidence from Long-term Care Insurance". NBER Working Paper 11039.

Food and Drug Administration (2016). *Approved Drug Products with Therapeutic Equivalence Evaluations*.

Foster, Lucia, John C. Haltiwanger, and C. J. Krizan (2011). "New Developments in Productivity Analysis". In: ed. by Charles R. Hulten, Edwin R. Dean, and Michael J. Harper. University of Chicago Press. Chap. Aggregate Productivity Growth. Lessons from Microeconomic Evidence.

Frank, Richard G. and David S. Salkever (1992). "Pricing, Patent Loss and the Market for Pharmaceuticals". In: *Southern Economic Journal* 59.2, pp. 165–179.

Geruso, Michael (2013). "Selection in Employer Health Plans: Homogeneous Prices and Heterogeneous Preferences". unpublished draft.

Geweke, J. and Michael Keane (2001). "Computationally Intensive Methods for Integration in Econometrics". In: *Handbook of Econometrics*. Ed. by James Heckman and E.E. Leamer.

Geweke, John, Michael Keane, and David Runkle (1997). "Statistical Inference in the Multinomial Multiperiod Probit Model". In: *Journal of Econometrics*.

Grabowski, Henry et al. (2006). "Generic Competition in the US Pharmaceutical Industry". In: *International Journal of the Economics of Business* 13.1, pp. 15–38.

Grubb, Michael and Matthew Osborne (2015). "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock". In: *American Economic Review* 105.1.

Guryan, Jonathan, Kory Kroft, and Matthew J. Notowidigdo (2009). "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments". In: *American Economic Journal: Applied Economics* 1.4, pp. 34–68.

Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud (1996). "Simulation of Multivariate Normal Rectangle Probabilities and their Derivatives Theoretical and Computational Results". In: *Journal of Econometrics*.

Handel, Benjamin (2013). "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts". In: *American Economic Review* 103.7.

Handel, Benjamin, Igal Hendel, and Michael Whinston (2015a). "Equilibria in Health Exchanges: Adverse Selection vs. Reclassification Risk". In: *Econometrica* 83.4.

— (2015b). "The Welfare Impacts of Long-Term Contracts". unpublished draft.

Hartman, M. et al. (2013). "National Health Spending in 2011: Overall Growth Remains Low, but some Payers and Services Show Signs of Acceleration". In: *Health Affairs* 32.1, pp. 87–99.

Hendel, Igal and Alessandro Lizzeri (2003). "The Role of Commitment in Dynamic Contracts: Evidence from Life Insurance". In: *Quarterly Journal of Economics*.

Hendren, Nathaniel (2013). "Private Information and Insurance Rejections". In: *Econometrica* 81.5, pp. 1713–1762.

Herring, Bradley and Mark Pauly (2006). "Incentive-compatible Guaranteed Renewable Health Insurance Premiums". In: *Journal of Health Economics*, pp. 395–417.

Hofmann, Annette and Mark Browne (2013). "One-sided Commitment in Dynamic Insurance Contracts: Evidence from Private Health Insurance in Germany". English. In: *Journal of Risk and Uncertainty* 46.1, pp. 81–112.

Hoot, N. R. and D. Aronsky (2008). "Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions". In: *Annals of emergency medicine* 52.2, pp. 126–136.

Hyslop, Dean (1999). "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women." In: *Econometrica*.

Institute of Medicine (2007). *Hospital-Based Emergency Care: At the Breaking Point*. Washington, DC: National Academies Press.

Jackson, C Kirabo and Elias Bruegmann (2009). "Teaching Students and Teaching each other: The Importance of Peer Learning for Teachers". NBER Working Paper 15202.

Kandel, Eugene and Edward P Lazear (1992). "Peer Pressure and Partnerships". In: *Journal of Political Economy* 100.4, pp. 801–17.

Keane, Michael (1993). "Simulation Estimation for Panel Data Models with Limited Dependent Variables". In: *Handbook of Statistics* 11.

— (1994). "A Computationally Practical Simulation Estimator for Panel Data". In: *Econometrica*.

— (1997). "Modeling Heterogeneity and State Dependence in Consumer Choice Behavior". In: *Journal of Business and Economic Statistics* 15.3, pp. 3100–327.

— (2013). "Panel Data Discrete Choice Models of Consumer Demand". Prepared for The Oxford Handbooks: Panel Data.

Ketcham, Jonathan D. et al. (2012). "Sinking, Swimming, or Learning to Swim in Medicare Part D". In: *American Economic Review*.

Kline, Patrick (2012). "The Impact of Juvenile Curfew Laws on Arrest of Youth and Adults". In: *American Law and Economics Review* 14.1, pp. 44–67.

Knittel, Chris and Peter Huckfeldt (2012). "Pharmaceutical Use Following Generic Entry: Paying Less and Buying Less".

Koch, Thomas (2011). "One Pool to Insure them All?: Age, Risk and the Price(s) of Medical Insurance". U.C. Santa Barbara Working Paper.

Krueger, Dirk and Harald Uhlig (2006). "Competitive Risk Sharing Contracts with One-Sided Commitment". In: *Journal of Monetary Economics*.

Lakdawalla, Darius, Tomas Philipson, and Y. Richard Wang (2006). "Intellectual Property and Marketing". NBER Working Paper 12577.

Lichtenberg, F. R. and G. Duflos (2009). *Time Release: The effect of Patent Expiration on U.S. Drug Prices, Marketing, and Utilization by the Public*. Tech. rep. Manhattan Institute.

Marquis, Susan et al. (2006). "Consumer Decision Making in the Individual Health Insurance Market". In: *Health Affairs* 25.3.

Mas, Alexandre and Enrico Moretti (2009). "Peers at Work". In: *American Economic Review* 99.1, pp. 112–45.

Ministerio de Salud (2013). *Medicamentos en Chile: Revision de la Evidencia del Mercado Nacional de Farmacos*. Tech. rep. Ministry of Health, Chile.

Nedler, J.A. (1989). *Generalized Linear Models*. Chapman and Hall, New York.

OECD (2000). *Competition and Regulation Issues in the Pharmaceutical Industry*. Tech. rep. OECD.

— (2013). *Health at a Glance 2013: OECD Indicators*. Tech. rep. OECD.

Pardo, Cristian and Whitney Schott (2012). "Public versus Private: Evidence on Health Insurance Selection". In: *International Journal of Health Care Finance and Economics* 12, pp. 39–61.

— (2013). "Health Insurance Selection: A Cross-sectional and Panel Analysis". In: *Health Policy and Plannig*.

Patel, Vip and Mark Pauly (2002). "Guaranteed Renewability and the Problem of Risk Variation in Individual Health Insurance Markets". In: *Health Affairs*.

Pauly, Mark and Bradley Herring (1999). *Pooling Health Insurance Risks*. Washington, D.C.: The AEI Press, Publisher for the American Enterprise Institution.

Pauly, Mark, Howard Kunreuther, and Richard Hirth (1995). "Guaranteed Renewability in Insurance". In: *Journal of Risk and Uncertainty*.

Petrin, Amil and Kenneth Train (2009). "A Control Function Approach to Endogeneity in Consumer Choice Models". In: *Journal of Marketing Research*.

Shepard, Mark (2015). "Hospital Network Competition and Adverse Selection: Evidence from the Massachusetts Health Insurance Exchange". Job Market Paper.

Skinner, Jonathan and Douglas Staiger (2009). "Technology Diffusion and Productivity Growth in Health Care". NBER Working Paper 14865.

Train, Kenneth (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Wooldridge, Jeffrey (2005). "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity". In: *Journal of Applied Econometrics*.

# Appendix A

# Lock-in in Dynamic Health Insurance Contracts

## A.1 Multinomial logit for destination company

I specify the multinomial logit among switchers, for the probability that individual $i$ chooses firm $k$ upon switching, as:

$$p_i^k = \frac{e^{X_i'\beta^k}}{\sum_{l=1}^{K} e^{X_i\beta^l}}$$

where $\beta^k$ are firm-specific coefficients and $X_i$ are individual-specific regressors that include pre-switching health status as well as other demographics. Table A.1 shows the estimated $\beta$ coefficients as well as the $\chi^2$ statistic for the null that all health coefficients are equal to zero. Column (1) corresponds to a specification that only includes a dummy $preex = 1$ of a preexisting condition (equal to one if at any point in the past the individual received treatment related to any condition). Column (2) adds age, gender and (the log) wage. Column (3) replaces the preexisting condition dummy by a dummy of a preexisting condition in the three months prior to switching, $preex_{3m} = 1$. Finally (4) replaces this variable by $log(1 + healthpreex_{3m})$, where $healthpreex_{3m}$ summarizes the total health expenditures related to preexisting conditions in the previous three months. The tables shows the $X_4^2$ statistic and p-value for the Wald test that all variables related to pre-switching health expenditures are equal to zero.

Table A.1: Multinomial logit for destination company among switchers

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Firm = B | | | | |
| $\mathbb{1}(preex)$ | 0.139** | 0.091 | | |
| | (0.054) | (0.063) | | |
| $\mathbb{1}(preex_{3m})$ | | | -0.070 | |
| | | | (0.121) | |
| $log(1 + healthpreex_{3m})$ | | | | 0.044 |
| | | | | (0.176) |
| Firm = C | | | | |
| $\mathbb{1}(preex)$ | -0.096 | -0.319*** | | |
| | (0.081) | (0.098) | | |
| $\mathbb{1}(preex_{3m})$ | | | -0.627*** | |
| | | | (0.212) | |
| $log(1 + healthpreex_{3m})$ | | | | -0.685* |
| | | | | (0.402) |
| Firm = D | | | | |
| $\mathbb{1}(preex)$ | -0.004 | -0.009 | | |
| | (0.059) | (0.066) | | |
| $\mathbb{1}(preex_{3m})$ | | | -0.245* | |
| | | | (0.132) | |
| $log(1 + healthpreex_{3m})$ | | | | -0.228 |
| | | | | (0.226) |
| Firm = E | | | | |
| $\mathbb{1}(preex)$ | -0.228*** | -0.241*** | | |
| | (0.059) | (0.068) | | |
| $\mathbb{1}(preex_{3m})$ | | | -0.299** | |
| | | | (0.132) | |
| $log(1 + healthpreex_{3m})$ | | | | -0.664*** |
| | | | | (0.234) |
| Demographics | No | Yes | Yes | Yes |
| $N$ | 16943 | 12971 | 12971 | 12427 |
| $\chi^2_4$ | 47.0 | 37.6 | 13.1 | 12.6 |
| $(p - value)$ | (0.00) | (0.00) | (0.01) | (0.01) |

*Notes*: Mulitnomial model estimated by ML for the probability of switching to each firm B-E among switchers, as a function of health conditions and other demographics. A is the base category. if the individual was treated for a preexisting condition at any point in time before switching. $\mathbb{1}(preex_{3m})$ calculates this indicator using the 3 months before switching. $log(1 + healthpreex_{3m}$ is the log of 1 plus the total health expenditures related to preexisting conditions in the 3 months prior to switching. $\chi^2_4$ if for null that all coefficients related to health status are jointly equal to zero. The corresponding p-value is in parenthesis. Robust standard errors for multinomial-logit coefficients in parentheses * p<0.10, ** p<0.05, *** p<0.01

## A.2 Market shares for each geographical region

To understand the variation of preferences across geographic areas I investigate the role of "in-network providers". I exploit the geographic variation in the presence of in-network providers of different companies, and show that providers matter for the decision to enroll in a given insurance company.[1] Specifically, I investigate the relationship between in-network providers and market share within each district by estimating:

$$ln\left(mshare_{kd}\right) = \eta_k + \beta NP_{kd} + \epsilon_{kd} \tag{A.1}$$

where $\eta_k$ are insurer fixed-effects, $NP_{kc}$ is an indicator variable that is equal to 1 if insurer $j$ has a network provider in district $d$ and $\epsilon_{kc}$ is an error capturing other determinants of market share.

Column (1) of the following Table shows the results of estimating equation A.1 for all 1934 districts. Column (2) restricts the sample to districts in which all ISAPREs have at least one client (78% of the districts) while column (3) restrict the sample to districts in which also the market is higher than 500 individuals (894 districts).

Table A.2: Market share and in-network providers

| | (1) | (2) | (3) |
|---|---|---|---|
| $NP$ | 0.258*** | 0.259*** | 0.321*** |
| | (0.048) | (0.049) | (0.062) |
| sample | all | all clients $> 0$ | market $> 500$ |
| $N$ | 1934 | 1620 | 894 |

*Notes*: Table shows OLS estimates of equation A.1. Standard errors in parentheses * p<0.10, ** p<0.05, *** p<0.01

In all three specifications I find that a network provider significantly increases the market share of an insurance company, between 26 % and 32 % depending on the specification. In the following Table I investigate the same relationship with narrower sources of identification by adding region fixed effects (column 2) and province fixed effects (column 3) to equation A.1. Using only within-region variation or within province variation I find smaller but still significant effects.

[1] I identify the presence of in-network providers with the claims data, since claims are classified either done at an "in-network" or at an "out-network" provider.

Table A.3: Market share and in-network providers

| | (1) | (2) | (3) |
|---|---|---|---|
| NP | 0.321*** | 0.189*** | 0.118** |
| | (0.062) | (0.047) | (0.054) |
| FE | No | Region | Province |
| *N* | 894 | 894 | 894 |

*Notes*: Table shows OLS estimates of equation A.1. Standard errors in parentheses * p<0.10, ** p<0.05, *** p<0.01

## A.3 Multinomial logit for initial choice: Testing forward-looking behavior

I test for forward looking behavior regarding insurance company enrollment, by testing future health shocks on the current decision of health insurance company using a multinomial logit specification.

Table A.4: Multinomial logit: Active choice as a function of future health expenditures

| | Firm | | | |
|---|---|---|---|---|
| | A | B | C | D |
| $log(1+h_{i,t})$ | 0.011 | -0.004 | 0.089** | 0.015 |
| | (0.029) | (0.026) | (0.030) | (0.026) |
| $log(1+h_{i,t+1})$ | 0.02 | -0.016 | 0.021 | 0.014 |
| | (0.017) | (0.017) | (0.024) | (0.017) |
| $log(1+h_{i,t+2})$ | 0.011 | 0.024 | 0.024 | 0.017 |
| | (0.019) | (0.017) | (0.023) | (0.016) |
| wage | 0.000 | 0.000 | -0.000** | 0.000*** |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| age | -0.031*** | -0.029* | -0.034*** | -0.029*** |
| | (0.009) | (0.011) | (0.008) | (0.007) |
| gender | 0.302 | 0.428 | 2.024*** | -0.275 |
| | (0.359) | (0.472) | (0.286) | (0.283) |
| stgo | -0.365 | -0.22 | -0.113 | -1.713*** |
| | (0.251) | (0.288) | (0.231) | (0.173) |
| Plan Controls | Yes | Yes | Yes | Yes |

*Notes*: Table shows ML estimation of a multinomial logit for the active choices of company as a function of future health expenditures and other controls. Standard errors in parentheses * p<0.10, ** p<0.05, *** p<0.01

## A.4 GHK algorithm

Here I give details on the steps to apply the GHK algorithm to a setting with varying choice sets. I adapt the methodology outlined by John Geweke, Keane, and Runkle, 1997 and Train, 2009.

Assume the following model for $U_{it}^{jk}$, where $j$ denotes plans and $k$ companies.

$$U_{it}^{jk} = \alpha_i^k + V_{it}^{jk} + u_{it}^{jk}$$

where the $\alpha_i^k$ are treated as random utility $\alpha_i^k \sim N\left(\mu^k, \sigma^k\right)$. Here I incorporate all the (deterministically) time-varying portion of preference heterogeneity discussed in the text in $V_{it}^{jk}$.

I define $\epsilon_{it}^{jk} = \alpha_i^k + u_{it}^{jk}$ as the composite random error term. I assume that $u$ is an AR(1) process, where I allow autocorrelation within the same plan, and also autocorrelation across plans within a same company. Therefore the composite error has cross-sectional correlation for plans of the same company and time-series correlation for plans of the same company and within the same plan.

Since in the choice set there is only one plan per insurance plus the guaranteed-renewable plan, I drop from here on the $j$ subscript, and denote the GR plan as the $GR$ option. I also drop the $i$ subscript to simplify notation.

Let $\tilde{s} = \{k_1, k_2\}$ be the sequence of chosen options in period 1 and 2 and $E(s)$ the define the vector of stacked error terms across time periods for each individual as

$$E(\tilde{s}) = \left(\epsilon_1^1, ..., \epsilon_1^K, \epsilon_2^1, ..., \epsilon_2^K, \epsilon_2^{GR}(\tilde{s}), \epsilon_3^1, ..., \epsilon_3^K, \epsilon_3^{GR}(\tilde{s})\right)'$$

This vector depends on $\tilde{s} = \{k_1, k_2\}$ because the $\epsilon_2^{GR}$ and $\epsilon_3^{GR}$ are defined by the individual's choice sequence. Similarly define

$$A(\tilde{s}) = \left(\alpha_1^1, ..., \alpha_1^K, \alpha_2^1, ..., \alpha_2^K, \alpha_2^{GR}(\tilde{s}), \alpha_3^1, ..., \alpha_3^K, \alpha_3^{GR}(\tilde{s})\right)'$$

and

$$U(\tilde{s}) = \left(u_1^1, ..., u_1^K, u_2^1, ..., u_2^K, u_2^{GR}(s), u_3^1, ..., u_3^K, u_3^{GR}(\tilde{s})\right)'$$

I can write succinctly,

$$E(\tilde{s}) = A(\tilde{s}) + U(\tilde{s})$$

Let $\Omega(\tilde{s}) = cov(E(\tilde{s}))$. Since $u$ and $\alpha$ are assumed to be uncorrelated

$$\begin{aligned}\Omega(\tilde{s}) &= cov(A(\tilde{s})) + cov(U(\tilde{s})) \\ &= \Gamma(\tilde{s}) + \Sigma(\tilde{s})\end{aligned}$$

Let $E^K(\tilde{s}) = M^K \times E(\tilde{s})$ where $M^K$ is the matrix such that I am taking differences with respect alternative $K$ in each period. The $M^K$ matrix is defined as

$$M^K = \begin{bmatrix} M_1^K & 0 & 0 \\ 0 & M_2^K & 0 \\ 0 & 0 & M_3^K \end{bmatrix}$$

where $M_1^K$ is a $K-1$ identity matrix with an added column of $-1's$ in the $K$th position. Similarly, $M_2^K$ and $M_3^K$ is a $K$ identity matrix with an added column of $-1$s in the $K$ position (periods 2 and 3 include in the last column the guaranteed-renewable contract). Let $\Omega^K(\tilde{s})$ be the corresponding covariance matrix

$$\Omega^K(\tilde{s}) = M^K \Omega(\tilde{s})$$

For each sequence $\tilde{s}$, calculate the Choleskly factor $L^K(\tilde{s})$ such that

$$\Omega^K(\tilde{s}) = L^K(\tilde{s})' L^K(\tilde{s})$$

Then I calculate for each sequence $\tilde{s}$ the Cholesky factor of the undifferentiated errors by adding a row of zeros in the $K^{th}$ row corresponding of each period, resulting in matrix $L(\tilde{s})$. (Train, 2009).

For each choice $k$ in $t$ I define the matrix $M_k^t$ as the $N_t$ identity matrix with an extra column of $-1$'s added in the $K^{th}$ column. Note that $N_1 = K-1$ and $N_t = K\ \forall t > 1$ (since the choice in every period after the first includes the choice from each company and the guaranteed-renewable contract). For a given sequence $s = \{k^1, k^2, k^3\} = \{\tilde{s}, k^3\}$, I define the matrix $M(s)$ as

$$M(s) = \begin{bmatrix} M_{k_1}^1 & 0 & 0 \\ 0 & M_{k_2}^2 & 0 \\ 0 & 0 & M_{k_3}^3 \end{bmatrix}$$

I calculate for each sequence $s$ the covariance matrix as

$$\Omega(s) = (M(s) L(\tilde{s}))(M(s) L(\tilde{s}))'$$

Finally, I take the Choleksy decomposition of $\Omega(s) = L(s)' L(s)$. [2]

The matrix $L(s)$ is a $K-1+(T-1)\times K$ lower-triangular matrix.

$$L(s) = \begin{bmatrix} L^{11} & 0 & 0 \\ S^{21} & L^{22} & 0 \\ S^{31} & s^{32} & L^{33} \end{bmatrix}$$

The first $(K-1)\times(K-1)$ elements, $L^{11}$, correspond to the Cholesky decomposition of the differentiated errors in period 1 for an individual that chose $k_1$ consistent with $s$. Then $S^{21}$ is the Cholesky decomposition of the error terms in period 2 with respect to period 1 errors, and so on. Therefore, the the stacked $E(i)$, can be write as a function of series of vector $K-1+(T-1)\times K$ iid errors $\eta_i$ as

$$\epsilon_i = L(s) \times \eta_i$$

where $\eta_i$ is a normal iid.

Then, for period 1, I perform the following steps (see John Geweke, Keane, and Runkle, 1997)

| **step** | | | | |
|---|---|---|---|---|
| (1) | draw | $\eta_1^{1,r}$ | s.t. | $\tilde{V}_1^1\left(\eta_1^{1,r}\right) < 0$ |
| $\vdots$ | | | | |
| $(c_1-1)$ | draw | $\eta_1^{c_1-1,r}$ | s.t. | $\tilde{V}_1^{c_1-1}\left(\eta_1^{1,r},...,\eta_1^{c_1-1,r}\right) < 0$ |
| $(c_1)$ | skip | $\eta_1^{c_1,r}$ | | |
| $(c_1+1)$ | draw | $\eta_1^{c_1+1,r}$ | s.t. | $\tilde{V}_1^{c_1+1}\left(\eta_1^{1,r},...,\eta_1^{c_1-1,r},\eta_1^{c_1+1,r}\right) < 0$ |
| $\vdots$ | | | | |
| $(K-1)$ | draw | $\eta_1^{K-1,r}$ | s.t. | $\tilde{V}_1^{K-1}\left(\eta_1^{1,r},...,\eta_1^{c_1-1,r},\eta_1^{c_1+1,r},...\eta_1^{K-1,r}\right) < 0$ |

[2] Note that with this procedure, the $L(K,K,K) = L_K$, the Cholesky decomposition of the matrix $\Omega_K$ I used to parametrize the model

Similarly, for period 2,

| **step** | | | | |
|---|---|---|---|---|
| (1) | draw | $\eta_2^{1,r}$ | s.t. | $\tilde{V}_2^1\left(\eta_1^{1,r},...,\eta_1^{K-1,r},\eta_2^{1,r}\right)<0$ |
| $\vdots$ | | | | |
| $(c_2-1)$ | draw | $\eta_2^{c_2-1,r}$ | s.t. | $\tilde{V}_2^{c_2-1}\left(\eta_1^{1,r},...,\eta_1^{K-1,r},\eta_2^{1,r},...,\eta_2^{c_2-1,r}\right)<0$ |
| $(c_1)$ | skip | $\eta_2^{c_2,r}$ | | |
| $(c_1+1)$ | draw | $\eta_2^{c_2+1,r}$ | s.t. | $\tilde{V}_2^{c_2+1}\left(\eta_1^{1,r},...,\eta_1^{K-1,r},\eta_2^{1,r},...,\eta_2^{c_2-1,r},\eta_2^{c_2+1,r}\right)<0$ |
| $\vdots$ | | | | |
| $(K)$ | draw | $\eta_2^{K,r}$ | s.t. | $\tilde{V}_2^K\left(\eta_1^{1,r},...,\eta_1^{K-1,r},\eta_2^{1,r},...,\eta_2^{c_2-1,r},\eta_2^{c_2+1,r},..,\eta_2^{K,r}\right)<0$ |

and for period 3

| **step** | | | | |
|---|---|---|---|---|
| (1) | draw | $\eta_3^{1,r}$ | s.t. | $\tilde{V}_3^1\left(\eta_1^{1,r},...,\eta_2^{K,r},\eta_3^{1,r}\right)<0$ |
| $\vdots$ | | | | |
| $(c_2-1)$ | draw | $\eta_3^{c_3-1,r}$ | s.t. | $\tilde{V}_3^{c_2-1}\left(\eta_1^{1,r},...,\eta_2^{K,r},\eta_3^{1,r},...,\eta_3^{c_3-1,r}\right)<0$ |
| $(c_1)$ | skip | $\eta_3^{c_3,r}$ | | |
| $(c_1+1)$ | draw | $\eta_3^{c_3+1,r}$ | s.t. | $\tilde{V}_3^1\left(\eta_1^{1,r},...,\eta_2^{K,r},\eta_3^{1,r},...,\eta_3^{c_3-1,r},\eta_3^{c_3+1,r}\right)<0$ |
| $\vdots$ | | | | |
| $(K)$ | draw | $\eta_3^{K,r}$ | s.t. | $\tilde{V}_3^K\left(\eta_1^{1,r},...,\eta_2^{K,r},\eta_3^{1,r},...,\eta_3^{c_3-1,r},\eta_3^{c_3+1,r},..,\eta_3^{K,r}\right)<0$ |

Calculate the simulated probability as

$$P^r = P\left(\tilde{V}_1^1<0\right)\times\Pi_{K>1,K\neq c_1}Pr\left(\tilde{V}_1^K<0\right)\times\Pi_{K,K\neq c_2}Pr\left(\tilde{V}_2^K<0\right)\Pi_{,K\neq c_3}Pr\left(\tilde{V}_3^K<0\right)$$

and

$$P_{GHK}=\frac{1}{R}\sum P^r$$

**$\Sigma_K$ and $\Gamma_K$**

The first variance I am concerned about is the following:

$$\Sigma(s) = cov\begin{pmatrix} u_1^1 \\ \vdots \\ u_1^K \\ u_2^1 \\ \vdots \\ u_2^K \\ u_2^{GR_2} \\ u_3^1 \\ \vdots \\ u_3^K \\ u_3^{GR_2} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & & & & \\ \Sigma_{21} & & & & \\ \Sigma_1^{GR_2} & \Sigma_2^{GR_2} & \Sigma^{GR_2,GR_2} & & \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} & & \\ \Sigma_2^{GR_3} & \Sigma_2^{GR_3} & \Sigma^{GR_3,GR_2} & \Sigma^{GR_3,GR_2} & \Sigma^{GR_3,GR_3} \end{pmatrix}$$

The matrix $\Sigma(s)$ depends on the chosen sequence $s$, as each sequence defines the company to which the guaranteed renewable plan en each period belongs to and the corresponding covariances. To normalize the utility of each period, I take differences with respect to alternative $K$, resulting in vector $\tilde{u}$ and covariance matrix

$$cov(\tilde{u}) = \Sigma_K(s) = \begin{pmatrix} \tilde{\Sigma}_{11} & & & & \\ \tilde{\Sigma}_{21} & & & & \\ \tilde{\Sigma}_1^{GR_2}(s) & \tilde{\Sigma}_2^{GR_2}(s) & \tilde{\Sigma}^{GR_2,GR_2}(s) & & \\ \tilde{\Sigma}_{31} & \tilde{\Sigma}_{32} & \tilde{\Sigma}_{33} & & \\ \tilde{\Sigma}_2^{GR_3}(s) & \tilde{\Sigma}_2^{GR_3}(s) & \tilde{\Sigma}^{GR_3,GR_2}(s) & \tilde{\Sigma}^{GR_3,GR_2}(s) & \tilde{\Sigma}^{GR_3,GR_3}(s) \end{pmatrix}$$

Similarly, I define $\tilde{\alpha}_k = \alpha_K - \alpha_K$, and the corresponding covariance

$$cov(\tilde{\alpha}) = \Sigma_K(s) = \begin{pmatrix} \tilde{\Gamma}_{11} & & & & \\ \tilde{\Gamma}_{21} & & & & \\ \Gamma_1^{GR_2}(s) & \tilde{\Gamma}_2^{GR_2}(s) & \tilde{\Gamma}^{GR_2,GR_2}(s) & & \\ \tilde{\Sigma}_{31} & \tilde{\Gamma}_{32} & \tilde{\Gamma}_{33} & & \\ \tilde{\Gamma}_2^{GR_3}(s) & \tilde{\Gamma}_2^{GR_3}(s) & \tilde{\Gamma}^{GR_3,GR_2}(s) & \tilde{\Gamma}^{GR_3,GR_2}(s) & \tilde{\Gamma}^{GR_3,GR_3}(s) \end{pmatrix}$$

To set the scale and level of the utility, I normalize $var\left(u_t^K\right) = 1$, and $\sigma_K = \sigma_{K-1} = 0$, which leads to $Var\left(\epsilon_K^t\right) = Var\left(\epsilon_{K-1}^t\right) = 1$.

**Cohort 0**

I observe individuals from cohort 0 for 2009, 2010, and 2011. In all these periods they have access to the GR plan as well as the spot plans. The correlation of errors across

alternatives in 2009 depends on the entire choice path since individuals entered the market, which unfortunately I do not observe. However, the GHK algorithm makes evident that the choice in 2009 provides enough information about the correlation of errors to estimate the choices in the following years.

Therefore, for cohort 0 I perform a "conditional" GHK estimation, taking their choice in 2009 as given and calculating the likelihood of their choice in 2010 and 2011 given their observed choice in 2009.

Note that cohort 0 in 2009 is comprised by individuals that either stayed in the their contract or switched within their company between 2008 and 2009. Since I cannot observe the GR contract for individuals that switched (and thus cannot construct the menu available to them in 2009), I restrict the sample of cohort 0 only to those that did not switch plans between 2009 and 2010, and thus picked the GR contract in 2009. I identify which are the ones that did not switch based on the tenure of their plans.

- For the sample defined above, I construct the variance-covariance $\Omega\left(k_0, k_1, k_2\right)$ considering the choice in 2009, 2010, and 2011, and the corresponding Cholesky decomposition $L\left(k_0, k_1, k_2\right)$.
- Draw the error terms for each option in 2009 consistent with $k_0$
- Draw the error terms for each option in 2010 and 2011 consistent with $k_1$ and $k_2$
- Calculate the simulated conditional probability, given the choice in 2009, as

$$P^r|c_1 = \Pi_{K\neq c_2} Pr\left(\tilde{V}_2^K < 0\right) \Pi_{K\neq c_3} Pr\left(\tilde{V}_3^K < 0\right)$$

And therefore

$$P_{GHK}|c_1 = \frac{1}{R}\sum P^r|c_1$$

## Average cost curves with preference heterogeneity

Here I provide a simple proof that preference heterogeneity decreases the average cost curve at any price. Assume that there are two preferences, $u_1(h,\epsilon)$ and $u_2(h,\epsilon)$ that generate the same demand curve but such that $u_1$ entails a entails a higher degree of preference heterogeneity (and normalize $u_1(0,0) = u_2(0,0) = 0$). Under my definition, this means that $\partial u_1/\partial\epsilon > \partial u_2/\partial\epsilon \geq 0$.

For a given preference $r = 1, 2$ I define $u_r(h_r^*, \epsilon) - P \equiv 0$. It follows that $h_2^* > h_1^*$ (both are assumed to exist). Let $f()$ be the marginal distribution of $\epsilon$ with corresponding cdf $F()$. The average cost at price $P$ is given by

$$AC\left(P\right) = \int_{-\infty}^{\infty} E_h(h|h > h^*(\epsilon, P))f(\epsilon)d\epsilon \tag{A.2}$$

which is increasing in $h^*$.