Essays on Health Insurance and Annuities

Essays on Health Insurance and Annuities

A dissertation presented

by

Mark Shepard

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

May 2015

Dissertation Advisors:
Professor David Cutler
Professor Ariel Pakes

Author:
Mark Shepard

Essays on Health Insurance and Annuities

Abstract

Insurance creates an important source of economic well-being by providing for beneficiaries in times of need. But because a variety of forces may inhibit the proper functioning of insurance markets, governments are deeply involved through regulation, subsidies, and direct provision of insurance. This dissertation studies insurance demand, supply, and the role of policy in two types of markets of direct interest to policymakers: health insurance and annuities. I highlight the importance of both traditional market failures (adverse selection and moral hazard) and less standard factors like limited competition (market power) and puzzlingly low insurance demand to influence insurance market outcomes.

In the first chapter, I study how health insurers compete in individual insurance markets like those established in the Affordable Care Act. I focus on the role of an increasingly important benefit: plans' networks of covered medical providers. Using data from Massachusetts' pioneer insurance exchange, I show evidence of substantial adverse selection against plans covering the most expensive and prestigious academic hospitals. Individuals loyal to the prestigious hospitals both select plans covering them and are more likely to use these hospitals' high-price care. Standard risk adjustment does not capture their higher costs driven by preferences for using high-price providers. To study the welfare implications of network-based selection, I estimate a

structural model of hospital and insurance markets and use the model to simulate insurer competition on premiums and hospital coverage in an insurance exchange. I find that with fixed hospital prices, adverse selection leads all plans to exclude the prestigious hospitals. Modified risk adjustment or subsidies can preserve coverage, benefitting those who value the hospitals most but raising costs enough to offset these gains. I conclude that adverse selection encourages plans to limit networks and star academic hospitals to lower prices, with the welfare implications depending on whether those high prices fund socially valuable services.

Chapter 2 also studies health insurance exchanges and the competitive effect of a policy design choice: how the level of subsidies is determined. In the Affordable Care Act exchanges and other programs, subsidies depend on prices set by insurers – as prices rise, so do subsidies. I show that these "price-linked" subsidies incentivize higher prices, with a magnitude that depends on how much insurance demand rises when the price of uninsurance (the mandate penalty) increases. To estimate this effect, I use two natural experiments in the Massachusetts subsidized insurance exchange. In both cases, I find that a $1 increase in the relative monthly mandate penalty increases plan demand by about 1%. Using this estimate, my model implies a sizable distortion of $48 per month (about 12%). This distortion has implications for the tradeoffs between price-linked and exogenous subsidies in many public insurance programs. I discuss an alternate policy that would eliminate the distortion while maintaining many of the benefits of price-linked subsidies.

Chapter 3 studies demand for annuities – insurance products that protect retirees against outliving their assets. Standard life cycle theory predicts that individuals facing uncertain mortality will annuitize all or most of their retirement wealth. Researchers seeking to explain why retirees rarely purchase annuities have focused on imperfections in commercial annuities –

including actuarially unfair pricing, lack of bequest protection, and illiquidity in the case of risky events like medical shocks. I study the annuity choice implicit in the timing of Social Security claiming and show that none of these can explain why most retirees claim benefits as early as possible, effectively choosing the minimum annuity. Most early claimers in the Health and Retirement Study had sufficient liquidity to delay Social Security longer than they actually did and could have increased lifetime consumption by delaying. Because the marginal annuity obtained through delay is better than actuarially fair, standard bequest motives cannot explain the puzzle. Nor can the risk of out-of-pocket nursing home costs, since these are concentrated at older ages past the break-even point for delayed claiming. Social Security claiming patterns, therefore, add to the evidence that behavioral explanations may be needed to explain the annuity puzzle.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my advisors David Cutler, Jeffrey Liebman, and Ariel Pakes for extensive guidance, comments, and support in writing this dissertation. I also thank Sonia Jaffe, my co-author for the second chapter. Without their help, this dissertation would not have been possible.

This research benefitted from helpful comments from and discussions with Katherine Baicker, Amitabh Chandra, Raj Chetty, Jeffrey Clemens, Keith Ericson, Martin Feldstein, Amy Finkelstein, Jerry Green, Jon Gruber, Kate Ho, Sonia Jaffe, Scott Kominers, David Laibson, Tim Layton, Robin Lee, Greg Lewis, Day Manoli, Tom McGuire, Joe Newhouse, Phil Oreopoulos, Daria Pelech, Amanda Starc, Karen Stockley, Rich Sweeney, Jacob Wallace, Tom Wollmann, Ali Yurukoglu, and participants in the Harvard Industrial Organization, Health Care Policy, and Labor lunches. I thank Che-Lin Su for helpful conversations and for making available a license for the Knitro optimization program.

I would like to thank the Massachusetts Health Connector – and particularly Michael Norton, Sam Osoro, Nicole Waickman, and Marissa Woltmann – for assistance in providing and interpreting the Massachusetts health insurance exchange data (used in chapters 1 and 2). I also acknowledge RAND for their cleaned version of the Health and Retirement Study, which I use in chapter 3.

I gratefully acknowledge Harvard's Lab for Economic Applications and Policy, whose financial support paid for the Massachusetts Connector data. I also acknowledge Ph.D. funding support from National Institute on Aging Grant No. T32-AG000186 (via the National Bureau of Economic Research), the Rumsfeld Foundation, and the National Science Foundation Graduate Research Fellowship.

Finally, I want to thank my wife, parents, sister, and the rest of my family and friends. Your encouragement and love throughout this process was the best support of all.

Chapter 1

Hospital Network Competition and Adverse Selection:

Evidence from the Massachusetts Health Insurance Exchange

1.1 Introduction

Public programs increasingly use regulated markets to provide health insurance. These markets (often called "exchanges") now cover more than 75 million people and comprise over $300 billion in public spending in U.S. programs including the Affordable Care Act (ACA), Medicare, and Medicaid. Exchanges can improve welfare by giving enrollees a choice among competing plans with varying levels of benefits and prices. But a perennial concern in insurance markets is adverse selection. Adverse selection occurs when consumer costs vary in ways that cannot be priced (e.g., because of pricing regulations), and high-cost consumers differentially choose generous plans. This raises the costs and prices of these plans, leading to inefficient sorting and potentially pushing insurers to drop generous benefits. Because of these concerns, exchanges attempt to address adverse selection by regulating plan benefits and risk adjusting payments to compensate plans attracting sicker people.[1] Whether selection is a problem even with these policies is an important question and a matter of active research.

As the ACA has regulated plans' covered medical services and patient cost sharing, insurers have differentiated on an alternate dimension: covered networks of hospitals and doctors. The first year of the ACA saw a large number of "narrow network" plans, which now comprise almost half of all exchange

[1] The ACA specifies "essential benefits" that must be covered and requires insurers to offer plans fitting into one of four generosity tiers. Risk adjustment uses observed demographics and medical diagnoses to predict each enrollee's expected cost (or "risk"). Plans attracting high-risk enrollees receive transfers from plans attracting low-risk enrollees intended to compensate for the expected cost difference between these plans' enrollees.

plans (McKinsey 2014).[2] These plans have generated controversy, including calls for broader network requirements, partly because they tend to exclude the most prestigious (and expensive) academic hospitals, or "star" hospitals.[3] An important issue for this debate is whether adverse selection discourages insurers from covering top hospitals, a question on which there is little direct evidence.[4]

In this paper, I show evidence of adverse selection against plans covering star hospitals through a channel that is theoretically distinct from the usual selection story and therefore poses a challenge for standard policies like risk adjustment. Typically, adverse selection in health insurance is associated with *sicker* people choosing generous plans. But in addition to medical risk, consumer costs can differ because of varying *preference* for using expensive medical providers when sick. Health care features substantial variation in prices and costs across providers, with star hospitals tending to be some of the most expensive (Ho 2009). Importantly, insurers typically cover the bulk of these cost differences – often thousands of dollars per hospital admission – with patients paying a small fee that differs little across in-network hospitals.[5] As a result, patients who choose star hospitals when sick are more costly to cover than patients who use less expensive alternatives. When a plan covers a star hospital, I find that it attracts consumers who incur higher costs because they use the star hospital for more of their care. Risk adjustment does not offset these costs, which are driven by provider choices rather than medical risk. As a result, coverage of the hospitals can unravel even if many people value access to them above cost.

In some ways, the implications of this alternate channel for adverse selection are standard – inefficient sorting and potentially unravelling of generous coverage. Because regulations prevent plan

[2] "Narrow networks" were defined as plans covering less than 70% of area hospitals. The McKinsey report documented a sharp rise in narrow network plans in 2014 relative to individual insurance markets in 2013.

[3] Ho (2009) uses the term "star hospitals" to refer to hospitals with strong reputations who use those reputations to bargain for high prices from insurers. In this paper, I use the term to refer to the top Massachusetts academic medical centers that *U.S. News* ranks most highly, which I show have high prices.

[4] An older literature studying the rise of HMOs found that HMOs attracted healthier customers than traditional insurance (Miller and Luft 1997). But HMOs differed from both via limited networks and a variety of managed care restrictions. By contrast, today's competition involves exclusively managed care plans.

[5] In the Massachusetts exchange setting I study, hospital copays were required to be identical across in-network hospitals, and coinsurance (in which patients pay a percentage of the overall bill) was not allowed. However, even in other markets that have coinsurance or "tiered" copays, insurers typically cover the bulk of cost differences.

premiums from varying based on consumer-specific use of star hospitals, plan sorting can be inefficient. For example, some people might want the option value of accessing a star hospital if they get a serious cancer. But to buy a plan covering it, they have to pool with people who use star providers for all their health care needs. Plans covering star hospitals differentially attract these high users, forcing them to raise premiums on all customers. The result is a cycle of adverse selection in which plans continually raise premiums and lose low-cost consumers, with this process either stabilizing or leading to full unravelling.

But selection on use of star hospitals also has non-standard implications for two reasons. First, it is fundamentally linked to the tradeoff between risk protection and moral hazard in health care. When insurers cover a star hospital, patients' costs increase because they can now use it (instead of cheaper alternatives) without paying the full bill. Coverage of star hospitals therefore generates a form of moral hazard.[6] What I find is that the people most likely to use star hospitals when covered (i.e. highest moral hazard) tend to select into plans that cover them. Thus, this pattern is an example of the idea of "selection on moral hazard" documented by Einav et al. (2013). Because moral hazard is involved, the welfare implications of unravelling are ambiguous and depend on whether the lost value of access exceeds or falls short of the cost savings. It also implies that policies to reduce moral hazard (e.g., higher "tiered" copays for expensive hospitals) may also reduce adverse selection. However, as long as the higher copays do not cover the full extra cost of star hospitals, some moral hazard and adverse selection is likely to remain as a tradeoff to providing risk protection.

A second difference from the standard analysis is that the selection is linked to a service (care at star hospitals) whose prices are not set competitively. Instead, hospital prices are set in negotiations with insurers driven by market power. This market power complicates policy responses to selection. For instance, a mandate to cover the hospitals – one standard response to selection – would be problematic because it would give star hospitals extreme power to raise prices. Other policies that subsidize plans

[6] The moral hazard terminology can be confusing because contract theory typically refers to moral hazard as a hidden or non-contractible agent action that is costly to the principal. Here, the action is using a star hospital, and while it is not hidden, regulation prevents writing an insurance contract whose price varies with star hospital use.

covering star hospitals would have a conceptually similar (but less extreme) effect. A key question is whether government is willing to subsidize star hospitals' high prices to ensure that exchange enrollees have access to them. Because these academic hospitals' high prices partly fund teaching, medical research, and care for the poor, the answer to this question is not obvious.

To study these issues, I use data from a market that was a key precursor to the ACA: Massachusetts' subsidized insurance exchange.[7] The Massachusetts exchange provides a nice setting for studying networks and selection. Covered services and patient cost-sharing rules are fixed by regulation for all plans, so provider networks are the only significant benefit that differs across plans. Further, the exchange has excellent administrative data on all consumers' plan choices and insurance claims. These detailed data let me estimate a flexible model of demand and costs to capture heterogeneity driving adverse selection. Finally, Massachusetts has a clear set of star hospitals: the Partners Healthcare System. Partners is centered around Mass. General and Brigham & Women's hospitals, which are consistently ranked by *U.S. News & World Report* as the top two hospitals in Massachusetts and among the top 10 hospitals in the nation. Consistent with state reports (e.g., Coakley 2013), I find that Partners hospitals are extremely expensive. I estimate that their risk-adjusted prices per admission are almost twice the average of other hospitals and over $5,000 (or 33%) more than the average of other academic medical centers.

I start by testing for adverse selection against plans covering Partners using reduced form methods. I show that these plans attract a group who appear to strongly prefer Partners: people who have used Partners hospitals in the past, either for inpatient or (more often) outpatient care. Compared to an average enrollee, these past Partners patients are (1) 15% costlier even after risk adjustment, (2) 80% more likely to select a plan that covers Partners, and (3) more than twice as likely to use Partners for subsequent hospitalizations. These facts suggest that Partners patients are loyal to their preferred hospitals

[7] Massachusetts' subsidized exchange (which I study) is distinct from its unsubsidized exchange that has been studied by Ericson and Starc (2013, *forthcoming*). The only research I am aware of using the subsidized exchange is by Chandra, Gruber, and McKnight (2010, 2011, *forthcoming*) studying the effect of the individual mandate on market enrollment and the effects of cost-sharing changes in 2008 on utilization of care.

and select plans based on their desire to use Partners in the future.[8] I find that this loyalty to past-used hospitals is true more broadly across all hospitals in my data, suggesting that it is a general phenomenon likely to drive plan selection in health insurance markets.[9]

I next study how this selection played out in a case in 2012 when a large plan dropped Partners and several other hospitals from its network. This type of network change provides a natural source of evidence that past research has rarely studied. Consistent with selection, I find a sharp increase in enrollees switching away from the plan that dropped Partners, driven almost entirely by a nearly 40% switching rate by Partners patients. While exchange enrollees typically switch plans less than 5% of the time (c.f. Handel 2013), they are much more willing to switch plans when a preferred provider is dropped. Using my model (discussed below) to decompose the plan's large risk-adjusted cost reduction after dropping Partners, I find that selection can account for between 36% and 50% of the change.

The reduced form analysis suggests the importance of adverse selection based on hospital networks. However, both the welfare and policy implications of this selection are less clear. To investigate these issues, I estimate a structural model of consumer preferences and insurer costs. The model – which follows a structure used in past work (Ho 2009; Capps, Dranove, and Satterthwaite 2003) – consists of three pieces: (1) a hospital demand system capturing hospital choices under different plan networks, (2) an insurance demand system capturing plan choice patterns, and (3) a cost system estimated from the insurance claims data. Relative to past work, the main innovation is to allow for detailed preference heterogeneity and use the individual-level data to capture the correlations among hospital choices, plan preferences, and costs – which are critical for adverse selection. In addition, I pay special attention to the identification of the premium and network coefficients in plan demand, using only within-

[8] Past Partners patients also appear to be unobservably sick, since they have above-average risk-adjusted costs even in plans that do not cover Partners. However, their costs are much higher in plans that cover Partners.

[9] It is less clear how much of this loyalty is driven by state dependence (a preference for hospitals used in the past) versus more durable preference heterogeneity. Both are valid channels for the short-run adverse selection results I find. But state dependence implies lower long-run welfare impacts of unraveling of Partners coverage, since patients need only incur a one-time cost of switching providers. Disentangling the roles of state dependence versus heterogeneity in loyalty to providers is an important question for future research.

plan variation to identify them. For premiums, I use variation across consumers driven by Massachusetts' subsidy rules. For networks, I use variation in how different consumers value a given hospital network.

My demand estimates imply that individuals value both lower prices and better hospital networks, though with significant heterogeneity in this tradeoff. Consistent with the reduced form evidence, I find that past patients of a hospital are particularly likely to use it again and to select plans that cover it. These effects are particularly strong for past patients of Partners hospitals. Thus, the demand estimates are consistent with significant selection based on coverage of the prestigious Partners hospitals.

I use the model to study the competitive, welfare, and policy implications of network-based selection. I simulate equilibrium in a game where insurers first choose whether or not to cover the Partners hospitals (holding fixed other hospital coverage) and then compete in a static Nash-in-prices game. I model exchange policies similar to those in the ACA, which differ in several ways from those used in Massachusetts. The key limitation of these simulations is that they hold hospital prices fixed at their observed values, not modeling hospital-insurer price bargaining. At Partners' observed high prices, I find that adverse selection leads all insurers drop them from network. As in the reduced form results, a plan deviating to cover Partners loses money both through higher costs for its existing enrollees (moral hazard) and by attracting high-cost enrollees who particularly like Partners (adverse selection). I use the model to decompose the adverse selection into traditional selection on levels of cost and selection on use of Partners. Of the 27% higher risk-adjusted costs for the group that most highly values Partners, about 60% is a higher level of risk-adjusted costs even in a plan that does not cover Partners, and 40% is larger cost increases when a plan covers Partners. Thus, both selection on unobserved health risk and selection on likelihood to use Partners appear to be quantitatively important.

Finally, I use my model to analyze policy changes to address adverse selection. I find that modified risk adjustment and differential subsidies for higher price plans can reverse the unraveling. These policies give plans a greater incentive to cover Partners even though doing so requires raising prices and attracting high-cost enrollees. However, I highlight two tradeoffs. First, covering Partners creates moral hazard. When Partners is covered, not only people who value it highly can use it but also

those who value it little. My model's estimates imply that past Partners patients have greater value of access than costs, but other enrollees on average do not. Because the latter group is much larger, I find a net decrease in social surplus when the government changes policy to encourage Partners coverage.

A second tradeoff of these policy changes is that they encourage both insurers and Partners hospitals to raise prices. My current model does not capture the higher Partners prices (which are held fixed). But I find important increases in insurance prices and markups, leading to a government-funded increase in insurer profits. This analysis aligns with recent work finding that adverse selection leads plans to reduce markups in imperfectly competitive markets (Starc 2014; Mahoney and Weyl 2014). Adverse selection gives insurers an incentive to keep prices low to attract low-cost consumers. Policies that offset this effect encourage plans to raise price markups. In exchanges, higher plan prices mean higher government subsidies, which are set based on these prices.

These results suggest that standard policies used to address adverse selection (e.g., risk adjustment and subsidies) are less effective at improving welfare with selection based on star hospital use. These policies compensate insurers for attracting high-cost enrollees but do not address the fundamental issue of efficiently sorting patients across hospitals. Policies that address this sorting challenge directly – e.g., higher "tiered" copays for high-price hospitals or payment incentives for doctors to steer patients to lower-cost hospitals (see Song, et al. 2011; Ho and Pakes 2014) – may be more effective and are a fruitful subject for future research.

The remainder of this paper is organized as follows. Section 1.2 outlines a simple model that captures the main intuition for network-based selection. Section 1.3 presents background on the Massachusetts exchange and hospital market and introduces the data. Section 1.4 shows reduced form results, and Sections 1.5-1.6 present the structural model and estimates. Section 1.7 analyzes the model's implications for adverse selection, and Section 1.8 presents the equilibrium and counterfactual policy simulations. The final section concludes.

1.2 Basic Theory

Economists have long known that insurance markets are subject to adverse selection. When consumers' costs vary but firms cannot observe or price based on this variation, market competition can be inefficient. Insurers may have incentives to cut benefits to avoid high-cost types (Rothschild and Stiglitz 1976) and in the extreme, the market can unravel entirely (Akerlof 1970) or to the minimum quality option (Cutler and Reber 1998). Most past theories have focused on selection based on medical risk (e.g., having an expensive illness). As a result, the standard policy to address selection is risk adjustment – transfers to plans attracting sicker people to offset their higher expected costs.

In this section, I present a simple model to illustrate an alternate source of adverse selection. This selection arises because certain "star" hospitals (usually academic medical centers) negotiate high prices that insurers cover (i.e., do not pass onto sick patients) when they include the hospitals in network. Whether insurers *should* cover star hospitals depends on the classic tradeoff between insurance protection (against the risk of needing advanced care) and moral hazard. What I show is that whether profit-maximizing insurers *will* cover star hospitals is also influenced by adverse selection. Consumers who strongly prefer using star hospitals will both be high-cost and more likely to choose plans covering the hospitals. Standard risk adjustment is unlikely to capture this form of selection, giving insurers an incentive to drop the star hospitals. I use the model to analyze when this unravelling is inefficient and to discuss the policy implications, which are complicated by insurer and hospital market power.

Simple Model

Consider a model with two hospitals. One hospital *S* is a "star" hospital, which can use its reputation to bargain for a high price τ_S per visit. A second "non-star" hospital *NS* has a lower price, $\tau_{NS} < \tau_S$. I assume that the star hospital's price is above its marginal cost mc_S and that its markup above costs exceeds the markup for the non-star hospital: $\tau_S - mc_S > \tau_{NS} - mc_{NS}$. Consumers differ in two

dimensions: (1) their *risk* (or probability) of hospitalization, $r_i \in [0,1]$, and (2) their relative *value* (in dollar terms) of using hospital *S* when sick, v_i^S. If patient copays are identical across hospitals (as in the Massachusetts exchange), a patient with access to both hospitals will choose *S* if $v_i^S \geq 0$ and *NS* if $v_i^S < 0$.[10] Let $s_i \equiv 1\{v_i^S \geq 0\}$ be an indicator of consumer *i*'s preference for *S*. Defining $\Delta\tau \equiv \tau_S - \tau_{NS}$, under these assumptions, an insurer's expected total cost for consumer *i* is:

$$C_{ij} = \begin{cases} r_i \cdot \tau_{NS} & \text{if a plan does not cover } S \\ r_i(\tau_{NS} + s_i \cdot \Delta\tau) & \text{if a plan covers } S \end{cases}$$

Consider an insurance market with two plans (*A* and *B*). For simplicity, plan *B* is a non-strategic actor who covers only hospital *NS* and sets its premium at average cost.[11] Plan *A* also covers hospital *NS* but chooses whether to cover *S* and strategically sets its premium to maximize profits. I assume consumers' utility for a plan not covering *S* is normalized to 0. Their (monetized) utility for plan *A* if it covers *S* equals their expected value of access to *S*: $U_{iA}^{CoverS} = r_i \cdot s_i \cdot v_i^S$. The key assumption here is that consumers' utility for a plan covering the star hospital (U_{iA}^{CoverS}) is correlated with their likelihood of using that hospital ($= r_i \cdot s_i$).

Following Massachusetts' rules, assume that each plan *j* sets a single premium P_j that cannot vary across consumers.[12] Although prices cannot vary, the exchange risk adjusts payments based on consumer observables Z_i so a plan in total receives $P_j + RA(Z_i)$ for consumer *i*.[13] The risk adjustment

[10] Negative values capture the fact that hospital *NS* may be closer and more convenient for many patients.

[11] If instead plan *B* set its premium to maximize profits, all of the basic adverse selection intuition of the model would carry through, but I would need to consider how the policies I discuss below would affect its markup.

[12] Assume that any subsidies are a flat amount for both plans so that the difference in consumer premiums equals the difference in the prices plans receive.

[13] Risk adjustment methods vary, and in general, the exchange could also make risk adjustment a function of prices. This was done in Massachusetts so that $RA_i^{Mass} = (\varphi(Z_i) - 1)P_j$, where $\varphi(Z_i)$ was a risk score and the plan's total payment was $\varphi(Z_i)P_j$. The ACA's risk adjustment is closer to the simple model, since its transfer is based on an enrollee risk score and the average price in the market.

function is set to offset a consumer's expected extra costs, so $RA(Z_i) = E(c_{ij} \mid Z_i) - \bar{c}$ (where $\bar{c}$ is overall average cost). If risk adjustment captured costs perfectly, a plan's profit margin would be a constant $P_j - \bar{c}$ for all consumers. However, heterogeneous preferences for hospitals can confound risk adjustment in two ways. First, consumers selecting a plan covering *S* may have higher *unobserved risk* – a problem of imperfect risk adjustment. Second, consumers may select based on *preference* for hospital *S*, an independent source of cost variation that standard risk adjustment does not capture.

Table 1.1 Simple Example

	Expected Enrollee Costs		Avg. Cost (for risk adj.)
	Prefer Hospital *NS*	Prefer Hospital *S*	
Low Risk	$r_L \cdot \tau_{NS}$ **\$500**	$r_L \cdot \tau_S$ **\$1,000**	$r_L\left(\tau_{NS} + \bar{s}_L \cdot \Delta\tau\right)$ **\$750**
High Risk	$r_H \cdot \tau_{NS}$ **\$5,000**	$r_H \cdot \tau_S$ **\$10,000**	$r_H\left(\tau_{NS} + \bar{s}_H \cdot \Delta\tau\right)$ **\$7,500**

Parameter Assumptions: $r_L = 0.05$, $r_H = 0.50$, $\tau_{NS} = \$10\text{k}$, $\tau_S = \$20\text{k}$, $\bar{s}_L = \bar{s}_H = 0.5$

To see this, suppose that the exchange observes risk perfectly ($Z_i = r_i$) and consider a simple example with two risk types and specific parameters shown in Table 1.1. The table shows costs for each enrollee type in plan *A* if it covers hospital *S*. Because risk adjustment only captures average costs for each risk, the costs of consumers who prefer *S* exceed the risk-adjusted payment, and they are therefore less profitable to cover. Notice also that the interaction between risk and preference is important. For *S*-preferring types, risk-adjusted payments are too low by \$250 for low risks and by \$2,500 for high risks. This follows from high risks using the expensive hospital more and therefore having a *differential* cost increase in a plan covering *S*. I argue below that these differential costs induce inefficient sorting, since a premium increase sufficient to pay for high risks' extra costs may cause low risks to leave the plan. This inability of a homogenous premium to efficiently sort consumers with varying differential costs is related to a point made by Bundorf, Levin, and Mahoney (2012) and Handel, Hendel, and Whinston (2013).

Equilibrium and Welfare Implications

To study the equilibrium, I simplify by assuming the exchange risk adjusts based on the cost of *NS*, so $RA(Z_i) = \left(E\left(r_i \mid Z_i\right) - \overline{r}\right)\tau_{NS}$.[14] Notice that I again allow for risk to be observed imperfectly. Define the risk adjustment error as $e_i \equiv \left(r_i - E\left(r_i \mid Z_i\right)\right)\tau_{NS}$, which is positive for the unobservably sick and negative for the unobservably healthy. Under these assumptions, plan *B* always charges a price (assumed to equal average costs) of $P_B = \overline{r} \cdot \tau_{NS} + \overline{e}_B$, where $\overline{e}_B$ is plan *B*'s enrollees average risk adjustment error. If plan *A* does not cover hospital *S*, the two plans are undifferentiated. Both charge $P_A^0 = P_B^0 = \overline{r} \cdot \tau_{NS}$, attract a random sample of enrollees, and earn zero profits. If plan *A* does cover *S*, it can raise its price and potentially earn profits. However, by doing so, it also increases its costs. Defining the price difference as $\Delta P_A \equiv P_A - P_B$, plan *A* will cover *S* if:

$$\text{Cover } S \text{ if:} \quad \max_{\Delta P_A} \left\{ \underbrace{\Delta P_A}_{\text{Price Difference}} - \underbrace{E\left[r_i \Delta\tau \mid r_i v_i^S > \Delta P_A\right]}_{\text{Insurer Cost Increase}} - \underbrace{\left(\overline{e}_A - \overline{e}_B\right)}_{\text{Unobs. Risk Selection}} \right\} \geq 0 \qquad (1.1)$$

Adverse selection on use of the star hospital shows up in (1.1) in the conditional expectation – the people who choose *A* tend to be those who most value (and use) hospital *S*. More traditional selection on unobserved risk occurs through the $\overline{e}_A - \overline{e}_B$ term. If people preferring hospital *S* are unobservably sicker $(\overline{e}_A - \overline{e}_B > 0)$ – which I find to be true empirically – this is an additional channel for adverse selection.[15]

Now compare this equilibrium to the efficient outcome. For now, I assume that hospital markups are pure transfers, so the true social cost of using each hospital is mc_S and mc_{NS}.[16] The first-best outcome is for patients to choose the plan covering *S* if and only if $v_i^S \geq \Delta mc$. This first-best is unlikely to be attainable. Even if plan *A* covers *S*, consumers will choose plan *A* and get access to *S* only if $r_i v_i^S \geq \Delta P_A$.

[14] This assumption does not change any of the intuition of the results but makes the math much simpler.

[15] In theory, these individuals could be unobservably healthier, creating a source of advantageous selection that offsets some of the adverse selection on use of the star hospital.

[16] As I discussed in the introduction, how to value the star hospital's markup is not obvious. If it funds socially valuable services like teaching and medical research, it might have more than the valuation at cost I assume here.

These conditions cannot coincide as long as there is consumer risk heterogeneity. With heterogeneous risks, some low-risk types who would highly value star hospital access if they did become sick will not choose plan A because doing so involves pooling with high-risk types who frequently use hospital S. Conversely, some high-risk types will inefficiently select into plan A. These errors reflect inefficient sorting with a single premium and heterogeneous costs.

Even if the first-best is unattainable, we can ask how plan A's incentives compare to second-best efficiency given that consumers sort based on a single premium difference. It is socially optimal for plan A to cover hospital S (at some differential premium ΔP) if:

$$\text{Efficient to Cover } S \text{ if:} \quad \max_{\Delta P} \left\{ \underbrace{E\left(r_i v_i^S \mid r_i v_i^S > \Delta P\right)}_{\text{Cons. Value of Star Hospital}} - \underbrace{E\left(r_i \cdot \Delta mc \mid r_i v_i^S > \Delta P\right)}_{\text{Social Cost Increase}} \right\} \geq 0 \qquad (1.2)$$

Comparing conditions (1.1) and (1.2), we can say that *the plan has too little incentive to cover the star hospital* due to three factors:

(a) Selection on unobserved risk ($\bar{e}_A - \bar{e}_B > 0$), which the plan treats as a disincentive to cover S even though it is not a social cost;

(b) The star hospital's higher markup ($\Delta\tau > \Delta mc$), which makes the private cost of covering S exceed the social cost; and

(c) Consumer surplus for the plan when it covers hospital S: Plan A's price increase equals the value to the marginal consumer (for whom $r_i v_i^S = \Delta P_A$), not the larger gain to the average consumer. This corresponds to the standard monopoly quality problem of Spence (1975).[17]

Thus, the plan will sometimes fail to cover the star hospital when it would be socially efficient to do so. This is the first inefficiency. Notice that selection on preference for using hospital S enters this inefficiency indirectly – by exacerbating the extra cost associated with the star hospital's markup.

[17] Unlike Spence (1975), the monopoly quality distortion can be signed here because coverage of hospital A is a discrete good.

A second inefficiency is that even if the plan covers hospital S, *its premium will be inefficiently high*. To maximize social welfare defined by (1.2), the optimal premium difference is $\Delta P^* = \Delta c \cdot E\left(r_i \mid r_i v_i^S = \Delta P^*\right)$. But for the plan to not lose money when covering *S*, we know by (1.1) that $\Delta P_A \geq \Delta\tau \cdot E\left(r_i \mid r_i v_i^S > \Delta P_A\right) + \left(\overline{e}_A - \overline{e}_B\right)$. The plan's premium will be too high because of selection on unobserved risk ($\overline{e}_A - \overline{e}_B > 0$), the star hospital's markup ($\Delta\tau > \Delta c$), and selection on use of the star hospital (since $E\left(r_i \mid r_i v_i^S > \Delta P\right)$ exceeds $E\left(r_i \mid r_i v_i^S = \Delta P\right)$).

To summarize, under standard risk adjustment, several forces combine to give the plan a greater incentive to exclude the star hospital than is socially optimal. And if the plan does cover the star hospital, its price will be too high partly because it selects the highest-cost enrollees. Both inefficiencies suggest the potential for policy changes to improve the equilibrium.

Market Power and Policy Implications

How should exchange policy respond to these inefficiencies? Two natural policies are to modify subsidies or risk adjustment to encourage plan *A* to cover the star hospital and to reduce its relative premium if it does. In doing this, policymakers should consider potential negative competitive side effects. This consideration is important because when insurers have market power, adverse selection can have the positive effect of lowering price markups (Starc 2014; Mahoney and Weyl 2014). Policies used to offset selection may therefore lead insurers to raise prices and profits. In exchanges, higher prices are a public policy problem because subsidies are set based on prices[18] – so higher prices raise government costs, with an associated excess burden of taxation.

To see how this works, consider the condition for plan *A*'s profit-maximizing price:

[18] Specifically, ACA subsidies are linked to the price of the second-cheapest silver tier plan. In Massachusetts, subsidies were linked to the price of the cheapest plan. These policies ensure that the cheapest plans will be affordable to consumers even if plan prices are higher than expected.

$$P_A^* = \underbrace{\left[\bar{r}\tau_{NS} + \bar{r}_A\Delta\tau + \bar{e}_A\right]}_{\text{Risk-Adjusted Avg. Cost}} + \underbrace{\frac{1 - \frac{d\bar{r}_A}{dP_A}\Delta\tau - \frac{d\bar{e}_A}{dP_A}}{\eta_{D_A}}}_{\text{Plan Price Markup}} \tag{1.3}$$

where $\bar{r}_A$ is the average risk of people enrolling in plan A and $\eta_{D_A} \equiv -\frac{1}{D_A}\frac{dD_A}{dPrem_A}$ is the premium semi-elasticity of demand. As is standard with imperfect competition, the plan's price equals its risk adjusted costs plus a markup. Adverse selection (after risk adjustment) affects both of these terms. It first implies that plan A attracts worse risks (higher $\bar{r}_A$ and $\bar{e}_A$), raising its risk-adjusted costs. But it also implies that the plan has an incentive to hold down its markup, since by raising price it attracts even worse risks ($\frac{d\bar{r}_A}{dP_A} > 0$ and $\frac{d\bar{e}_A}{dP_A} > 0$). Policy changes that offset adverse selection push the other way: lowering risk-adjusted costs for the adversely selected plan and raising its price markups. The net effect on its price is ambiguous. In Section 1.8, I analyze the effect of specific risk adjustment and subsidy changes to offset selection using my empirically estimated model.

A second competitive effect comes from the star hospital's market power. So far, I have been assuming that hospital prices are exogenous, but now suppose that the price of S can adjust, either by adjusting costs or markups. Another way to interpret the hospital coverage condition (1.1) is as a *maximum price* that S can charge while still being covered:[19]

$$\tau_S^{Max} = \tau_{NS} + \frac{P_A^* - \bar{r}\tau_{NS} - \bar{e}_A}{\bar{r}_A} \tag{1.4}$$

Adverse selection raises $\bar{r}_A$ and $\bar{e}_A$, both of which lower the maximum price the star hospital can charge, although increases in plan A's price push the other direction. If the first effects dominate, adverse selection will spur the star hospital to cut its price, while policies to offset selection will do the opposite.

[19] This would be the hospital's price in a Nash bargaining model in which the star hospital had all of the bargaining power. In this simple model, the hospital might want to charge a lower price because by doing so, plan A can charge a lower premium and attract more consumers who use hospital S. In a more realistic model with many hospitals and non-hospital costs, this effect would be sufficiently small that the maximum bargaining price would hold.

1.3 Massachusetts Exchange Background and Data

I study the subsidized Massachusetts health insurance exchange – called Commonwealth Care, or CommCare. Created in Massachusetts' 2006 health insurance reform, it operated from November 2006 to December 2013, after which it shifted form to comply with ACA rules. Like the ACA exchanges, CommCare offered subsidized coverage to low-income people (below 300% of poverty) not eligible for employer-sponsored insurance or other public programs. Higher-income and otherwise ineligible people could buy unsubsidized plans in a separate exchange ("CommChoice") – a market studied by Ericson and Starc (2012, 2013).[20] In CommCare, enrollees could choose among competing private plans in a centralized marketplace. Over the 2010-2013 period I focus on, the exchange had five competing insurers and about 170,000 enrollees per month. This size makes it comparable to a very large employer insurance plan but still small relative to Massachusetts' overall population of 6.6 million.

CommCare's regulator was aware of the possibility of adverse selection and took several steps to counteract it. Generous subsidies (covering over 90% of premiums on average) and an insurance mandate encouraged broad participation to avoid just the sick from selecting into the market. To address adverse selection within the market, CommCare used several policies. First, it standardized nearly all benefits, requiring plans to follow pre-specified covered service and cost-sharing rules. The only major benefit left flexible were provider networks.[21] Second, the exchange used risk adjustment to compensate plans for attracting the sick. This worked by assigning each enrollee a risk score φ_i based on his expected costliness. A plan with price P_j would receive $\varphi_i P_j$ for covering enrollee i. CommCare used sophisticated diagnosis-based methods to set risk scores, similar to those in the ACA and Medicare.[22]

[20] The ACA differs from Massachusetts by pooling subsidized and unsubsidized enrollees into the same market. The ACA also differs in subsidizing people up to 400% of poverty but excluding people below 133% of poverty, who will be eligible for Medicaid in states expanding the program.

[21] In addition, insurers could flexibly set two more minor benefits: (1) a few "extra benefits" like gym membership discounts, and (2) prescription drug formularies, provided they covered at least two drugs per class.

[22] One difference was that Massachusetts used *prospective* risk adjustment (which uses only past observed diagnoses), whereas the ACA uses a *concurrent* method incorporating diagnoses observed in the current year.

Notably, risk adjustment does not incorporate people's past provider choices, since it is not intended to capture people's likelihood to use expensive hospitals. Although there is conflicting evidence on how well risk adjustment has worked in Medicare (see Brown et al. 2011; Newhouse et al. 2015), these methods nonetheless represent the state of the art.[23]

A final policy affecting adverse selection was Massachusetts' subsidy rules. Unlike the ACA which applies a flat subsidy to all plans, CommCare differentially subsidized higher-price plans. In theory, these "marginal" subsidies address selection by narrowing price differences and encouraging plans to compete less on price and more on quality. In practice, CommCare's subsidies were not intended to address selection nor were they well designed to do so. The main marginal subsidies were for enrollees below poverty, for whom federal Medicaid rules required all plans to be fully subsidized (i.e., zero consumer premiums). Because of this, the exchange had to set a maximum price to prevent arbitrary price increases. When binding, this maximum price in turn discourages plans from covering expensive hospitals, since they cannot increase their price when they do so.

Whether adverse selection is a significant concern, despite these policy steps, is an unknown but important question. Like CommCare, the ACA has generous subsidies and a mandate, risk adjustment and reinsurance, and benefit standardization relative to the pre-ACA market (though not as stringent as CommCare). A key question is whether provider networks, one of the few flexible benefits, create residual selection problems. Because of its network variation in isolation of other benefit differences, CommCare is a nice setting to study the effects of limiting provider networks.

Massachusetts' method meant that risk adjustment for new enrollees could only be based on age and sex, since no past diagnoses information was available. In practice, I find the selection results hold just as strongly even for the subsample for whom diagnosis-based risk adjustment was used.

[23] Like the ACA, CommCare also used reinsurance for very high-cost enrollees (above $150,000 per year), but this accounted for just 0.03% of enrollees and about 1% of costs.

Provider Networks and Plan Background

The Massachusetts provider market in which CommCare operates includes a large number of providers – about 80 hospitals and 28,000 active doctors (AAMC 2013). But these providers are organized into a few dominant provider systems, most centered around large academic hospitals. Table 1.2 shows statistics on general acute care hospitals and hospital systems based on my CommCare data.[24] The top panel lists the most expensive hospitals based on the average per-admission cost to insurers. I also show a risk-adjusted price measure (which adjusts for patient severity, as I describe in Section 1.6.1) and the average severity of their patients (where the mean across all patients in all hospitals is normalized to 1.0). The bottom panel shows the most used hospital systems.

All of the five most expensive hospitals are academic medical centers – a designation given by the Massachusetts government to the state's six most significant academic hospitals. These academic hospitals both have high prices and treat sicker than average patients (severity > 1.0). But even among academic hospitals, the Partners Healthcare System is unique. Partners is both the most-used system and includes the two most expensive hospitals – Mass. General Hospital and Brigham & Women's Hospital, both in Boston.[25] These two are clear examples of what Ho (2009) called "star hospitals." *U.S. News & World Report's* rankings list them as the top two hospitals in Massachusetts, and they have consistently been listed among the top 10 hospitals in the U.S. (with Mass. General ranking #1 in 2013-14). The Partners hospitals' high prices have been repeatedly documented (see Coakley 2013; CHIA 2013) and have sparked anti-trust investigations by the U.S. Department of Justice and the Mass. Attorney General.

[24] Because I exclude specialty hospitals and merge together different campuses of the same hospital (which are often not separately identified in the claims data), my data has a total of 64 unique hospitals.

[25] Partners also included (as of 2012) five community hospitals in Eastern Massachusetts and network including more than 1,100 primary care physicians (BCBS Foundation of Massachusetts 2013).

Table 1.2 Massachusetts Hospital and Plan Network Statistics

Panel A: Most Expensive Hospitals

	Hospital	System	Hospital Type	Per-Admit Insurer Cost Stats: Total Payment	Risk-Adj. Price	Severity Index
1	Brigham & Women's	Partners	Acad. Med Ctr	$23,278	$20,474	1.12
2	Mass. General	Partners	Acad. Med Ctr	$21,428	$19,550	1.09
3	Boston Med. Ctr.	BMC	Acad. Med Ctr	$16,850	$15,919	1.05
4	Tufts Med. Ctr.	Tufts	Acad. Med Ctr	$15,328	$14,038	1.10
5	UMass Med. Ctr.	UMass	Acad. Med Ctr	$14,941	$14,111	1.07
6	Charlton Memorial	Southcoast	Community	$14,411	$14,210	1.03
7	Baystate Med. Ctr.	Baystate	Teaching	$13,715	$12,223	1.11
8	Lahey Clinic	Lahey	Teaching	$13,430	$11,742	1.13
9	Beth Israel Deaconess	CareGroup	Acad. Med Ctr	$12,971	$11,787	1.08
10	St. Vincent	Vanguard	Teaching	$11,881	$11,455	1.03
	All Other Hospitals	---	---	*$8,232*	*$8,585*	*0.95*

Panel B: Most Used Hospital Systems

	Hospital/System	Share of Admissions	Num. Hospitals	Hosp. Beds	Payment per Admission
1	Partners	13.6%	7	2,488	$18,016
2	Steward	12.2%	10	1,627	$7,695
3	UMass	8.5%	5	947	$13,140
4	CareGroup	7.7%	5	1,088	$11,481
5	Boston Med. Ctr.	6.6%	1	474	$16,850
6	Southcoast	5.7%	3	815	$12,642
7	Baystate	4.9%	3	774	$12,361
	All Others	*40.7%*	*30*	*5,391*	*$8,649*

NOTE: These tables show the most expensive hospitals and the most used hospital systems in the Massachusetts exchange (CommCare) data. The most expensive hospitals in Panel A are ranked by the average insurer payment per in-network hospital admission. I also show averages for risk-adjusted prices and an index of patient severity, both of whose estimation is described in Section 1.6.1. Hospital type is a classification defined by the Massachusetts state government. Panel B shows the most used hospital systems (by share of admissions in the Massachusetts data), including some other statistics on the hospitals. In this table, payment per admission is again the average insurer payment per in-network hospital admission.

Table 1.3 shows hospital network coverage for each of the five exchange plans from 2009-2013, based on information the exchange posted online for enrollees to review. Because of its unique status, I focus attention on coverage of Partners, listing only the statewide share of all other hospitals covered (weighted by number of beds). Plans always cover/exclude the two Partners academic medical centers in tandem. Coverage of Partners' five community hospitals is more flexible but correlates strongly with coverage of the academic medical centers. As of 2011, three plans covered Partners – Network Health, Neighborhood Health Plan (NHP), and CeltiCare. The other two plans had specific reasons not to cover Partners: Boston Medical Center (BMC) plan is owned by a competitor academic hospital, and Fallon is centered in central-Massachusetts and does not offer coverage in most of Boston where Partners is based.

Over time, several changes occurred in Partners coverage. In 2012, the exchange introduced a limited choice policy that gave the lowest-price plans exclusive access to new below-poverty enrollees (for whom all plans were free). In response to this incentive to price low, CeltiCare and Network Health cut prices by 11% and 15%, respectively, to become the two cheapest plans. While CeltiCare already had low costs (partly through its very limited coverage of non-Partners hospitals), Network Health needed to cut costs to make possible this price reduction. To do so, it dropped almost all of the Partners system plus eight other hospitals from network.[26]

As I show in Section 1.4, these changes helped Network Health reduce per-member costs by 21%. But a significant portion came from high-cost enrollees leaving Network Health and shifting to NHP and CeltiCare. In response, CeltiCare excluded the main Partners primary care physicians at the start of fiscal 2014, though it continued to cover the hospitals for care if patients were referred through an

[26] Network Health had covered all of Partners in 2011 and dropped all except two small hospitals on the islands of Nantucket and Martha's Vineyard. It also excluded the affiliated Partners physician groups. The eight non-Partners hospitals dropped were mostly community hospitals but included one less prestigious academic medical center (Tufts), one teaching hospital (St. Vincent).

Table 1.3 Hospital Network Coverage by Exchange Plans

Hospital Coverage by Year and Plan

Hospital Group	2009	2010	2011	2012	2013
	Boston Medical Center Plan (BMC)				
Partners: Academic Hospitals	No	No	No	No	No
Other Partners Hospitals	2/5	1/5	1/5	1/5	1/5
Share of Non-Partners Hospitals	*88%*	*92%*	*92%*	*92%*	*92%*
	Network Health				
Partners: Academic Hospitals	Yes	Yes	Yes	No	No
Other Partners Hospitals	5/5	5/5	5/5	2/5	2/5
Share of Non-Partners Hospitals	*74%*	*75%*	*81%*	*78%*	*85%*
	Neighborhood Health Plan (NHP)				
Partners: Academic Hospitals	Yes	Yes	Yes	Yes	Yes
Other Partners Hospitals	2/5	4/5	4/5	4/5	5/5
Share of Non-Partners Hospitals	*72%*	*84%*	*87%*	*87%*	*87%*
	CeltiCare *(new in 2010)*				
Partners: Academic Hospitals	---	Yes	Yes	Yes	Yes
Other Partners Hospitals		3/5	3/5	3/5	3/5
Share of Non-Partners Hospitals	---	*25%*	*33%*	*40%*	*40%*
	Fallon Health Plan *(central Mass only)*				
Partners: Academic Hospitals	No	No	No	No	No
Other Partners Hospitals	0/5	0/5	0/5	1/5	0/5
Share of Non-Partners Hospitals	*20%*	*21%*	*12%*	*10%*	*10%*

NOTE: This table shows statistics on the hospital network coverage of the five Massachusetts exchange plans in each plan year. For each plan, I list statistics separately for the Partners Healthcare System (the star hospital system this paper focuses on) and all other Massachusetts general acute care hospitals. For Partners, I list whether the plan covers its two star academic medical centers (Mass. General and Brigham & Women's hospitals) and the number of other Partners hospitals covered. For all other hospitals, I list the share of hospitals covered, weighted by number of hospital beds. The largest network change in the exchange history is for Network Health in 2012 when they dropped almost all of the Partners system, as well as eight other hospitals. I focus on this change in my reduced form analysis in Section 1.4.

outside physician.[27] By contrast, NHP continued to cover Partners but had special reason to do so: Partners purchased it effective in fiscal year 2013. Thus, as CommCare transitioned into the ACA mid-way through fiscal 2014, only one plan fully covered Partners and that through a vertical ownership relationship.

Data: Plan Choices and Insurance Claims

To study CommCare, I use a comprehensive administrative dataset on plan enrollment and insurance claims for all plans from 2006-2013.[28] For each de-identified enrollee, I observe demographics, plan enrollment history, and claims for health care services while enrolled in the market. The claims include information on patient diagnoses, services provided, the identity of the provider, and the actual amounts the insurer paid for the care.

I use the raw data to construct two datasets for model estimation. The first dataset is for hospital demand and costs. From the claims, I pull out all instances of inpatient stays at general acute care hospitals in Massachusetts during fiscal years 2008-2013, the period over which I have data on plan networks. I add on hospital characteristics from the American Hospital Association (AHA) Annual Survey and define patient travel distance using the Google Maps driving distance from the patient's home zip code centroid to the hospital.[29] For each hospitalization, I sum up the insurer's total payment while the patient was admitted (including both the hospital facility fees and physician professional service fees) and use this to estimate the hospital price model described in Section 1.6.1.

[27] Testimony from CeltiCare's CEO to the Massachusetts Health Policy Commission supports this interpretation: "For the contract year 2012, Network Health Plan removed Partners hospital system and their PCPs [primary care physicians] from their covered network. As a result, the CeltiCare membership with a Partners PCP increased 57.9%. CeltiCare's members with a Partner's PCP were a higher acuity population and sought treatment at high cost facilities. … A mutual decision was made to terminate the relationship with BWH [Brigham & Women's] and MGH [Mass. General] PCPs as of July 1, 2013." (HPC 2013)

[28] The data was obtained via a data use agreement with the Massachusetts Health Connector, the exchange regulator. To protect enrollees' privacy, the data was purged of all identifying variables.

[29] I thank Amanda Starc and Keith Ericson for sharing Google Maps distance data between hospitals and zip codes.

The second dataset is for insurance plan demand and costs. Using the enrollment data, I construct a dataset of available plan choices, plan characteristics (including premium), and chosen options during fiscal years 2008-2013. I consider plan choices made at two distinct times: (1) when an individual initially enrolls in CommCare or re-enrolls after a gap in coverage, and (2) at annual open enrollment when premiums change and current enrollees are allowed to switch plans. A key difference between these two situations is their default choice. New and re-enrollees must make an active choice to receive coverage – if they do nothing, they are not enrolled and do not appear in the data.[30] By contrast, current enrollees who do not respond to open enrollment mailings and emails are defaulted to their current plan. Over 95% of current enrollees do not switch, suggesting the likely importance of the default. For the remainder of the year following each plan choice, I sum up each individual's total costs in the claims data, distinguishing between inpatient hospital costs (included in the hospital cost model) and non-inpatient costs. I use the non-inpatient costs to estimate an additional plan cost model described in Section 1.6. The tables in Appendix A.1 show summary statistics for both the hospital and plan choice samples.

1.4 Reduced Form Adverse Selection Evidence

In this section, I present reduced form evidence of adverse selection against plans covering the Partners hospitals. I also show evidence consistent with the key predictions of the theory in Section 1.2. Specifically, I document a factor that seems to capture enrollee preference for Partners – past use of care at a Partners facility. These past users are (1) higher cost (even after risk adjustment), (2) more likely to select a plan that covers Partners, and (3) more likely to use Partners for subsequent hospitalizations. These facts suggest that past users *select* plans based on their desire to *use* Partners, whose high prices mean higher costs for insurers. I show that these enrollees' loyalty gives Partners substantial market power, since these enrollees will switch plans to retain access.

[30] An exception to this rule prior to fiscal year 2010 was that the exchange auto-assigned the poorest new enrollees who failed to make an active choice. I observe these auto-assignees and exclude them from the plan demand dataset.

I present two types of evidence to show these facts. First, I use the standard positive correlation test for asymmetric information. Second, I examine evidence around a large plan's exclusion of Partners from network in 2012.

Positive Correlation Test

My analysis starts with the positive correlation test (Cardon and Hendel 2001; Chiappori and Salanie 2000). This method tests whether individuals who select generous plans (here, plans covering Partners) are more costly in ways not captured by factors that prices can vary on. This is a joint test for adverse selection and moral hazard, since either factor may drive the correlation. Indeed, in my setting, the selection and moral hazard effects of covering Partners are connected.

I follow the "unused observables" approach of Finkelstein and Poterba (2013) for the correlation test. This procedure works by separately regressing plan selection and cost on factors on which prices vary plus "unused" factors that I observe but were not included in pricing or risk adjustment. In the 2011-2013 years I analyze, insurers were not allowed to vary prices *at all* across enrollees (i.e. full community rating) but their payment for an enrollee equaled price times the enrollee's risk score. Therefore, the only "used" factor in the regression is the risk score.[31]

I include an unused factor that is likely to capture individuals' preferences for Partners: whether they have used a Partners hospital in the past, either for inpatient or outpatient care. Outpatient care – which includes visits to a doctor whose practice is in a hospital – is much more common and accounts for most of the past use. Past use is, of course, not an immutable characteristic but a proxy to pick up *future* preference for Partners. It may do so both through time-invariant heterogeneity and provider loyalty (state dependence), and I do not attempt to separate the two. The positive correlation test requires only a correlation between cost and plan choice, and I use past Partners patient status to test for this correlation.

[31] I also include plan-year dummies (for costs) and year dummies (for plan choice) to net out any constant effects across plans or time. All results are robust to excluding these.

Table 1.4 shows the test results. The first two columns show regressions for plan choice (where the outcome is a dummy for choosing a plan covering Partners), and the remaining columns show regressions for costs. Among all consumers (column 1), past Partners users are 32.8% points more likely to choose a plan that covers it, a very large effect relative to the mean probability of 41%. One concern is that this estimate partly reflects consumer inertia in plans covering Partners. To address this, Column 2 restricts the sample to re-enrollees who I know are making an active choice.[32] Based on Partners use in their earlier enrollment spell, these enrollees are still 26.8% more likely to choose a plan covering Partners.

The next set of columns shows the results for costs. For the full sample, past Partners use is significantly associated with $56.2 higher risk-adjusted monthly costs, a 15% increase relative to mean costs of $387.[33] As the theory suggested, past Partners users may be more expensive either because they are unobservably sicker or because for the same sickness, they use expensive providers. I examine this in several ways. First, columns 4 and 5 separate the sample by whether the enrollee's plan covers Partners. In plans not covering Partners the estimate is a smaller but still significant $36.5, while in plans covering Partners, the estimate is a much larger $62.3 (and the difference between these is statistically significant). This pattern is consistent with *both* unobserved sickness and use of Partners driving the higher costs. Enrollees who prefer Partners are more costly even in plans that do not cover it. But they are particularly costly when their plan gives them access to Partners facilities.[34]

[32] I restrict the sample to people with a break of at least three months to rule out cases where the break represents a short lapse in paying premiums. This estimate is robust to limiting to longer breaks. Even among enrollees gone for more than two years, past Partners users are 21% points more likely to choose a Partners-covering plan.

[33] I perform a robustness check to test whether a limitation in the risk adjuster – its lack of diagnosis information for new enrollees – is driving these results. Limiting the sample to individuals with full diagnosis risk adjustment, past Partners use still significantly predicts higher risk-adjusted costs, with a magnitude similar to the full sample.

[34] An alternate interpretation is that only unobserved sickness is driving the results but that past Partners users who do not choose a plan covering Partners are less unobservably sick. Against this interpretation, however, I show below that these individuals are much more likely to choose Partners hospitals when hospitalized.

Table 1.4 Positive Correlation Test Regressions

Panel A: Plan Selection and Costs

VARIABLES	**Dep. Var.: Choose Plan Covering Partners**		**Dep. Var.: Costs (per month)**		
	All Enrollees (1)	Re-Enrollees (2)	All Plans (3)	Plans Not Covering Partners (4)	Plans Covering Partners (5)
Past Partners Use	0.328***	0.268***	56.22***	36.45***	62.31***
	(0.002)	(0.004)	(5.54)	(8.04)	(7.67)
Control Variables					
Risk Score	-0.006***	-0.015***	405.03***	392.71***	419.18***
	(0.001)	(0.002)	(7.94)	(10.08)	(12.58)
Year Dummies	X	X	---	---	---
Plan-Year Dummies	---	---	X	X	X
Dependent Var. Mean	0.409	0.420	$387.0	$372.8	$407.5
Observations	843,779	131,120	843,779	498,773	345,006
R-Squared	0.126	0.201	0.10	0.11	0.10

NOTE: Panel A shows results from the positive correlation test for asymmetric information discussed in Section 1.4. It tests whether a factor not used in plan pricing (here, whether an enrollee has used Partners in the past) correlates with both probability of choosing a plan covering Partners and with risk-adjusted costs per month. The tests are performed for 2011-2013, the years for which I have full risk adjustment data. The cost regressions are performed separately for enrollees in all plans and in plans that cover and do not cover Partners.

Panel B: Other Dependent Variables and Tests

VARIABLES	Sample: Plans Covering Partners			
	Hospitalization Rate (1)	Prob. Choose Partners Hosp. (2)	Cost per Hospitalization (3)	Costs: Diagn. Risk Adj Only (4)
Past Partners Use	0.0005	0.463***	1,953***	76.64***
	(0.0016)	(0.010)	(454.9)	(8.97)
Risk Score	0.0574***	-0.003	750.6***	379.85***
	(0.0020)	(0.002)	(169.6)	(13.70)
Plan-Year Dummies	X	X	X	X
Dependent Var. Mean	0.0439	0.244	$12,543	$394.50
Observations	345,006	11,759	11,759	172,731
R-Squared	0.042	0.208	0.016	0.13

*** p<0.001, ** p<0.01, * p<0.05

NOTE: Panel B explores why past Partners users may be higher cost by using a similar regression structure as Panel A with other dependent variables. The final column shows a robustness check on the cost results in Panel A by limiting the sample to individuals for whom the exchange used diagnosis-based risk adjustment.

To further interpret these results, the bottom half of Table 1.4 shows other attributes of past Partners patients in plans that cover Partners. The estimates show that controlling for risk score, they are 46.3% points more likely to choose a Partners hospital when hospitalized and $1,953 more costly per hospitalization. Again, these results are consistent with individuals' decisions to use Partners directly contributing to their higher costs. However, their risk-adjusted hospitalization rate is no higher – although their absolute rate is higher, the risk score picks up the difference.

Evidence from Plan Network Changes in 2012

A second way to test for selection is to study plan network changes. I focus on changes in 2012 that were both the largest in CommCare's history and the only time when the main Partners hospitals were dropped. As discussed in Section 1.3, this change occurred after the exchange introduced new incentives rewarding the lowest-price plans. In response, Network Health cut its price by about 15% and, to cut costs, excluded the Partners system and eight other hospitals from its network. Other plans also changed prices but did not make significant network changes at the time.

I start by studying plan choice patterns, again using past use as a proxy for Partners preference. Figure 1.1 shows the share of current Network Health enrollees who switched plans just before the start of each plan year. The average switching rate is usually very low (about 5%), but it spikes in 2012 to just over 10%. All of this spike is driven by patients of the hospitals Network Health dropped. Almost 40% of past Partners patients switched away from Network Health in 2012, a more than *seven-fold* increase from adjacent years. One factor behind this increase may be that Partners providers encouraged their patients to switch plans.[35] Most of these switchers moved to CeltiCare and Neighborhood Health Plan (NHP), the two remaining plans covering Partners. Patients at the eight other dropped hospitals were also more likely

[35] By chance, I observed Partners doing so during a tour of Brigham & Women's Hospital in late 2013 when a Medicaid managed care plan was about to drop Partners from its network. The finance department was calling past patients to let them know they needed to switch plans to maintain access to Partners providers.

to switch but only about half as frequently as Partners patients. This is consistent with Partners' star power giving it much greater ability to influence plan choices than non-star hospitals.

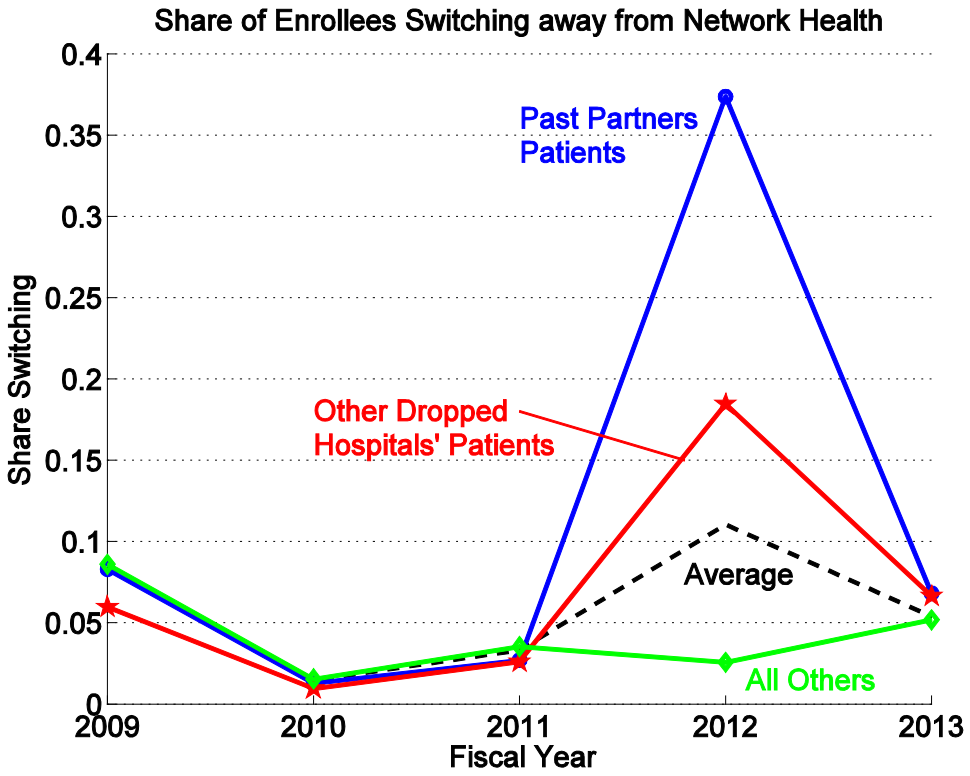
Because the Partners patients are a high-cost group, these switching patterns had important cost implications. Table 1.5 shows the change in unadjusted and risk-adjusted costs for Network Health between 2011 and 2012. Overall, its per-member-month costs fell by 21% (or 15% after risk adjustment), a huge decline in the health insurance industry where costs rarely fall. However, among a fixed population of "stayers" in Network Health in both years, costs fell by just 6%. The remainder of the change came through selection of enrollees leaving and joining the plan. The most expensive group was those who switched away from Network Health in 2012 – their 2011 risk-adjusted costs were $509 per month, almost 40% above the plan's average.[36]

The bottom panel of the table breaks down costs for switchers and stayers into past Partners patients (as of the start of 2012) and all others. It makes clear that Partners patients drove the high costs among switchers away from Network Health. They represented 68% of all switchers and had risk-adjusted costs of $571 per month in 2011 (54% above the plan average), whereas all other switchers had below-average costs. In comparison, the Partners patients who stayed with Network Health were somewhat less expensive – only $472 per month in 2011. They also experienced a substantial 26% decline in costs in 2012, accounting for *all* of the cost declines among stayers. Although mean reversion surely plays a role, this pattern is also consistent with dropping Partners having a *differential* cost impact for the people most likely to use it, as in the model.[37]

[36] In addition to the switchers, the group exiting the market was high-cost. While the reasons are unclear, exiting enrollees appear to be high-cost in other years and plans as well, not just in Network Health in 2011-2012.

[37] One fact suggesting reversion does not fully explain the difference is that the Partners patients who left Network Health had a much smaller 13% cost decline, despite having even higher baseline costs in 2011. Of course, this difference could also reflect selection. To further separate mean reversion from treatment effects, I could attempt to decompose the changes into quantity of services (e.g., number of hospitalizations) vs. price per service.

Figure 1.1 Plan Switching after Network Health Dropped Partners in 2012



NOTE: This figure shows the share of enrollees in Network Health plan who switch to another plan at the start of each fiscal year (when all exchange enrollees are given an opportunity to switch plans). The black dashed lines show the average switching rate for all enrollees; the colored solid lines decompose this average into subgroups. In most years, switching rates are quite low, but in 2012, switching spiked after Network Health dropped the star Partners hospitals and eight other less prestigious hospitals. The graph shows a large switching spike among past patients of Partners (in blue) and a smaller spike among patients of the other dropped hospitals (in red). There was little change in switching rates among all other enrollees (in green). These results suggest that many patients are willing to switch plans to keep access to their regular hospital provider. As I show elsewhere, the past Partners patients were a particularly high-cost group, so these switching patterns contributed to favorable selection for Network Health when it dropped Partners.

Table 1.5 Analysis of Network Health's Cost Changes from 2011-2012

Network Health Cost Changes, 2011-2012

Enrollee Group	Avg. Costs			Risk Adj. Costs			Group
	2011	2012	%Δ	2011	2012	%Δ	Size*
All Enrollees	**$386**	**$306**	**-21%**	**$370**	**$313**	**-15%**	**---**
Stayers	$323	$303	-6%	$317	$300	-6%	36,768
Left Plan in 2012							
Switched Plans	$670	*[$616]*	---	$509	*[$425]*	---	4,640
Exited Market	$470	---	---	$459	---	---	22,617
Joined Plan in 2012							
Switched Plans	*[$283]*	$288	---	*[$303]*	$309	---	15,062
Entered Market	---	$315	---	---	$334	---	51,109

NOTE: This table shows the changes in medical costs per member-month for Network Health from 2011 (when it covered the star Partners hospitals) to 2012 (when it dropped them). The first set of columns show unadjusted costs. The next columns show risk-adjusted costs, defined as a group's average cost divided by average risk score (a measure defined by the exchange for risk adjustment). Group size is the number of enrollees in the relevant group during the year(s) they were enrolled in Network Health. Overall, Network Health's costs fell by 21%, or 15% after risk adjustment. The next rows break costs into enrollee subgroups: a fixed group of "stayers" (people in the plan in both years) and enrollees who left or newly joined the plan in 2012. These show that costs for stayers fell by just 6%. Selection out of Network Health by high-cost enrollees explains a large portion of the fall in costs.

Panel B: Breakdown by Partners Patient Status

Enrollee Group	Avg. Costs			Risk Adj. Costs			Group
	2011	2012	%Δ	2011	2012	%Δ	Size*
Stayers							
Partners Patients	$533	$397	-26%	$472	$332	-30%	5,308
All Others	$284	$289	2%	$285	$294	3%	31,460
Switched from Network Health in 2012							
Partners Patients	$778	*[$674]*	---	$571	*[$438]*	---	3,169
All Others	$403	*[$490]*	---	$335	*[$391]*	---	1,471
Switched to Network Health in 2012							
Partners Patients	*[$547]*	$435	---	*[$502]*	$364	---	1,303
All Others	*[$261]*	$275	---	*[$284]*	$302	---	13,759

* Number of enrollees during the relevant year they were enrolled in Network Health.

NOTE: Panel B breaks down cost changes for stayers and switchers (see note above) in Network Health into people who had ever been a patient at a Partners hospital (as of the start of 2012) and all other enrollees. Exit by high-cost Partners patients fully explains high costs among people who switched out of Network Health in 2012. Among stayers, Partners patients' costs fell sharply, while all other enrollees' costs rose slightly.

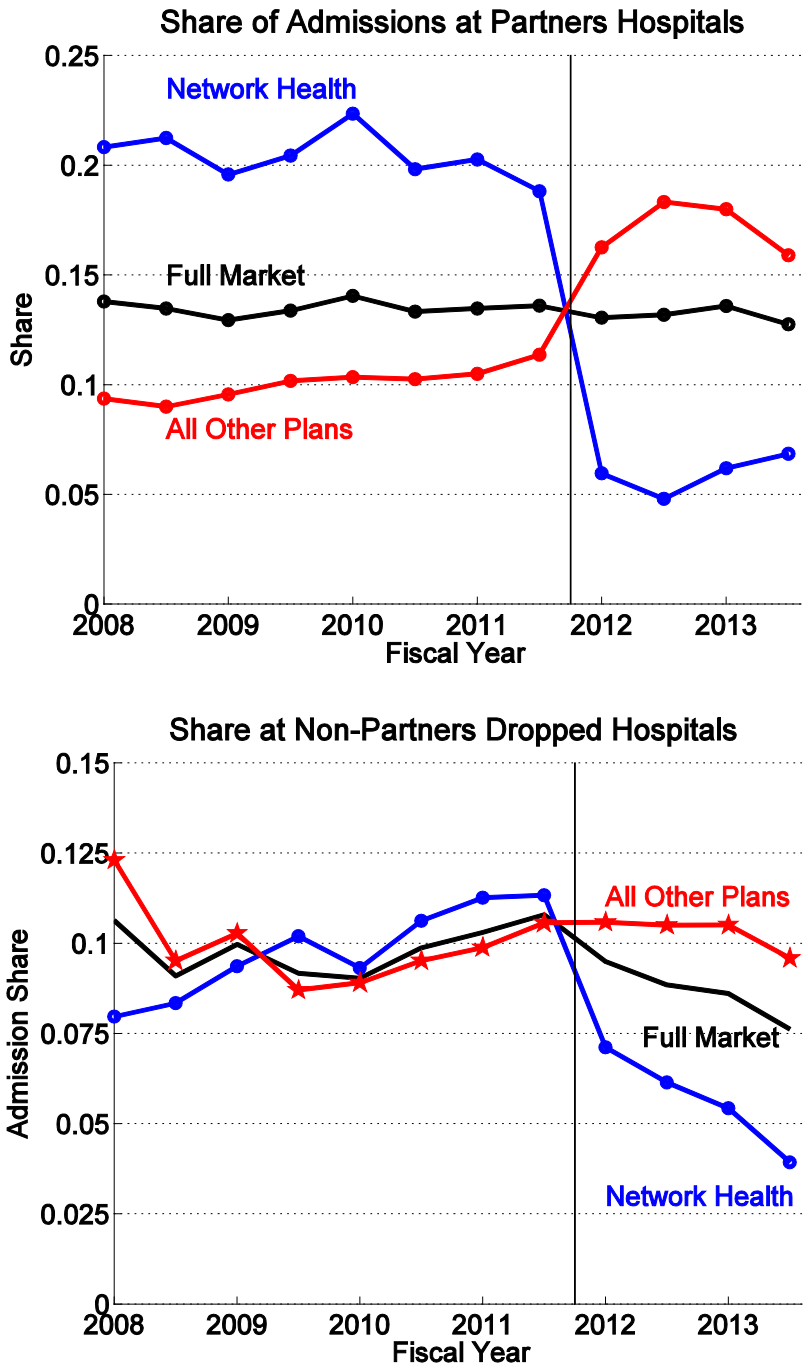
These network changes and selection patterns had important impacts on hospital choice patterns and costs. Figure 1.2 shows the impact on admission shares at the Partners hospitals (top panel) and the other eight dropped hospitals (bottom panel). In both cases, admission shares among Network Health enrollees fell sharply.[38] But for Partners hospitals, admissions also rose sharply at other plans – despite their Partners coverage not changing. Selection by enrollees most likely to use Partners appears to drive this increase. For the market as a whole, Partners admission share barely budged – despite being dropped by the second largest plan. The same was not true for the other, less prestigious dropped hospitals. Their shares at other plans did not increase much, and their overall market share fell. These hospital use patterns again illustrate the unique status of Partners in this market. Figure 1.3 shows the implications of these hospital use changes for plans' costs per hospital admission. Network Health's costs fell sharply in 2012 by about 15%, with the drop *entirely* driven by less use and lower prices at Partners hospitals.[39] Conversely, per-admission costs at the two plans still covering Partners (NHP and CeltiCare) began rising in 2012, as more of their patients used the expensive hospitals.[40]

[38] Notably, these declines, while substantial, were less than 100% because patients can still use out-of-network hospitals in an emergency or if given prior authorization by their plan.

[39] Per-admission costs also fell sharply at Partners, with a likely explanation being exchange rules for out-of-network reimbursements when patients use them in an emergency. In these cases, plans are allowed to pay the hospital at the state's low Medicaid payment rates. Because this rule does not apply to non-emergency admissions that the plan authorizes, costs (and prices) at Partners are still higher than for an average hospital.

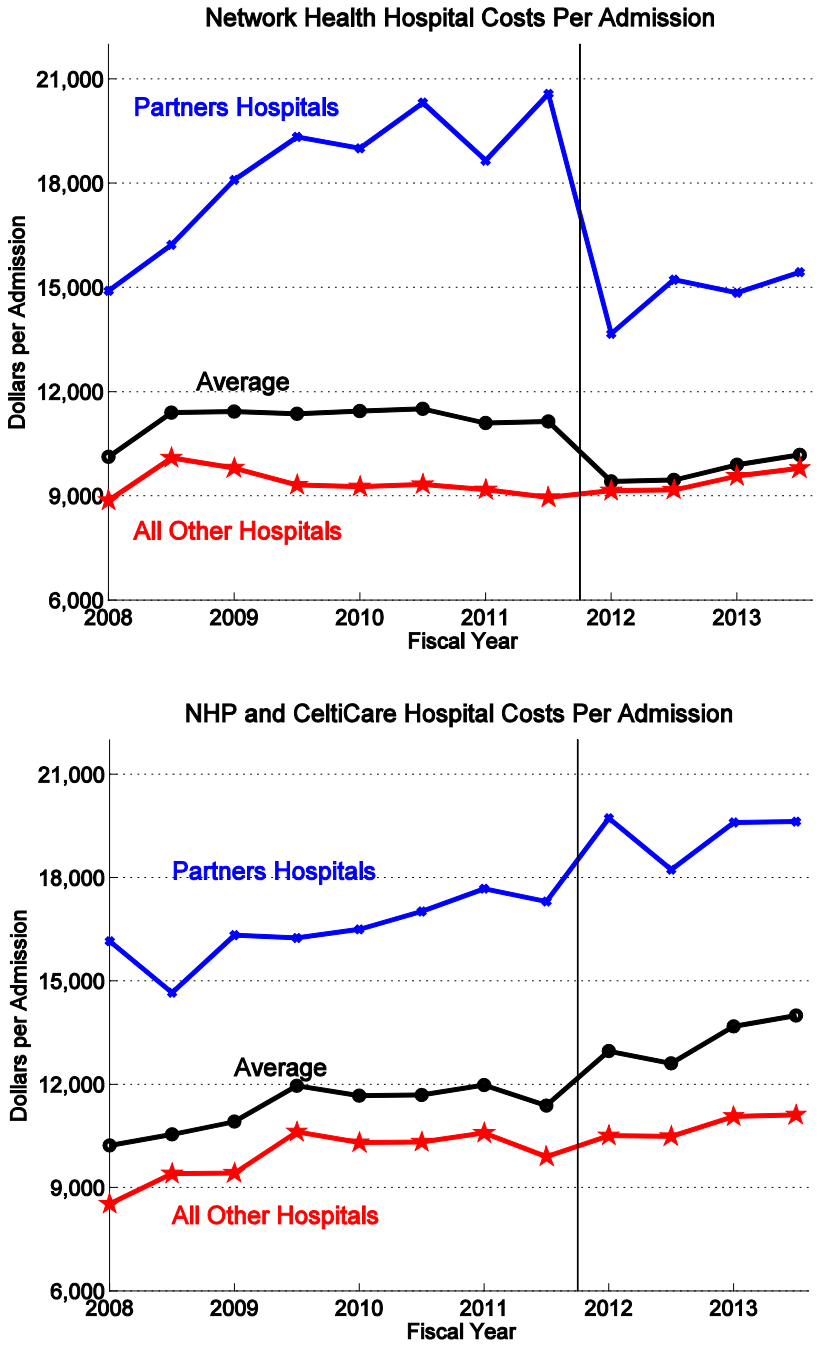
[40] Partners also does not appear to have given these plans price discounts after being dropped by Network Health. Costs per admission at Partners actually rise, but this is driven by a compositional shift among Partners hospitals, with little change in hospital-specific prices.

Figure 1.2 Admission Shares at Hospitals Dropped by Network Health in 2012



NOTE: These figures show the share of hospital admissions at hospitals that Network Health plan dropped from its network in 2012. The top figure shows shares for the star Partners hospitals; the bottom figure shows the eight other dropped hospitals. In both cases, shares fell sharply among Network Health enrollees in 2012. However, Partners' shares rose sharply at other plans (even though their coverage did not change), as the enrollees most likely to use Partners switched to plans that covered them. The same was not true for the less prestigious dropped hospitals, whose shares from other plans barely changed.

Figure 1.3 Changes in Cost per Hospital Admission around 2012Changes



NOTE: These figures show average costs per hospital admission for enrollees of two sets of plans: Network Health (top figure), which dropped the star Partners hospitals in 2012, and NHP and CeltiCare (bottom figure), which continued to cover them. Each figure shows the plans' overall average (in black) and also separate averages for Partners (blue) and all other hospitals (red). Network Health's per-admission costs fell 15% in 2012, with all of the reductions explained by less use and lower prices at the Partners hospitals. Costs for NHP and CeltiCare increased in 2012 as more enrollees used the high-price Partners facilities. All averages were calculated after winsorizing per-admission costs at $150,000 (above the 99.9th percentile) to reduce the influence of outliers.

1.5 Structural Model: Hospital and Insurance Plan Demand

The reduced form evidence suggests that consumers select into plans covering the star Partners hospitals based on their preference for using those hospitals. Understanding the competitive and welfare implications of this selection requires estimating a structural model that can capture this correlation. In this section, I present and estimate the hospital and insurance demand portion of this model. I follow a method introduced by Capps, Dranove, and Satterthwaite (2003) to capture how different consumers value plans' hospital networks. I first estimate a hospital demand model that captures how patients weigh different factors (e.g., distance, hospital characteristics) when selecting hospitals. This hospital demand model generates an expected "network utility" metric capturing the attractiveness of each plan's network to a specific consumer. I then estimate an insurance plan demand model, including network utility as a covariate. If patients choose plans based on their hospital networks, the coefficient on network utility should be positive. Because of the importance of past Partners users in the reduced form results, I allow preferences in both the hospital and plan demand models to vary based on which hospitals an individual has previously used. This section proceeds by estimating hospital demand (Section 1.5.1) and deriving network utility (Section 1.5.2). I then present and estimate plan demand (Section 1.5.3-1.5.4).

1.5.1 Hospital Demand

I use the micro-data on inpatient hospital use to estimate a multinomial logit model capturing how patients choose hospitals. My method and specification are similar to much past work (e.g., Town and Vistnes 2001, Gaynor and Vogt 2003, Ho 2006). The main covariates are distance and hospital characteristics, and I allow preferences to vary with patient observables. I do not include patient fees as a covariate, since CommCare's copays are constant across in-network hospitals and therefore drop out.[41] In

[41] Recent work by Ho and Pakes (2014) also finds that hospital price matters for referral patterns in plans where doctors are paid by capitation. Unlike their California setting, CommCare insurers pay doctors almost exclusively fee-for-service, with capitation accounting for less than 5% of physician service fees.

addition, I do not include an outside option, since I am focusing on patients sick enough to need hospital care and Massachusetts is a relatively complete hospital market.[42]

My model differs from past work in two main ways. First, based on the reduced form results, I allow hospital preferences to vary with whether a patient has used the hospital in the past (either for inpatient or outpatient care).[43] Although its interpretation is not obvious – it captures both heterogeneity and state dependence – I include past hospital use because of its importance as a channel for adverse selection. Second, because I observe a non-trivial number of out-of-network hospitalizations covered by plans (e.g., see Figure 1.2), I include out-of-network hospitals in the choice set. This captures the fact that patients can sometimes get plan authorization to see an out-of-network provider. To capture the associated hassle costs, I estimate a plan-specific out-of-network cost in the hospital choice model.[44] This specification generalizes previous work that disallows out-of-network admissions, which is equivalent to assuming an infinite hassle cost.

Consider an admission at time *t* for individual *i* (in plan *j*) who has principal diagnosis *d*. I specify the following model for the latent utility for hospital *h*:

$$u_{i,d,t,j,h} = \underbrace{\delta\left(Z_{i,d,t}\right)Dist_{i,h}}_{\text{Distance}} + \underbrace{\gamma\left(Z_{i,d,t}\right)X_h + \eta_h}_{\text{Hospital Characteristics}} + \underbrace{\lambda \cdot PastUse_{i,h}}_{\text{Past Use Dummy}} - \underbrace{\kappa_j \cdot 1\left\{h \notin N_{j,t}\right\}}_{\text{Out-of-Network Hassle Cost}} + \varepsilon_{i,d,t,h} \qquad (1.5)$$

where $Dist_{i,h}$ is patient travel distance (and distance-squared), X_h are observed hospital characteristics, η_h is an unobserved characteristic (captured by hospital dummies), $PastUse_{i,h}$ are past use indicators, and $1\left\{h \notin N_{j,t}\right\}$ is an out-of-network dummy (and κ_j the hassle cost). I allow coefficients on distance and characteristics to vary with patient observables $Z_{i,d,t}$ to allow for preference heterogeneity. Finally, $\varepsilon_{i,d,t,j,h}$ is an i.i.d. Type 1 extreme value error that generates the logit demand form.

[42] The only significant exception is spillover of patients in Southeastern Mass. to hospitals in Providence, RI.

[43] To rule out immediate readmissions, I require that the past use occurred more than 60 days beforehand.

[44] Patients can also use any hospital in an emergency (without needing plan authorization) but may need to be transferred once stabilized, creating a different type of hassle. I allow for an interaction between emergency status and the out-of-network cost but find little evidence that the cost is lower in emergencies.

Because all of the covariates are observed, I estimate the model by maximum likelihood. Table 1.6 shows the results. Consistent with previous papers' estimates, patients have a disutility of traveling to more distant hospitals, with the estimates implying that an extra 10 miles distance reduces a hospital's share by 31% on average. The model estimates a sizeable hassle cost for out-of-network hospitals that reduces their shares by 63% on average.[45] The table shows the largest hospital service x diagnosis interactions; the remaining coefficients are almost all significantly positive.

Two sets of coefficients have implications for adverse selection. First, teaching hospitals and particularly the largest academic medical centers (including the two star Partners hospitals) attract the sickest patients – where severity is based on an index of the costliness of a patient's diagnoses defined in Section 1.6.1. A one standard deviation increase in severity (a change of 0.3) increases the likelihood of using an academic medical center by 47%. Second, the past use dummies are very strong predictors of future hospital choices. For instance, patients who have previously used a hospital for outpatient care choose the same hospital in future admissions about 40% of the time. The model implies that this 40% share is an almost 5-fold increase above what would be expected otherwise.

The model fit is quite good, particularly when past hospital use variables are included. Calculating hospital shares at the service area-plan-year level,[46] the model explains 74% of the variance in shares, despite the absence of any year-specific interactions in the model. This indicates that conditional on network, hospital use patterns are fairly stable in the market over time. The left half of Table 1.6 shows estimates and fit from a simpler model (with only distance, out-of-network cost, and hospital dummies) for comparison. This simple model can also pick up 64% of the variance in shares. Most of the increase in fit from moving to the more complex model comes by adding the previous use covariates.

[45] A 63% reduction from being out of network may seem low. However, it is consistent with a basic statistic from the data: only 25% of hospital choices are out of network but 8% of admissions are at out-of-network facilities.

[46] Service areas are subregions defined by the exchange as the areas at which plans can choose whether or not to offer coverage. The five regions are divided into 38 service areas.

Table 1.6 Hospital Demand Estimates

	Simple Model		Full Model	
VARIABLE	Coeff.	Std. Error	Coeff.	Std. Error
Distance to Hospital:				
Distance in Miles (avg. coeff.)	-0.189***	(0.001)	-0.144***	(0.001)
Distance^2 (avg. coeff.)	0.0013***	(0.0000)	0.0009***	(1e-5)
Distance Interactions:				
x Income > Poverty			-0.006***	(0.0006)
x Age / 10			-0.003***	(0.0002)
x Severity Weight			-0.002	(0.0011)
x Emergency			-0.015***	(0.0006)
Out-of-Network Disutility				
Out-of-Network x BMC	-1.327***	(0.016)	-1.117***	(0.034)
Out-of-Network x CeltiCare	*(same for all plans)*		-1.464***	(0.058)
Out-of-Network x Fallon			-1.583***	(0.059)
Out-of-Network x NHP			-0.543***	(0.049)
Out-of-Network x Network			-1.011***	(0.036)
Out-of-Network x Emergency			0.010	(0.034)
Past Use of this Hospital *(>60 days before)*				
Inpatient Care			1.417***	(0.020)
Outpatient Care			2.202***	(0.013)
Hospital Characteristics				
Hospital Dummies	Yes		Yes	
Severity x Academic Med. Ctr. (avg).			2.076***	(0.044)
Severity x Teaching Hosp			1.026***	(0.045)
Diagnoses x Hospital Services (largest coeffs.):				
Mental: Psych. Services			1.844***	(0.040)
Pregnancy: Obstetrics Services			1.122***	(0.076)
Injury: Level 1 Trauma Center			0.805***	(0.037)
Cancer: Oncology Services			0.704***	(0.084)
Model Statistics:				
Pseudo-R^2 (McFadden's)	0.463		0.569	
R^2 in Shares (Area-Plan-Yr Level)	0.643		0.742	
Num. Choice Instances	74,383		74,383	

Std. Errors in parentheses. * = 5% sign., ** = 1% sign., *** = 0.1% sign.

NOTE: The table shows estimates for the multinomial logit hospital choice model described in Section 1.5.1. The left columns show a simple model, while the right columns show the full model used for all further analyses. The logit coefficients shown are interpretable as entering the latent utility function describing hospital choice. Past use variables are dummies for whether a patient has used each specific hospital at least 60 days before the current admission. Severity is an estimated summary measure of costs described in Section 1.6.1. In addition to the variables shown, the model includes: distance interacted with exchange region, detailed income group (by 50% of poverty), and gender; severity interacted with separate dummies for each academic medical center; and five additional diagnosis x hospital service interactions (circulatory diagnosis interacted with cath lab, interventional cardiology, and heart surgery services; pregnancy diagnosis x NICU; and musculoskeletal diagnosis x arthritis services).

One concern with the out-of-network costs is that they are based on the network of a patient's chosen plan. Plan selection on observables (such as distance and past use) is okay, but if there is selection on unobservable hospital tastes, the out-of-network cost will be biased upward. This problem could be addressed econometrically by estimating the plan and hospital demand models jointly, allowing for unobserved hospital tastes to enter into plan choices (see Crawford and Yurukoglu 2012; Lee 2013). I have not implemented this method because of its computational complexity. One suggestion that any bias may not be too severe is that the model credibly matches hospital use patterns around Network Health's 2012 change in network (see Section 1.6.4). Nonetheless, the absence of plan selection on unobserved hospital preferences is a limitation of the model.

1.5.2 Hospital Network Utility

To generate a measure of network utility for plan demand, I follow the method of Capps, Dranove, and Satterthwaite (2003) and Ho (2006, 2009). I define network utility based on the expected utility metric from the hospital demand system. Conditional on needing to be hospitalized, a consumer's utility of access to network $N_{j,t}$ in plan j is:

$$HospEU_{i,d,t,j}\left(N_{j,t}\right) \equiv E\max_{h}\left\{\hat{u}_{i,d,t,j,h}\left(N_{j,t}\right)+\varepsilon_{i,d,t,j,h}\right\} = \log\left(\sum_{h}\exp\left(\hat{u}_{i,d,t,j,h}\left(N_{j,t}\right)\right)\right) \tag{1.6}$$

where $\hat{u}_{i,d,t,j,h}\left(N_{j,t}\right) \equiv u_{i,d,t,j,h} - \varepsilon_{i,d,t,j,h}$. At the time of plan choice, however, consumers do not know their hospital needs. Instead, they have expectations of their hospital use frequency for each diagnosis d over the coming year, which I denote $freq_{i,d,t}$. Given this expectation, the *ex-ante* expected network utility is:

$$NetworkUtil_{i,j,t}\left(N_{j,t}\right) \equiv \sum_{d} freq_{i,d,t} \cdot HospEU_{i,d,t,j}\left(N_{j,t}\right) \tag{1.7}$$

This network utility in (1.7) is what I include in plan demand. To calculate it, I first use my data to estimate a Poisson regression of the annual number of hospitalizations for each diagnosis on

individuals' age and demographics.[47] I use the predicted values from these regressions for $freq_{i,d,t}$. Next, I calculate the value of $HospEU_{i,d,t,j}\left(N_{j,t}\right)$ for each plan and diagnosis, using the individual's location and demographics at the time of plan choice.[48] Finally, I input these values into equation (1.7) to calculate network utility. Because network utility does not have natural units, I normalize it so that 1.0 is the average decrease in utility for Boston-region residents when Network Health dropped Partners in 2012.

1.5.3 Plan Demand Model

I next estimate plan demand to capture how plan premiums and hospital networks influence consumers' choices. These estimates are important for capturing the extent of both market power (which is based on the price elasticity of demand) and adverse selection (which is based on the correlation between demand and cost). The demand estimates also generate a revealed-preference welfare measure capturing how individuals trade off generous networks against lower prices when choosing plans.

I use the dataset described in Section 1.3 to estimate a multinomial logit plan choice model for both new and current enrollees (allowing inertia for the latter, as I discuss below). I treat individuals' timing of entry/exit from the exchange as exogenous and model just their choices among exchange plans.[49] For new/re-enrollee *i* making a choice at time *t*, the model for utility of plan *j* is:

$$U_{ijt} = \underbrace{\alpha\left(Z_i\right)\cdot Prem_{j,t,Reg_i,Inc_i}}_{\text{Plan Premium}} + \underbrace{Network_{ijt}}_{\text{Hospital Network Vars.}} + \underbrace{\xi_{ijt}}_{\text{Unobs. Quality}} + \underbrace{\varepsilon_{ijt}^{Plan}}_{\text{Logit Error}} \tag{1.8}$$

where:

[47] I choose not to use diagnoses in this regression because past diagnoses are unavailable for new enrollees. I plan to explore a robustness check in which for current enrollees I use past diagnoses and for new enrollees, I use a separate model including chronic disease diagnoses observed in the subsequent plan year.

[48] The two hospitalization variables that remain to be filled in are severity and emergency status. For emergency status, I use the average emergency probability for each diagnosis to take an average of the values of *EU* for each possibility. For severity, I regress severity in the hospitalizations data on age-gender groups and emergency status and use the predicted value from this regression for each individual.

[49] Because exogenous factors like income and job status determine exchange eligibility and generous subsidies incentivize participation, this assumption seems reasonable. This assumption implies that in my model, changes in plan prices and networks do not induce people to substitute into/out of the market. Although it would be nice to weaken this assumption, I do not have sufficient data on people choosing the outside option (largely uninsurance) to estimate a model incorporating it as a choice.

$$Network_{ijt} = \beta_1(Z_i) \cdot NetworkUtil_{ijt} + \beta_2(Z_i) \cdot CoverPastUsed_{ijt}$$
$$\xi_{ijt} = \xi_{j,Reg_i,Inc_i} + \xi_{j,t,Reg_i}$$

and ε_{ijt}^{Plan} is an i.i.d. Type 1 extreme value error that gives demand its logit form. Plan utility depends on three sets of plan attributes: premiums, networks, and unobserved quality. Premiums – which vary across plans and within-plan across years, regions, and income groups – are observed, and I include them directly. Hospital networks are more difficult because while observed, the value of a given network varies across individuals. To capture this heterogeneity, I include two terms: the consumer-specific network utility measure (see Section 1.5.2) and a direct variable for whether the plan covers a consumer's previously used hospitals (or the share covered if there are multiple). Of course, these two variables are related, since past use entered hospital demand and therefore influenced network utility. However, the direct variables may predict demand beyond their impact on hospital network utility for several reasons. First, they may capture loyalty to doctors, who in Massachusetts are often hospital-affiliated and covered/dropped along with the hospital.[50] Second, it may be picking up error in hospital demand or the sickness frequency prediction. Finally, it may matter simply because plan and hospital choices are driven by different things. People may choose plans based on whether it covers their regular provider but hospitals based on many other factors (e.g., which hospital is closest in an emergency).

The third set of covariates in plan demand (ξ_{ijt}) are plan dummies capturing unobserved quality – e.g., customer service and plan reputation.[51] To aid identification of the premium coefficient (see discussion below), I allow these to vary at a detailed region-year and region-income group level.

Preference heterogeneity enters this model in two ways. First, I allow observed heterogeneity by income, age, and gender groups for the premium coefficient and by income group for network utility. Second, the network variables also incorporate heterogeneity, since (for the same plan) they vary by

[50] Though I do have information on physician networks and utilization, I have not yet modeled physician demand or network utility because of its complexity.

[51] Past work has found reputation to be an important driver of demand in the Medigap insurance market (Starc 2011), and based on my discussions with market participants, reputation is also important in CommCare.

consumer location, sickness, and past relationships with providers. This heterogeneity is useful for capturing substitution patterns and adverse selection.

Current Enrollees and Inertia: The model so far has applied to new/re-enrollees, who I can be sure are making active choices. A final issue is how to treat current enrollees, who can switch plans at annual open enrollment but are defaulted into their current plan if they take no action. There is growing evidence that defaults and inertia matter in health insurance (Ericson 2014, Handel 2013), and consistent with this, I find that fewer than 5% of enrollees switch plans each year. However, how to interpret this low switching rate is less clear. It may reflect a combination of true inertia/switching costs (a form of state dependence) and preference heterogeneity causing optimal choices to be serially correlated.[52]

While I am not able to fully separate these factors, I want the model to capture switching behavior because of its implications for selection. To do so, I take a reduced form approach. In addition to the terms in equation (1.8), current enrollees' utility includes a dummy for their current plan. I allow the coefficients on this dummy to vary with observed demographics and (based on the evidence in Section 1.4) whether the plan has just dropped a previously used hospital. These inertia coefficients can be interpreted as either switching costs or reduced form coefficients capturing the likelihood of consumers being passive/inattentive in their switching choice, and I report statistics for both interpretations.[53]

Including current plan dummies ensures that the model will match average switching rates for each group with a separate coefficient. However, the coefficients themselves will pick up both true inertia and any unobserved heterogeneity driving choice persistence. For matching static adverse selection, it is not clear that it is critical to distinguish these factors. Where the two specifications will primarily differ is in their implications for dynamic competition, which I do not study in my counterfactuals. However, in

[52] This low switching rate does not appear to only reflect heterogeneity. Enrollees who enter the exchange just after prices have changed end up with very different market shares overall than enrollees who entered just before the price change. This group-level share difference is strongly suggestive the true state dependence is involved.

[53] In Appendix A.2.2, I show how this maps into a particular two-step model of inattention, where the first step models whether an enrollee is passive or active and a second step models plan choice conditional on being active.

interpreting the inertia estimates, readers should keep in mind that these coefficients are also picking up unobserved heterogeneity.[54]

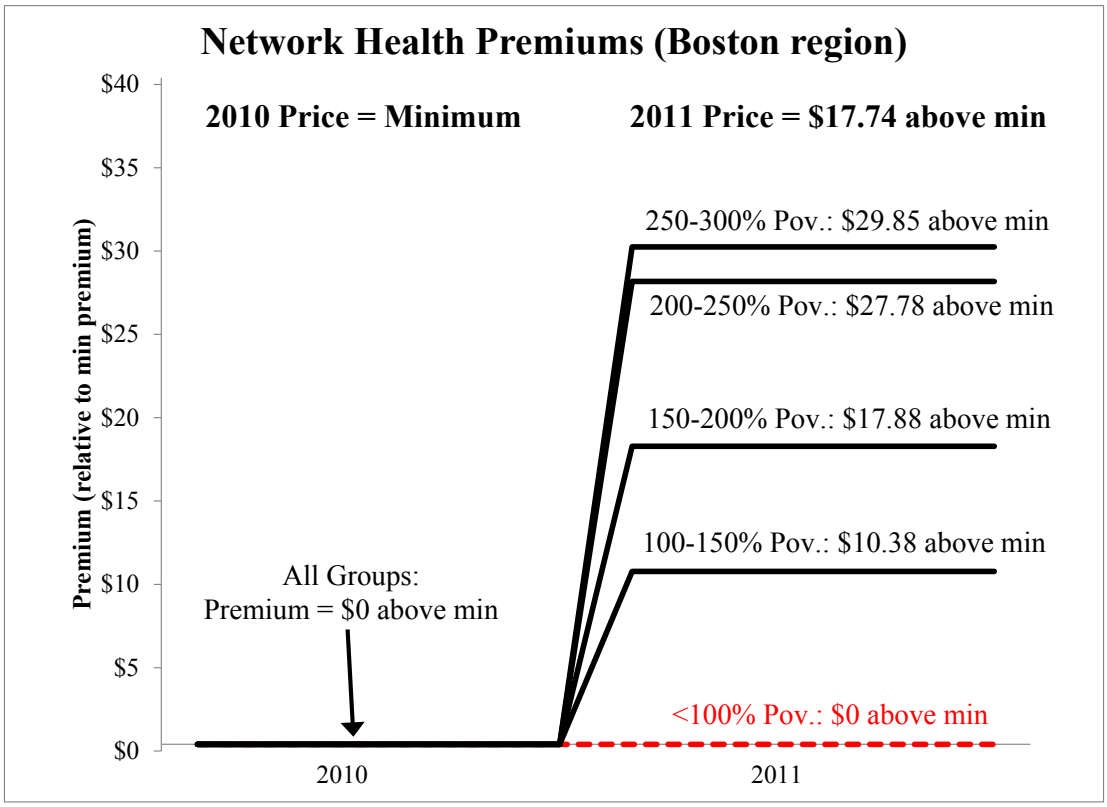
Identification and Estimation: I estimate the model using a micro-data method of moments estimator similar to Berry, Levinsohn, and Pakes (2004). A key difference in my setting is that the main plan attributes – premium and network utility – vary across individuals even for the same product in the same market and year. As a result (under assumptions discussed below), I can estimate the premium and network coefficients consistently from the micro-data alone, without needing instruments.

To identify the premium coefficients, I use within-plan variation induced by CommCare's subsidies. The key variation is that higher price plans have higher premiums for above-poverty enrollees but the same premium (always \$0) for fully subsidized below-poverty enrollees. This structure also creates differential premium changes across years, which I use for identification. Figure 1.4 shows how these differential changes work with an example from Network Health in the Boston region in 2010-2011. In 2010, Network Health was the cheapest plan for all groups. In 2011, its relative price increased but while above-poverty groups' premiums increased, below-poverty premiums were unchanged (still \$0).

I use these differential premium changes for identification by absorbing all other premium variation with a detailed set of plan dummies. Recall that because of regulation, premiums vary only across plans, years, regions and income groups. The first set of dummies (ξ_{j,Reg_i,Inc_i}) absorb any persistent demand differences for plan j across income groups (within a region). The second set of dummies (ξ_{j,t,Reg_i}) absorb demand differences across regions and over time. The remaining variation is from within-region differential premium changes across income groups. Because I allow a separate premium coefficient for each above-poverty group, the main identification comes from comparing demand changes for each above-poverty income group to those of below-poverty enrollees.

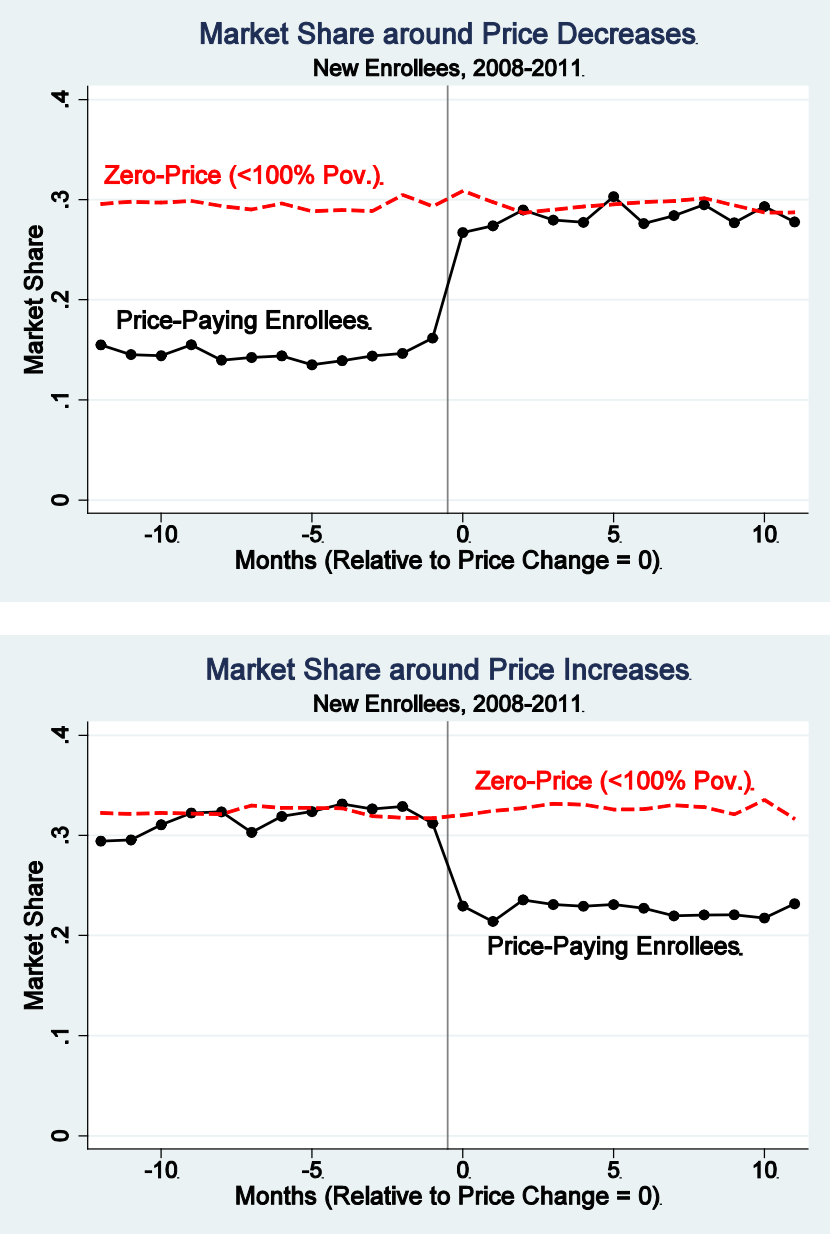
[54] In a future revision, I plan to do a robustness check with a demand model that includes time-invariant unobserved heterogeneity through random coefficients on premiums and plan dummies. I will use the choice patterns of re-enrollees to separately identify the random coefficient variances from the switching costs.

Figure 1.4 Premium Coefficient Identification Strategy



NOTE: This graph shows an example of the subsidy-driven premium variation for a single plan used to identify the premium coefficients in the plan demand model. The example is for Network Health plan in the Boston pricing region in 2011 and 2012. Network Health's premium in Boston was the cheapest in 2010, but in 2011 its relative premium increased. Because of subsidies, different income groups faced different premium increases. In particular, the below-poverty group faced *no increase* (its plans are always fully subsidized), while higher income groups faced different premium increases depending on the subsidy rules for each group. I use these within-region differential premium changes across income groups to identify the premium coefficients in plan demand. This identification approach is similar to difference-in-differences.

Figure 1.5 Premium Identification and Test of Parallel Trends Assumption



NOTE: These graphs show the source of identification for the premium coefficients in plan demand (see also Figure 1.4) and test the key parallel trends assumption for the difference-in-differences approach. Each graph shows average monthly plan market shares among new enrollees for plans that at time 0 decreased their prices (top figure) or increased their prices (bottom figure). Each point represents the shares for an independent set of new enrollees. The identification comes from comparing demand changes for above-poverty price-paying enrollees (for whom premium changes at time 0) versus below-poverty zero-price enrollees (for whom premiums are unchanged at $0). Consistent with the parallel trends assumption, trends in shares are flat and parallel for both groups at times other than the premium change but change sharply for price-payers only at the price change. The sample is limited to fiscal years 2008-2011. I drop 2007 because above-poverty enrollees did become eligible for the market until mid-way through the year and 2012+ because below-poverty new enrollees became subject to a limited choice policy that required them to choose lower-price plans.

This identification strategy is a nonlinear version of the standard difference-in-differences approach. Thus, the key assumption is that any changes in unobserved plan quality evolve in parallel for low- and high-income enrollees. This assumption seems reasonable because all groups have access to a plan under the same brand name, with the same provider network and member services. However, to test its validity, I employ the standard parallel trends test for difference-in-differences. This test compares trends for the outcome (market shares) around price changes for the treatment group (above-poverty) versus the control group (below-poverty). Figure 1.5 shows this test, plotting average market shares for new enrollees in each month around price changes, separately for price cuts (top graph) and price increases (bottom graph).[55] Consistent with the parallel trends assumption, market share trends are flat and parallel for both groups at all times except when prices change. At price changes, price-paying groups' shares jump sharply in the expected direction, while zero-price groups' shares are essentially unchanged.

The detailed plan dummies are also helpful for proper identification of the network utility coefficients. The potential identification threat is that plans with better networks also have better unobserved quality. However, with the plan dummies, the network utility coefficients are identified from *within-plan* variation across individuals in the same region and year. A key source of variation is enrollees' location relative to covered hospitals, since this strongly predicts hospital utility.

I estimate the model using moments similar to those used in Berry, Levinsohn, and Pakes (2004).[56] For plan dummies, I match observed market shares for the relevant plan and enrollee group. For plan characteristics (whose coefficients vary with observed demographics), I match the average interaction between the characteristic and the demographic among chosen plans in the data. Appendix A.2 shows the formulas for these moments.

[55] The analysis is restricted to fiscal years 2008-2011. I drop 2007 because above-poverty enrollees did not start enrolling in the market until mid-way through 2007. I drop 2012+ because below-poverty new enrollees become subject to a limited choice policy that required them to choose lower-price plans.

[56] I use method of moments rather than maximum likelihood for two reasons. First, my network utility covariates are not observed, and I employ a standard error correction that is valid for method of moments. Second, in future revisions, I plan to include random coefficients, for which simulated method of moments is more appropriate.

1.5.4 Plan Demand Estimates

The demand estimates are shown in Table 1.7. Premiums (in dollars per month) enter negatively and significantly for all income groups. (I normalize the average premium coefficient to -1.0, so the remaining coefficients can be interpreted as dollar values for an average enrollee.) Premium sensitivity decreases monotonically with income, with the highest-income group's coefficient less than half as large as the lowest-income group's. Premium sensitivity also decreases with age, although much less sharply. Overall, these estimates imply that new enrollees are quite premium sensitive. A $1 increase in monthly premium decreases the average plan's market share among premium-paying enrollees by 3.0%.[57]

Enrollees place positive and significant value on both measures of hospital network quality. Recall that network utility was normalized so that 1.0 was the average utility change for Boston-area enrollees when Network Health dropped Partners in 2012. Thus, for an average Boston enrollee with no previous Partners use, the estimates indicate a modest $6-8 monthly value of Partners access. This positive but modest average value of broader networks is consistent with the findings of Ho (2006, 2009), who estimated a similar model for employer-sponsored insurance. However, this average masks significant heterogeneity both in the network utility of Partners and in the marginal utility of money. In addition, I estimate substantial coefficients on the direct measure of whether a plan covers an enrollee's previously-used hospitals. For non-Partners hospitals, I estimate an additional value of $5.41 per month and for Partners hospitals, the total effect is $17.04 per month.

[57] Because prices are subsidized, there are two ways to convert this semi-elasticity into an elasticity. Relative to consumers' relatively low premiums (which average about $45 for premium payers), the elasticity averages a relatively modest -1.35. However, relative to insurers' full prices (about $400 on average) – the statistic relevant for insurers' markups – the demand elasticity is -11.9.

Table 1.7 Insurance Plan Demand Estimates

VARIABLE	Coeff.	Std. Error	
Premium: Avg. Coeff. (*normalized*)	**-1.000*** **	**(0.025)**	
x 0-100% Poverty -- *Omitted (no prems.)*	---		
x 100-150% Poverty	-1.340***	(0.038)	
x 150-200% Poverty	-0.935***	(0.024)	
x 200-250% Poverty	-0.712***	(0.015)	
x 250-300% Poverty	-0.656***	(0.016)	
x Age/5 (average effect)	0.029***	(0.002)	
Hospital Network			
Network Utility x <100% Poverty	6.355***	(0.885)	
Network Utility x 100-150% Poverty	7.371***	(0.939)	
Network Utility x 150-200% Poverty	7.453***	(0.962)	
Network Utility x 200-250% Poverty	7.736***	(1.270)	
Network Utility x 250-300% Poverty	8.541***	(1.878)	
Past-Used Hospitals Covered (share)	5.411***	(0.836)	
x Past-Used Partners Hospitals	11.631***	(0.773)	
Switching and Inertia			Passive Prob.
Average Inertia Coefficient	96.810***	(0.230)	94.6%
x Drops Past-Used Hospital (Non-Partners)	-29.905***	(1.142)	75.2%
x Drops Past-Used Hospital (Partners)	-51.048***	(0.962)	51.8%
Plan Brand Effects (average)			
BMC HealthNet (normalized)	0.000	---	
CeltiCare	-23.088***	(0.890)	
Fallon	14.021***	(1.023)	
Neighborhood Health Plan	-2.199***	(0.251)	
Network Health	-3.822***	(0.337)	
Model Statistics			
R^2 in Share (Area-Income-Year)	0.963		
Model w/ Only Avg. Plan Dummies	0.866		
No. Choice Instances	1,588,889		
No. Unique Individuals	611,455		

* = 5% sign., ** = 1% sign., *** = 0.1% sign.

NOTE: This table shows estimates for the multinomial logit plan choice model described in Section 1.5.3. Premium is the monthly plan price. (In addition to the interactions shown, the full model contains interactions with 5-year age groups and gender.) I normalize the average consumer's premium coefficient to -1.0, so all other coefficients are interpretable as dollar values. Network utility is the consumer-specific expected utility measure for a plan's hospital network, derived in Section 1.5.2. Past-used hospitals coverage is the share of an enrollee's previously used hospitals that a plan covers, with a separate interaction for the star Partners hospitals. Switching and inertia are coefficients on a dummy variable for the current plan. The coefficients are interpretable as "switching costs" in dollars per month; the passive probabilities are the implied share of enrollees who passively stick with their current plan. The plan brand effects are coefficients on dummies for each plan. I show average values; the full model contains region-year- and region-income group-specific plan dummies.

As expected, I find substantial inertia in consumers' plan switching decisions. Rationalizing observed switching rates requires an average switching cost of $96.8 per month, or equivalently, an average 94.6% probability of passively choosing.[58] Though large, these estimates are actually a bit smaller than the average switching costs found in an employer insurance setting by Handel (2013) of $2,032 per year (or $169 per month). What is most interesting for selection on networks is that estimated inertia decreases when a plan drops an enrollee's past used hospital from network. For dropped non-Partners hospitals, enrollees are 19% points less likely to be passive and for Partners hospitals, they are 43% points less likely to be passive. A possible explanation is that when an enrollee's regular provider is being dropped, the provider contacts the patient and encourages them to switch plans. Whatever the reason, this inertia reduction exacerbates adverse selection, consistent with the findings of Handel (2013). Here, the inertia reduction is particularly important because it occurs precisely among some of a plan's most expensive consumers, past patients of the Partners hospitals.

1.6 Structural Model: Insurer Cost and Profit Functions

The adverse selection implications of hospital networks depend on the interaction between demand and costs. In this section, I specify a model for insurer costs and (combining this with demand) derive the insurer profit function. The goal is to capture insurer incentives to cover or exclude high-price star hospitals like those in the Partners system. These incentives depend both on how covering Partners affects individual-level costs and how it affects plan selection by individuals of different costliness.

I start by modeling how individual-level costs would vary in plans with different hospital networks. Section 1.6.1 describes how I model insurer costs for hospital care, which uses my hospital demand model and a set of estimated hospital prices. Section 1.6.2 then presents my model for non-

[58] While these estimates are also picking up unobserved heterogeneity, a simple calculation suggests that the passive probability would still be about 90% with a realistic degree of heterogeneity (based on the 55% rate at which re-enrollees choose the same plan as they had before). If there were 55% persistence among current enrollees who were active choosers, 91% of people must have been making passive choices to explain a 96% non-switching rate.

hospital costs. In Section 1.6.3, I aggregate this individual-level cost model up to the insurer level (using plan demand to predict plan choices) and derive the insurer profit function. Finally, Section 1.6.4 considers model fit and analyzes the 2012 change when Network Health dropped Partners.

1.6.1 Hospital Prices and Insurer Costs for Hospital Care

To model insurer costs for inpatient hospital care, I start from an individual-level model. I condition on each person's set of observed hospitalizations (and their diagnoses) and ask how hospital choices and costs would have changed if the patient had been in a different plan with a different hospital network. An advantage of this approach is that it lets me capture the correlation between hospital use and enrollee attributes (which determine plan selection) in a rich, nonparametric way.[59] Nonetheless, this approach assumes networks do not affect the *number* of hospitalizations, only the hospitals chosen when sick.[60]

I first estimate the prices insurers pay to hospitals for inpatient care using the payment data in the insurer claims. Because actual payment rules are unknown (and likely quite complicated), there is a need for simplification. I follow past work (Gowrisankaran, Town, Nevo 2013) in estimating *average* payment factors that capture proportional differences across hospital-insurer pairs.[61] I estimate a Poisson regression (also known as a generalized linear model with a log link) of the form:

$$E\left[Payment_{i,j,h,t,a} \mid Diag_{ita}, Z_{ita} \right] = \exp\left(\rho_{j,h,t} + Diag_{ita}\lambda + Z_{ita}\gamma \right) \tag{1.9}$$

[59] The potential danger is over-fitting. Because I have a large sample and consider only insurer actions that affect a large set of individuals (prices and coverage of Partners), over-fitting is less of a concern for my purposes.

[60] This assumption is likely a good first approximation but is not perfect. Recent evidence from Gruber and McKnight (2014) finds small reductions in the number of hospitalizations in limited network plans. If applicable in my setting, my model will somewhat understate the cost savings from plans' limiting their networks.

[61] Following convention, I refer to these payment factors as "prices," although they are distinct from the actual negotiated prices. These payment factors capture both price differences and service quantity differences across hospitals (conditional on diagnosis) since both affect insurers' payment differences across hospitals.

where *a* indexes the admission, $Diag_{ita}$ is the principal diagnosis, and Z_{ita} is other patient covariates.[62] The key term is $\rho_{j,h,t}$, which is a coefficient that captures average payment differences across hospitals, insurers, and years.[63] This effect is assumed to be proportional across all types of admissions, which is surely not exactly right but should capture a valid average effect. Appendix A.2 discusses additional details on the hospital price estimation.

I use the estimates of (1.9) to define hospital prices as $\hat{P}_{j,h,t} \equiv \exp\left(\hat{\rho}_{j,h,t}\right)$ and an admission-specific severity measure as $\hat{\omega}_{i,t,a} \equiv \exp\left(Diag_{ita}\hat{\lambda} + Z_{ita}\hat{\gamma}\right)$. I scale $\hat{\omega}_{i,t,a}$ so that its mean is 1.0 and divide $\hat{P}_{j,h,t}$ by the same factor, so it can be interpreted as the hospital price for a patient of average severity. The average prices and severities for the 10 most expensive hospitals are shown in Table 1.2.

I use these severities and prices to model how hospital costs would differ in counterfactual plans and networks. As discussed above, I condition on each individual's observed admissions (or lack thereof) and severities $(\hat{\omega}_{i,t,a})$ and use hospital demand to predict how these admissions shift across hospitals. The hospital costs for individual *i* in year *t* in plan *j* (with network N_{jt}) is:

$$c_{ijt}^{Hosp}\left(N_{jt}\right) = \sum_{a=1}^{NAdmits_{it}} \hat{\omega}_{i,t,a} \cdot \left(\sum_{h} \hat{P}_{j,h,t} \cdot s_{i,d,t,j,h}^{Hosp}\left(N_{jt}\right) \right) \tag{1.10}$$

For most hospitals, I use only the plans' observed networks so hold hospital prices fixed at the estimated values. However, for Partners hospitals, I also consider adding/dropping them and therefore need a counterfactual price model. For this, I use a simple average of prices paid by insurers that actually

[62] For the principal diagnosis, I use the Clinical Classification Software (CCS) dummies defined by the U.S. government's Agency for Healthcare Research and Quality. The additional covariates include age, gender, income, and Elixhauser comorbidity dummies for the secondary diagnoses.

[63] As discussed in Appendix A.2.3, I specify a restricted model for $\rho_{j,h,t}$ to avoid over-fitting for hospital-insurer-year cells with small samples. I allow for flexible hospital-insurer and insurer-year dummies, separately by in- and out-of-network status, plus a separate insurer-year factor for each of the six largest hospital systems.

covered (excluded) the Partners hospital in a given year. The main limitation of this approach is that it does not capture insurer-hospital bargaining dynamics, something I have not yet modeled.[64]

1.6.2 Non-Hospital Costs

I complete the cost model by considering all costs other than inpatient hospital care. Unfortunately, I do not have a provider choice model for non-hospital care through which I could define costs analogously to my hospital cost model. Instead, I take a reduced form approach. I calculate monthly non-inpatient costs for each enrollee-year and use them to estimate the following Poisson regression model:

$$E\left(NonHospCost_{it} \mid Z_{it}\right) = \exp\left(\eta_{j(i),Reg(i),t} + Z_{it}\mu\right) \tag{1.11}$$

where Z_{it} are detailed enrollee diagnoses and demographics.[65] I use these estimates to define a region-year-specific plan effect $\hat{C}_{j,Reg,t} \equiv \exp\left(\hat{\eta}_{j,Reg,t}\right)$, an enrollee severity $\hat{\varsigma}_{it} \equiv \exp\left(Z_{it}\hat{\mu}\right)$, and an enrollee residual $\hat{v}_{it} \equiv NonHospCost_{it} / \left(\hat{C}_{j(i),Reg(i),t} \cdot \hat{\varsigma}_{it}\right)$. If an enrollee switches to plan *k*, I assume that his severity and residual are unchanged but that the plan effect switches to the counterfactual plan, so the enrollee's new cost is $\hat{C}_{k,Reg,t} \cdot \hat{\varsigma}_{it} \cdot \hat{v}_{it}$. This reduced form approach is clearly an approximation. However, the $\hat{C}_{j,Reg,t}$ estimates should capture a valid average plan effect on costs absent unobserved cost-based selection into plans. Given that I have documented unobserved selection based on the exchange's risk adjustment, this assumption is clearly imperfect.[66] If there is residual selection, I will understate costs for plans attracting residually healthier enrollees and overstate costs in the opposite case.[67] This will affect my estimates of

[64] Two facts suggest this approach may be a reasonable approximation in this setting. First, within-year price variation across insurers for the main Partners hospitals is small in practice – standard deviations for Brigham and Mass. General are just $359 and $846, respectively. Second, when Network Health drops Partners, I see little change over the next two years in Partners prices paid by the plans that still cover it.

[65] For diagnoses, I use the Hierarchical Condition Categories (HCC) defined by Medicare for its risk adjustment. I use HCCs observed in the *current* plan year so I can include diagnoses for new enrollees.

[66] The covariates in (2.11) will do somewhat better than the exchange risk adjustment because they include concurrently observed diagnoses, which allows for including diagnoses for new enrollees.

[67] To address this potential bias, I plan in a future revision to instrument for plan enrollment using the *timing* when an enrollee entered the exchange. Because of inertia, enrollees who enter just before a price change will have

the *level* of non-inpatient costs at observed networks but not the cost *difference* from network changes, which I specify separately next.

Networks may affect non-inpatient costs both through outpatient hospital care and through secondary effects on services like drugs and post-acute care. For the effect of adding/dropping Partners, I again specify a reduced form adjustment.[68] I first use the hospital cost model to calculate inpatient costs with and without the network change. I then assume that a plan's non-inpatient costs change in proportion to the *regional average* change in inpatient costs.[69] The final non-inpatient cost model is:

$$c_{ijt}^{NonHosp}\left(N_{jt}\right) = \underbrace{\hat{C}_{j,Reg,t}}_{\text{Plan Effect}} \cdot \underbrace{\left(\hat{\varsigma}_{it} \cdot \hat{\nu}_{it}\right)}_{\text{Ind. Severity and Residual}} \cdot \underbrace{\left(1 + \lambda \cdot \%\Delta HospCost_{j,Reg,t}\left(N_{jt}\right)\right)}_{\text{Network Cost Adjustment}} \tag{1.12}$$

Based on a risk-adjusted regression at the plan-region-year level, I find that each 10% increase in inpatient costs is typically associated with a 3.8% increase in non-inpatient costs and therefore set $\lambda = 0.38$. However, I can do robustness checks with alternate values of λ.

1.6.3 Total Costs and Insurer Profits

With a model for both individual-level inpatient hospital and other costs, I sum them to define total costs, $c_{ijt}^{Total}\left(N_{jt}\right)$. I also include in total costs a measure of variable plan administrative costs (e.g., for claims processing) based on plan financial reports to the exchange.[70] The final model step is to aggregate costs and revenue up to the plan level using the demand function. The annual profit function for plan j is:

$$\pi_{jt} = \sum_{i}\left(\varphi_{it}P_{jt} - c_{ijt}^{Total}\left(N_{jt}\right)\right) \cdot D_{ijt}\left(\mathbf{Prem}, \mathbf{N}\right) \tag{1.13}$$

---

different plan shares at a later date t than enrollees who enter just after the price change. Assuming that entry timing does not independently affect costs and that attrition is independent of unobserved costs, then entry timing in the exchange is a valid instrument for current plan enrollment.

[68] Past structural work on hospital networks has generally either ignored non-inpatient costs or assumed that they did not change with the hospital network. My reduced form method, though imperfect, improves on the past literature.

[69] A limitation with this method is that it does not capture *differential* percent changes for the people most likely to use Partners. In future revisions, I could consider using Network Health's 2012 dropping of Partners to estimate a more flexible model of non-hospital cost changes that varies with individual covariates.

[70] To do so, I estimate a regression of plan's reported administrative costs on their total enrollment. I find an almost perfect linear fit with a coefficient of about $30 per member-month, which I use for the model.

where P_{jt} is the plan's price, φ_{it} is the exchange's risk adjustment score for enrollee *i*, and $D_{ijt}(.)$ is the enrollee's demand for plan *j*. Demand is in units of member-months and is the product of two terms:

$$D_{ij}(\mathbf{Prem}, \mathbf{N}) = nMon_i \cdot S_{ij}(\mathbf{Prem}, \mathbf{N})$$

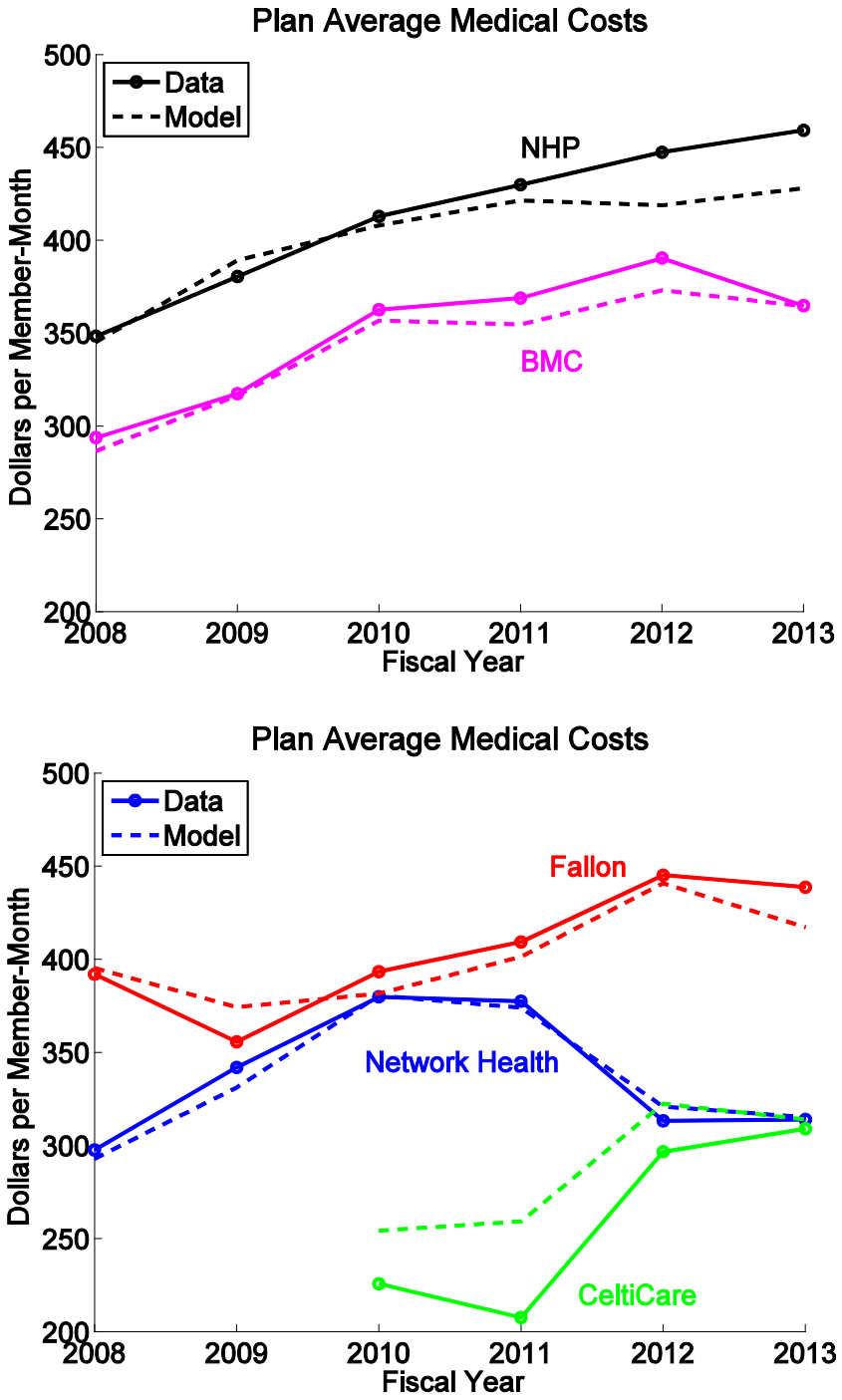
The first is the number of months an individual is enrolled in the exchange during the year. Many enrollees enter or leave in the middle of the year (e.g., because of a change in jobs that affects their eligibility), and I assume this enrollment churn is exogenous and hold $nMon_i$ fixed as observed. The second term is consumer *i*'s predicted share for plan *j* from the logit demand system.

1.6.4 Model Fit and Analysis of 2012 Network Health Change

Figure 1.6 shows the model fit for plans' average monthly medical costs per enrollee. The model averages are calculated using the model's cost and demand functions (as in (1.13)), creating two potential sources of errors versus the costs in the data. Nonetheless, the fit is quite good, with an R^2 at a plan-year level of 0.926. Importantly, the model captures very well the large fall in costs for Network Health in 2012 when it dropped Partners. The largest errors are predicting too high costs for CeltiCare in 2010 and 2011 (when it was a new plan and had very low enrollment), although the model does capture its large cost increase in 2012 after Network Health dropped Partners.

I next consider in more detail how well the model matches the cost and demand patterns for Network Health in 2012. Appendix A.3 shows a series of figures and tables similar to those analyzed in Section 1.4, with values predicted by the model added on. The model captures the variations in switching rates among Network Health's current enrollees quite well. Past Partners' patients switching rate is matched almost perfectly – since the model's interaction of switching costs with dropping Partners is largely identified from the 2012 change. It also captures the intermediate level of switching for patients of other dropped hospitals.

Figure 1.6 Model Fit for Plan Average Medical Costs



NOTE: These figures show each plan's average medical costs (per member-month) in each fiscal year in the data (solid lines) and in the model (dashed lines). The cost model, described in Section 1.6, is an individual-level model of costs, which is aggregated up to the plan level based on predicted shares from the plan demand model.

The next table shows how the model matches the cost change from 2011 to 2012. For the average costs and cost changes, the model matches almost perfectly. Breaking it into enrollee subgroups, the model captures the basic pattern that enrollees who left the plan in 2011 were much more expensive and that the cost decrease for stayers was smaller than the overall decrease, though it slightly underestimates the former and overestimates the latter.

The final set of figures analyze how the hospital model captures changes in admission shares and costs at Partners and other dropped hospitals.[71] In all cases, the fit is quite good. In particular, the model matches the striking fact that Partners admissions fell for Network Health, rose at other plans, and barely changed overall. It also matches Network Health's and other plans' costs per hospital admission in levels and trends (including the 15% drop for Network Health in 2012).[72]

Finally, I use the model to decompose how much of the 15% decline in Network Health's risk-adjusted costs was due to selection versus "real" cost reductions. One indication that selection played a large role is that costs declined just 6% on a fixed population of stayers in the plan in both 2011 and 2012 (see Table 1.5). However, this statistic does not capture the full effect of real cost cuts, which would also have applied to the people who switched plans had they not left. Instead, I use the model to decompose how changes in plan selection versus changes in the cost function affected costs. Formally, I can decompose the 2011-2012 change in costs into:[73]

$$Cost_{2012} - Cost_{2011} = \underbrace{\sum_i \left(c_{ij,2012} - c_{ij,2011}\right) \cdot D_{ij,2012}}_{\text{Cost Function Change}} + \underbrace{\sum_i \left(D_{ij,2012} - D_{ij,2011}\right) \cdot c_{ij,2011}}_{\text{Selection}}$$

where $j = \text{Network Health}$. Based on this decomposition, I find that selection explains 50% of Network Health's reduction in costs, with the rest due to a lower cost function for a fixed population. Notice that

[71] To focus on the hospital demand and cost model's ability to fit patterns, these figures condition on people's actual plans, rather predicting plan shares using the plan demand model.

[72] I have found that including the past hospital use variables in the hospital demand system are important to matching these patterns so well. Without these covariates, for instance, the model cannot match the sharp rise in Partners admissions for plans other than Network Health in 2012.

[73] Because this decomposition requires observing individuals in both years, I restrict the sample accordingly.

this decomposition calculates the cost function effect with 2012 shares and the selection effect with 2011 costs. If instead, I calculate the cost function effect with 2011 shares, the cost reductions are larger, and selection explains 36% of the decline. This difference implies that many of the people whose costs would have declined the most selected out of Network Health in 2012. Selection attenuated the cost-reducing effects of a change in networks. Either way, however, selection was important, explaining between 36-50% of Network Health's cost reduction.

1.7 Model Analysis: Heterogeneity in Value and Cost of Partners

Having estimated the model of insurance and hospital demand, I use the estimates to study heterogeneity in consumers' costs and value of Partners coverage. For simplicity, I focus on current enrollees in the exchange at the start of 2012, when Network Health dropped Partners. I define utility for Partners based on the difference in plan utility for Network Health, excluding switching costs (U_{ijt} in equation (1.8)), with and without Partners covered. I convert utilities into dollar values by dividing by each individual's marginal utility of money (the negative of their premium coefficient).[74] I calculate costs based on the cost function for Network Health with and without Partners.

Table 1.8 shows these estimates for all current enrollees in the exchange at the start of 2012.[75] The rows are sorted by the measure of Partners value. About 80% of enrollees have relatively little value for Partners coverage, with a monthly value of $4.30 or less – quite small compared to the typical variation in plan premiums of $20-60 per month. But value for Partners rises sharply in the top 10-20% of enrollees, with the top 5% valuing Partners at $46.80 per month. For these enrollees, almost all of whom are past Partners patients, Partners coverage plays a determinative role in their plan choices.

The remainder of Table 1.8 shows how these differences in value for Partners correlate with costs. I distinguish between two sources of adverse selection discussed in the theory in Section 1.2:

[74] I exclude below-poverty enrollees from this calculation because I cannot estimate their premium coefficient.

[75] I focus on current enrollees because their past hospital use (a key model covariate) is more likely to be observed.

selection on unobserved risk and selection driven by use of Partners' high-price care. Columns (2)-(3) suggest that unobserved risk is important. Even without Partners covered, people in the top decile of Partners value have risk-adjusted monthly costs of about $350, which is $50 (or 17%) higher than those who value Partners the least. Column (5) indicates that selection on use of Partners is also important. The $\Delta Cost$ from covering Partners rises from $8.0 (2.7%) for the lowest-value group to $48.5 (10.0% of a larger base) for the highest-value group. Combining both types of selection, the people in the top decile of Partners values are $84 (or 27%) more expensive (after risk adjustment) in a plan covering Partners than people with below-median values. Of this $84 difference, about 60% is due to selection on unobserved risk and the remainder due to differential use of the Partners system.

A final insight from Table 1.8 is that for each group, the estimated consumer value from access to Partners falls short of the increase in insurer costs. Even aside from adverse selection, this fact gives insurers a strong incentive to drop Partners. However, this does not prove that the welfare effect of covering Partners is negative for all groups. Part of plans' higher costs represent markups to the Partners hospitals, which may be used for socially valuable purposes like teaching and research. To account for these markups, I draw on a Massachusetts government estimate of the per-admission costs of Partners (CHIA 2014).[76] Based on these estimates, the cost per admission at the two star Partners hospitals in 2012 were about $12,500 (Mass. General) and $13,800 (Brigham), implying margins of about 30-35% relative to my estimated prices. Column (7) shows the net cost increase, subtracting the change in Partners net revenue for inpatient care from insurer costs.[77] After doing so, the net cost increase for people in the top decile of Partners values is substantially lower. Their value for Partners coverage now exceeds the estimate of net costs. However, value still falls short of net costs for people in the bottom 90% of Partners valuations.

[76] The measure is of hospitals' "inpatient cost per case mix adjusted discharge". The calculation, which is based on hospitals' cost reports to the state, is intended to be a comprehensive measure of average hospital costs (including fixed costs of facilities), excluding physician compensation and graduate medical education costs.

[77] Note that this values each $1 of Partners net revenue as $1 of social value. This calculation is imperfect because it excludes Partners markups for non-inpatient care and reductions in net revenue for non-Partners hospitals. The latter, however, are likely to be small; non-star hospitals often have low or negative margins (Ho 2009).

Table 1.8 Model Estimates: Relationship between Value and Cost of Partners Coverage

Consumer Value of Partners Covg.		Costs to Insurer (per month)					
		Not Covering Partners		Covering Partners			
Percentiles	Avg. Value ($/month)	Unadjusted Cost	Risk Adj. Cost	Risk Adj. Cost	ΔCost	%Δ	ΔCost - Partners Hospital Mkup.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
0-50%	$0.5	$300.0	$301.2	$309.2	$8.0	2.7%	$7.0
50-70%	$2.2	$269.6	$294.5	$308.6	$14.0	5.2%	$10.6
70-79%	$4.3	$264.3	$292.7	$310.8	$18.1	6.8%	$12.4
80-89%	$8.8	$300.1	$311.8	$335.3	$23.5	7.8%	$14.0
90-95%	$23.6	$455.7	$360.4	$398.3	$37.9	8.3%	$21.1
96-100%	$46.8	$482.3	$340.1	$388.6	$48.5	10.0%	$23.3
Average	***$5.7***	***$308.8***	***$305.6***	***$321.2***	***$15.6***	***5.0%***	***$10.6***

NOTE: This table shows the estimated model's implication for the relationship between enrollees' costs and their value coverage of the star Partners hospitals – the key relationship driving adverse selection. Consumers are sorted into percentiles of Partners value, and each row shows average values and costs for people in the relevant percentiles. All values and costs are calculated for current enrollees in 2012 based on the value and cost if Partners were added to Network Health plan's network. Value of Partners is defined as the extra plan utility (excluding switching costs) if Partners is covered, divided the marginal utility of money – based on plan utility estimates shown in Table 1.7. Because I cannot estimate marginal utilities for below-poverty enrollees, they are excluded. Costs are defined using the estimated cost function without Partners covered (columns 2-3) and with it covered (columns 4-7), both based on the plan cost model in Section 1.6.3. Column 7 subtracts from the increase in cost an estimate of how much of these higher costs are funding higher markups for Partners. The table shows that most enrollees value Partners coverage little, but the top 10-20% values Partners substantially. The table also decomposes two different reasons people with high values for Partners are high-cost. First, they have higher risk adjusted costs even if Partners is not covered, which suggests they are unobservably sicker. Second, they have a larger increase in costs when Partners is covered (column 5) because they use Partners hospitals more often.

1.8 Equilibrium and Analysis of Policy Solutions

This section uses the demand and cost estimates to simulate equilibrium in a model of insurance competition. I use this to examine the impact of different policies used to address adverse selection in insurance exchanges. In general, insurer competition on prices and hospital networks is extremely

complicated and subject to multiple equilibria.[78] To make progress, I focus on a static model where insurers compete only on price and coverage of the expensive Partners hospitals, holding hospital-insurer prices and other aspects of the network fixed. Although stylized, this model goes beyond most past empirical work on selection, which studies pricing holding fixed product characteristics.[79]

1.8.1 Equilibrium Simulations: Method and Results

Consider a model of insurance market equilibrium for a particular year (e.g., 2012) in the Massachusetts exchange. As in Massachusetts, I assume that each insurer offers a single plan with exchange-specified consumer cost sharing and covered service rules.[80] I condition on the plan's past history, including past network coverage and the set of current enrollees entering the year. I also hold fixed (at observed values) each plan's network and payment rates for all non-Partners hospitals. Before the year, the exchange announces policies (e.g., subsidy and risk adjustment rules). Insurers then compete in the following game:

Insurer Competition: 1. Insurers choose whether to cover Partners hospitals

2. Insurers set plan prices

Consumer Demand: 3. Consumers choose plans

4. Sick consumers choose providers (based on plan network)

I assume that insurers observe networks from stage 1 when setting prices and that they have full information on all demand and cost functions.

Insurers make choices to maximize profits, following the model for demand, costs and profits estimated in Sections 1.5-1.6. However, there is one additional simulation issue: how to incorporate

[78] For an innovative model incorporating hospital-insurer bargaining and network formation in a dynamic equilibrium framework, see Lee and Fong (2013).

[79] For example, see Einav, Finkelstein and Cullen (2010), Starc (2014), and Handel, Hendel, and Whinston (2013). One recent paper by Einav, Jenkins and Levin (2012) does consider the effect of selection on product design in consumer credit markets but does so in a setting with a monopolist firm.

[80] In the ACA, insurers can offer multiple plans with varying networks across four tiers of cost-sharing generosity. Unfortunately, my data (in which all plans have the same cost sharing) do not make it possible to model cost-sharing differentiation. However, in future analysis, I could study equilibrium when insurers can offer two plans that vary in whether they cover Partners.

dynamics arising because of enrollee inertia. When a plan lowers its price and attracts more enrollees today, it increases its future demand because some enrollees will passively stick with the plan in following years. This can lead to an invest-then-harvest equilibrium in which plans cycle between low and high prices. I choose not to specify a fully dynamic model, which would be both complicated and unrealistic unless it modeled uncertainty about policy changes (which occurred frequently in Massachusetts). Instead, I take a simple static approach that approximates a dynamic model. I assume that enrolling someone today increases future profits in proportion to the person's future duration in the market and the current profit margin on the individual. This "future profit effect" gives insurers an added incentive to keep prices low and helps offset the lower price elasticity of demand due to inertia. Appendix A.4 shows the modified pricing first-order conditions and lays out additional details for the simulation method.

In full-information Nash equilibrium, each insurer sets prices in step 2 to satisfy its first-order conditions given all other insurers' prices and networks. In step 1, they choose Partners coverage knowing the pricing equilibrium that will prevail for each network possibility. For Partners coverage, I assume a binary choice: either sticking with their actual coverage of Partners or adding/dropping all of the Partners hospitals. I do not model the vertical relationship between Partners and Neighborhood Health Plan (NHP) but allow it to flexibly cover/drop Partners.[81] Nash equilibrium occurs at a set of networks $\mathbf{N}$ if no insurer wishes to unilaterally deviate: $\pi_j\left(N_j, \mathbf{N}_{-j}\right) \geq \pi_j\left(\tilde{N}_j, \mathbf{N}_{-j}\right) \ \forall \tilde{N}_j, j$. While there is no guarantee of a unique equilibrium, I do not find multiplicity in my main results.

[81] To simplify, I also hold fixed the observed choice of Fallon (which is not available in most of the Boston area) not to cover Partners.

Table 1.9 Simulation Results

Equilibrium Simulation Results

Panel A: Mass. Exchange Population & Policies (2011)

			Insurance Plan				
Source	Year	Variable	BMC	Fallon	Network Hlth	NHP	CeltiCare
Observed	2011	Partners Covg.	No	No	Yes	Yes	Yes
		Price*	$424.6	$425.7	$422.6	$425.7	$404.9
Simulated	2011	Partners Covg.	No	No	No	No	Yes
		Price*	$425.7	$425.7	$425.7	$425.7	$404.9

** Exchange imposed maximum price of $425.7 and minimum price of $404.9*

Panel B: ACA-Like Population & Policies

			Insurance Plan				
Source	Year	Variable	BMC	Fallon	Network Hlth	NHP	CeltiCare
Simulated	2011	Partners Covg.	No	No	No	No	No
		Price	$407.2	$409.3	$389.4	$402.5	$318.8
Simulated	2012	Partners Covg.	No	No	No	No	No
		Price	$427.5	$464.5	$371.0	$417.6	$365.0
Simulated	2013	Partners Covg.	No	No	No	No	No
		Price	$437.2	$476.8	$432.9	$461.8	$419.4

NOTE: These tables show equilibrium results for the insurance market simulations described in Section 1.8.1. In the game, insurers first simultaneously choose whether or not to cover the Partners hospitals (holding fixed other hospital coverage) and then simultaneously choose their plan's price. The tables show their equilibrium choices of Partners coverage and price. Panel A shows simulations using the Massachusetts exchange's actual enrollee population and policies for 2011 – including required minimum and maximum prices – and compares simulated coverage and prices to the observed values. I do this comparison only for 2011 because of complications with analyzing other years. As discussed in Section 1.8.1, the model matches prices well but predicts even more dropping of Partners than actually occurred (although Network Health dropped Partners the following year).

Panel B conducts simulations with a population and policies closer to those in the ACA exchanges. Specifically, I exclude enrollees below poverty (who get Medicaid in the ACA), set subsidies as a flat amount for all plans (versus Massachusetts' higher subsidies for higher-price plans), and do not impose minimum and maximum prices. In these simulations, no insurer chooses to cover Partners partly because doing so attracts enrollees with high risk-adjusted costs and therefore lowers profits.

Table 1.9 shows equilibrium insurer choices for several simulations.[82] The top panel shows equilibrium under the actual Massachusetts subsidy and pricing policies in 2011, comparing these to the observed prices and networks.[83] The model's prices match extremely well. But this occurs largely because Massachusetts had a narrow allowed price range, and all plans bid at or near the range's min or max.[84] Nonetheless, the model captures well which insurers priced near the min versus the max. For networks, the model predicts just one plan (CeltiCare) willing to cover Partners, while in reality Network Health and NHP also covered Partners in 2011. However, Network Health did drop Partners in 2012, and Partners announced intentions to buy NHP in August 2011, a factor that I do not model. It is interesting that the model can rationalize CeltiCare's surprising decision (as the low-price plan) to cover Partners. In the model, CeltiCare is willing to do so because of the binding price floor. Without a price floor, CeltiCare instead cuts its price and drops Partners.

Because many of Massachusetts' distinct rules did not continue under the ACA, I perform the rest of the analysis using rules closer to those in the ACA. Specifically, I include only the 100-300% poverty population (those below poverty generally get Medicaid in the ACA), set subsidies as a flat amount for all plans, and do not impose minimum or maximum prices.[85] Panel B of Table 1.9 shows the simulation results. Under ACA-like policies, the model predicts that all plans drop Partners, and this result is robust across all the simulation years, 2011-2013. When an insurer deviates to cover Partners, its costs go up for

[82] To speed computing time, all of the simulations I report here have been conducted on a 10% random sample of enrollees. I will perform the simulations on the full sample in future revisions.

[83] While I would like to perform a similar model fit test for other years, data limitations and policy complexities make this difficult. Prior to 2010, the pricing process was much more complicated and involved some negotiation with the exchange. In 2010, I am missing the risk adjustment scores. And in 2012-13, the exchange introduced a limited choice policy that creates auction-like dynamics that I have not yet modeled.

[84] Massachusetts used maximum prices to lower costs given that it fully subsidizes below-poverty enrollees for any plan they choose. Minimum prices were imposed by federal actuarial soundness rules, which are designed to prevent insurers from pricing so low that they are unable to pay for the required medical benefits.

[85] The main remaining differences with the ACA are the lack of higher-income unsubsidized enrollees (who represent about 20% of ACA enrollees) and the absence of multiple plans per insurer across four cost-sharing generosity tiers. There is not much I can do to incorporate these factors, since I do not have data on higher income people or a way or estimating preferences for different levels of cost sharing. Therefore, the simulations should be seen as illustrative of the economic forces involved, not a prediction of what will occur in the ACA.

all of its enrollees, and it particularly attracts the enrollees who most value Partners and whose risk-adjusted costs are high. But by raising its price to compensate, it reduces demand among a large number of lower-cost enrollees. As a result, total profits go down when a plan covers Partners.

1.8.2 Welfare Analysis

An important question is whether this unravelling is socially inefficient. Answering this requires a welfare function, which is not obvious to define in this setting. My starting point is a social surplus approach, in which welfare equals consumer plan value (plan utility divided by marginal utility of money[86]) minus insurer costs. But I make several adjustments. First, I choose to exclude the switching cost, treating them as pure inattention. Recall that I estimated that switching costs were much lower when a plan dropped a consumer's hospital, and I do not want the welfare analysis to be driven by this difference.[87] Once I exclude switching costs, however, the standard inclusive value formula for expected utility in a logit model does not apply. Instead, I define expected plan value for consumer *i* as:

$$ConsValue_i = \frac{1}{-\alpha_i} \sum_j \hat{s}_{ij}^{Plan} \cdot \hat{U}_{ij}$$

where α_i is the premium coefficient, $\hat{s}_{ij}^{Plan}$ is the model's predicted share for consumer *i* choosing plan *j*, and $\hat{U}_{ij}$ is plan utility excluding switching costs and the logit error.

A second adjustment to social welfare is that I allow for an excess cost of government subsidies, to reflect the distortionary cost of tax financing. As a baseline, I assume an excess cost of government funds (ECF) of 30%, but I also consider an ECF of zero as in a textbook social surplus calculation. Finally, I add to social welfare an estimate of the markup of Partners' hospital prices above cost, based on the Massachusetts government estimate (discussed above in Section 1.7). As a starting point, I value each

[86] The marginal utility of money is the negative of the premium coefficient in the plan demand system. I do not need to worry about the premium coefficient for the below-poverty group (which I could not estimate) because they are excluded from the ACA-like population.

[87] I have also done the welfare analysis with switching costs included. The results are qualitatively similar, but past Partners' patients value of coverage is higher because of the switching cost interaction. However, this difference is not enough to change the net result of the welfare calculation.

dollar of markup as $1 of social welfare, although alternate assumptions are possible. How to value these markups depends on the social value of the hospital activities they fund, including teaching, research, and uncompensated care.

Table 1.10 compares welfare at the ACA-like equilibrium for 2012 (with no plans covering Partners) to a hypothetical in which all plans (except Fallon) cover Partners. The latter captures the hypothetical effect of ensuring full coverage of Partners (e.g., via a coverage mandate) if this could be done without Partners raising its prices. Plans, however, can adjust their prices, and consumers can re-sort across plans. Overall (in the last column), Partners access increases plan value by a relatively small $5.7 per member-month and increases costs (net of the Partners markup) by a slightly larger $11.2. Thus, on average, covering Partners lower net welfare, even without considering the increase in government costs. However, the averages mask important heterogeneity between people who strongly value Partners (again, proxied by past Partners use) and others. Past Partners patients value access by $30.2 per month, an order of magnitude more than all others. Although past Partners patients' net cost increases are also larger ($26.6 per month), the value gains are large enough to produce a small net welfare increase of $3.6 per month from giving them access. By contrast, all other consumers value Partners access relatively little but once Partners is covered, they use it and increase costs non-trivially. Because this broader group is 90% of the population, their effect dominates and social surplus falls.[88]

[88] One might expect these results to be different if just one or two plans covered Partners, since consumers could sort based on Partners preference. However, I find results to be qualitatively similar. Consumers choose plans based on many factors, so there are still consumers who value Partners little but end up in a plan covering it.

Table 1.10 Welfare Analysis of Partners Coverage

Welfare Analysis of Partners Coverage

ACA-Like Policies & Population (Sim. Yr. 2012), Units = $/member-month

Statistic		Enrollee Group: Past Partners Patients	Enrollee Group: All Others	Overall
Share of Enrollee Months		9.7%	90.3%	100.0%
Consumer ΔPlan Value		***$30.2***	***$3.0***	***$5.7***
Insurer Costs	No Partners Covg.	$521.8	$342.6	$360.0
	All Cover Partners*	$563.7	$355.6	$375.9
	Difference	***$41.9***	***$13.1***	***$15.9***
Partners Net Inpatient Revenue	No Partners Covg.	$3.5	$0.3	$0.6
	All Cover Partners*	$18.9	$3.8	$5.2
	Difference	***$15.3***	***$3.5***	***$4.6***
	Net Cost Difference	***$26.6***	***$9.6***	***$11.2***
Net Value - Cost of Partners Covg.		***$3.6***	***-$6.5***	***-$5.6***
Govt. Subsidy Cost Increase		*$26.2*	*$19.8*	*$20.4*
Additional Cost of Govt. Funds (ECF = 0.3)		***-$7.8***	***-$5.9***	***-$6.1***

** Except Fallon, which does not operate in most of the Boston area.*

NOTE: This table shows a welfare analysis of plans' coverage of Partners hospitals described in Section 1.8.2. It compares the consumer plan value and costs in two scenarios: no coverage of Partners and full coverage of Partners by all plans except Fallon (which does not operate in most of Boston). In each case, plan prices adjust into Nash equilibrium and consumers re-sort across plans based on the demand system. I show all values for the overall population and separately for past Partners patients and all other enrollees. Consumer plan value is defined in Section 1.8.2 and equals expected plan utility (excluding switching costs), divided by the marginal utility of money. Insurer costs are defined using my estimated cost function, and Partners net inpatient revenue is defined based on the hospital demand system, its estimated prices, and its per-admission cost estimates from the Massachusetts state government (CHIA 2014). The net cost difference equals the change in insurer costs, minus the change in Partners' net revenue, and the net value minus cost difference is consumer value minus this net cost. Overall, the results show that covering Partners increases value net of costs for past Partners patients, but raises costs for all others without increasing value much.

These welfare results highlight the fundamental issues involved with coverage of star hospitals. On the one hand, the group preferring Partners is high cost, so insurers have too great an incentive to avoid Partners because of adverse selection. On the other hand, covering Partners creates moral hazard among people who do not value it much but for whom expensive Partners hospitals are now available at no extra fee. Since Partners is very costly, the moral hazard costs are large, and reversing unravelling is

not socially optimal. The welfare calculus, however, might be different if Partners were less expensive, since moral hazard would be smaller but adverse selection would still deter covering Partners.

1.8.3 Policy Counterfactuals

The welfare analysis above studied a hypothetical in which plans covered Partners despite it being profitable to drop them. In reality, reversing unravelling would require policy changes to address adverse selection. In this section, I examine two policies to offset adverse selection and encourage coverage of the Partners hospitals: modified risk adjustment and subsidies. I examine how plans' prices and Partners coverage decisions change under alternate policies, continuing to hold Partners' hospital prices fixed.

The first policy modifies risk adjustment by increasing how much it compensates for high-risk types and reducing it for low risks – a form of the "over-payment" that Glazer and McGuire (2000) find to be optimal for risk adjustment. The logic for over-payment is that plans covering Partners attract consumers who are both observably and unobservably high-cost. The modified risk adjustment over-pays based on observed risk to compensate for the high unobserved risk of enrollees in plans covering Partners. To implement this, I multiply all risk scores above the mean by a factor $(1+\phi)$, divide all below-mean risk scores by the same factor, and renormalize the distribution to be mean 1.0. The potential downside of this policy is that insurers have incentives to avoid covering people with low observed risk (e.g., young people). If low risks are more price sensitive (as I found for young people in the plan demand estimates), insurers will respond by raising prices and markups.

The top of Table 1.11 shows the simulation results for modified risk adjustment. A ϕ of 50% is sufficient to reverse the unravelling, with NHP choosing to cover Partners. The change increases consumer surplus (by \$5.4 per member-month), insurer profits (by \$6.9), and Partners net revenue (by \$1.1, about one-fourth of the increase as when all insurers cover Partners in Table 1.10). However, the largest change is in government subsidy costs, which increase by \$14.4 per member-month (or 4.4%). Government costs increase because subsidies are set based on the lowest plan's price, which rises from

$365 to $381.[89] The low-price plan (CeltiCare) tends to selects low-risk people, and the modified risk adjustment penalizes it more for doing so. In addition, it has less incentive to keep markups low to attract the healthy, as discussed above. Therefore, CeltiCare raises its price. The cost of higher subsidies depends on whether there is an excess marginal cost of government funds (ECF). If there is no excess cost (ECF = 0), this is a pure transfer, and social surplus changes only slightly. With a more typical ECF of 30% (the final column in Table 1.11), social surplus falls more substantially.

I consider a second policy to address adverse selection: differentially subsidizing high-price plans. Rather than a fixed subsidy S_0 for all plans, a plan's subsidy equals $S_0 + \sigma \cdot \left(P_j - \min_k P_k\right)$, which is linked to its price P_j. I call this policy "marginal subsidies" because a plan's subsidy increases on the margin as it raises its price. Marginal subsidies decrease plans' incentive to compete on prices and therefore increase the incentive to raise quality (here, Partners coverage) – as shown by the classic analysis of Dorfman and Steiner (1954). Marginal subsidies also decrease the inefficiently high premiums a plan covering Partners charges because of selection, which I highlighted in the theory in Section 1.2. The downside is that plans have greater incentive to markup prices, regardless of whether they cover Partners.

The bottom part of Table 1.11 shows the simulations for marginal subsidies. Although conceptually different, subsidies have similar qualitative effects as risk adjustment. Marginal subsidy rates exceeding 25% induce BMC plan and at 50%, also NHP to cover Partners. Consumer surplus, insurer profit, and Partners net revenue increase at the expense of higher government spending. Relative to risk adjustment, however, consumer surplus increases less and insurer profits increase more. A key difference is the pattern of price increases across plans. With risk adjustment, the low-price plan raises its

[89] These simulations follow the Massachusetts rule of setting subsidies so that the cheapest plan's premium equals a pre-specified affordable amount. The ACA sets subsidies based on the second-cheapest silver-tier plan, which I do not follow because I do not have plans across multiple generosity tiers. Note that the increase in the lowest plan price is slightly larger than the increase in subsidies because the risk adjustment is not perfectly budget neutral.

Table 1.11 Counterfactual Policy Simulations

Risk Adjustment

Over-Adjustment Factor	Plan Statistics			Welfare Analysis (per member-month)					
	Covering Partners	Minimum Price	Avg. Price Other Plans	ΔCons. Surplus	Insurer Profit	Partners Net Rev.	Govt. Costs	ΔSocial Surplus ECF = 0	ECF=0.3
None	None	$365.0	$420.1	$0.0	$26.5	$0.6	$322.7	$0.0	$0.0
25%	None	$374.5	$420.9	$4.1	$30.0	$0.6	$330.7	-$0.4	-$2.8
50%	NHP Only	$381.3	$426.4	$5.4	$33.4	$1.7	$337.1	-$1.0	-$5.3

Marginal Subsidies

Marginal Subsidy Rate	Plan Statistics			Welfare Analysis (per member-month)					
	Covering Partners	Minimum Price	Avg. Price Other Plans	ΔCons. Surplus	Insurer Profit	Partners Net Rev.	Govt. Costs	ΔSocial Surplus ECF = 0	ECF=0.3
None	None	$365.0	$420.1	$0.0	$26.5	$0.6	$322.7	$0.0	$0.0
15%	None	$368.8	$427.6	$0.7	$33.4	$0.6	$331.1	-$0.8	-$3.3
25%	BMC Only	$372.1	$435.9	$0.7	$39.5	$1.0	$338.8	-$1.9	-$6.8
50%	BMC + NHP	$384.5	$469.4	$2.5	$65.5	$2.4	$370.2	-$4.1	-$18.4

NOTE: This table shows results of simulations of counterfactual policies to address the adverse selection, as discussed in Section 1.8.3. The top table shows simulations that modify risk adjustment by over-paying by the listed "over-adjustment factor" for people with above-average risk scores (and under-paying by the same factor for below-average risks). The bottom table shows simulations with "marginal subsidies" that narrow price differences across plans by the listed marginal subsidy rate. All simulations are for the ACA-like population and policies in 2012, so the baseline results (in the top row of each table) are the same as the 2012 equilibrium in Table 1.9. Each table lists which plans cover Partners, the minimum plan price, and average price of all other plans. They also list welfare statistics in units of dollars per member-month: the change in consumer surplus (with the baseline normalized to $0), insurer profit, Partners' net inpatient hospital revenue, and government subsidy costs. The final columns show the change in social surplus, with an excess government cost of funds (ECF) of either 0 or 0.3. The latter values each $1 of government subsidies as incurring a social cost of $1.3 because of the excess burden of tax financing.

price, while all higher-price plans raise prices relatively little, since they benefit from the greater compensation for their relatively sick consumers. However, with subsidies, all plans increase their prices in tandem. As a result, insurer profits increase a bit more, and social surplus falls a bit more.

This analysis points to a more general tradeoff involved with mitigating adverse selection in settings with imperfect competition, as shown in recent work by Starc (2014) and Mahoney and Weyl (2014). When sicker people differentially choose higher-price plans, insurers have an incentive to keep prices low to avoid the sick. If risk adjustment or subsidies offset this effect, insurers raise price markups. In insurance exchanges, higher markups may be a greater public policy concern than in typical markets, since government subsidies are linked to prices. Higher markups raise government costs, which create a direct efficiency cost because of the excess cost of tax-financed public funds.

An important limitation of this analysis is that I have throughout held fixed the prices of Partners hospitals. This may be sensible for the relatively small CommCare exchange (covering about 3% of Massachusetts' population), and indeed, I found that Partners did not change its prices much after Network Health dropped it in 2012. However, if plans in a broader array of markets dropped Partners, Partners would be forced to respond. Analyzing this response would require modeling hospital-insurer bargaining, something I have not yet done because of its complexity. However, part of the logic in such a model seems clear. Adverse selection that discourages plans from covering Partners should pressure Partners to lower its prices – while policies that offset selection should reduce this pressure. These effects are qualitatively similar to the effects on insurer prices discussed above.

With Partners, however, the welfare effects of higher prices are more complicated. Higher prices at star academic hospitals partly fund activities like teaching and medical research. Whether the government should subsidize plans to cover Partners depends on the social value of these activities. The above analysis has valued these at cost, but the true social value may be higher or lower. How to assess high star hospital prices is beyond the scope of this paper but an important topic for future research.

1.9 Conclusion

As health insurance programs like the ACA increasingly use exchanges to provide coverage, an important question is how well insurance competition will work. A key part of that question is whether adverse

selection is still a concern, despite exchange regulations and risk adjustment used to combat it. This paper has shown evidence from the Massachusetts exchange that there is meaningful residual selection against plans covering expensive star hospitals. Studying a 2012 case where a large plan dropped the star Partners hospitals, I find that selection explains between 35-50% of the plan's cost reductions. The selection is driven by people who strongly prefer the star hospitals and are willing to switch plans to maintain access to them. I find that this group has high risk-adjusted costs both because of greater unobserved risk and because conditional on medical risk, they are more likely to use the high-price hospitals. Improved risk adjustment can mitigate the selection on unobserved risk, but existing risk adjustment methods are not designed to address selection on use of high-price providers.

In many ways, the implications of this adverse selection are standard. Plans have disincentive to cover star hospitals. And when they do, their costs (and therefore prices) are increased in a way that sub-optimally allocates consumers across plans. For example, some people who would like to use star providers only for a severe disease like cancer must pay higher premiums that reflect the costs of people who use high-price providers for all their health care. This inability of a single premium to efficiently sort people with heterogeneous costs across plans is related to a point made in a different context by Bundorf, Levin, and Mahoney (2011). I show that this problem is also related to adverse selection, which gives plans an incentive to exclude star hospitals from network.

This inefficiency is fundamentally related to a sorting challenge: which patients should get access to the expensive services star academic hospitals provide? In standard markets, prices at the point of use create the sorting mechanism – only those willing and able to pay get access. In health insurance, plans cover all or most of hospitals' prices. Instead, people choose their hospital access when they choose plans. This system can lead to a type of moral hazard – when a plan covers star hospitals, its enrollees switch to using these high-price facilities rather than lower-price alternatives. Policies that reduce this moral hazard may also mitigate the adverse selection I find (see Einav et al. 2013). Examples include tiered patient copays (higher fees for more expensive providers) and supply-side incentives for doctors to steer patients

to lower-price facilities (e.g., partial capitation; see Song, et al. 2011; Ho and Pakes 2014). How best to sort patients across hospitals of varying costs is an important question for future research.

A key driver of the selection I find is the high prices of star hospitals. Researchers are increasingly recognizing the importance of provider prices in driving both cost increases and variations across providers (HCCI 2014; Newhouse et al. 2013). This study contributes an additional finding: providers with high prices create adverse selection against plans covering them.

This selection has implications for the health insurance exchanges in the ACA. It calls into question the efficiency of the sharp rise in limited network plans in the ACA's first year (McKinsey 2014). Narrow network plans (covering less than 70% of area hospitals) represented almost half of exchange plans and about 70% of the lowest-price plans. These plans, which are particularly likely to exclude academic hospitals, may grow because of favorable selection at the expense of broad network plans. This pressure on insurers may lead star providers to respond by cutting their prices and costs. It may also add to incentives for these providers to merge with or create an insurer – as Partners did with NHP in Massachusetts and as hospitals elsewhere have done or are considering (Frakt 2014).

The policy implications of my adverse selection findings, however, are less clear. On the one hand, selection against plans covering star hospitals suggests a benefit to subsidizing these plans, through modifications to risk adjustment or subsidies. However, as I showed in simulations, these policies reduce incentives for both insurers and the star providers to lower prices, worsening pre-existing market power. A key question for assessing this tradeoff is what high prices at star academic hospitals fund. If high prices fund valuable teaching, medical research, and uncompensated care for the poor, then pressure to reduce prices may be a public policy concern. If high prices fund higher physician salaries and fancier medical facilities, the policy calculus of subsidizing them would be different. Optimal policy also depends on whether there are more efficient means of subsidizing these activities than through the insurance system. These issues are important questions for future research.

Chapter 2

Price-Linked Subsidies and Health Insurance Markups[1]

2.1 Introduction

An increasingly important model for public health insurance programs is the coverage of enrollees through organized marketplaces offering a choice among subsidized private plans. Long used in Medicare's private plan option (Medicare Advantage), this model was adopted for the Medicare drug program (Part D) in 2006 and most recently, by the Affordable Care Act (ACA) exchanges in 2014. The goals of this program design are to leverage the benefits of choice and competition, while ensuring affordability through subsidies. In this paper, we argue that the method for setting subsidies can affect the strength of insurer price competition, leading to an important interaction between these two goals.

There are two basic approaches to setting subsidies. First, subsidies may be set "exogenously" – based on factors not controlled by market actors, such as an actuarial estimate of expected cost. While exogenous subsidies create clear-cut incentives, they risk leaving consumers with higher than expected premiums (when prices are higher than expected) or giving them windfalls at government expense (when prices are lower than expected). To remedy this problem, recent reforms (including Medicare Part D and the ACA) follow a second approach: setting subsidies endogenously as a function of insurers' prices. These "price-linked" subsides allow the state to ensure the affordability of insurance in the face of cost uncertainty. For instance, the ACA sets a consumer-specific subsidy so that consumers' post-subsidy premium for the second-cheapest "silver-tier" plan equals a specified "affordable" share of their income. This ensures that at least two silver plans will be affordable, even if prices grow faster than anticipated.

[1] This chapter is co-authored with Sonia Jaffe.

We point out an overlooked disadvantage of price-linked subsidies: they risk distorting firms' pricing incentives in imperfectly competitive markets. The basic intuition for the distortion is simple: if higher prices yield higher subsidies, firms have an incentive to raise prices. However, this intuition is only correct if higher prices increase the relative subsidies for a firm's plans. Since in the ACA there is a single flat subsidy,[2] if it applied to all options in the market, then (under standard assumptions) there would be no pricing distortion. However, though the subsidy applies to all plans, it does not apply to the "outside option" of not purchasing insurance. When the subsidy goes up, it decreases the cost of buying a plan relative to not buying insurance. Each firm will gain some of the consumers brought into the market by this price decrease; therefore, each firm has an incentive to raise the price of any plan it thinks might affect the subsidy. This has the potential to increase government subsidy costs, distort consumer choices, and raise prices for higher-income consumers who do not receive subsidies.[3]

We use a simple choice model based on the rules in the ACA exchanges to show this price distortion theoretically and derive sufficient statistics for its magnitude. We focus on the ACA case because of its timeliness and policy relevance during the early years of implementation. (We return to the broader implications for other markets in Section 2.4.) We show that the pricing incentive distortion depends on the price-responsiveness of consumers' decisions to purchase insurance in the exchange. Specifically, the price distortion for the subsidy-pivotal plan depends on its demand semi-elasticity with respect to the price of the outside option -- the fractional increase demand when the non-purchase price

[2] In other settings, subsidies vary across plans, and a plan's subsidy is directly increasing in its own price. These settings include Medicare Advantage -- which decrease subsidies by between 30-50 cents for each $1 of price decrease below a county benchmark -- and many employers that subsidize a fixed percent (e.g., 85%) of each plan's premium. In these settings with "marginal" subsidies, the intuition for a price distortion is even clearer than in the case with flat subsidies, which we focus on. Cutler and Reber (1998) show that marginal subsidies can be advantageous to offset mispricing due to adverse selection. In theory, the distortion we study could also mitigate adverse selection between insurance and uninsurance, but increasing the mandate penalty would achieve the same effect without distorting the relative prices of the plans most likely to be subsidy pivotal.

[3] If silver plan prices rise enough, the implied subsidy may exceed the price of some bronze plans, creating a dilemma of whether to allow negative consumer premiums or to penalize the cheapest bronze plans by capping their subsidies. The ACA does not allow negative premiums, and the phenomenon of subsidies exceeding the prices of some plans appears to have happened widely in the bidding for 2014 plans (the first year of exchanges). McKinsey Center for U.S. Health System Reform (2013) finds that 6-7 million uninsured Americans will have access to a plan whose post-subsidy premium is $0. This is a substantial fraction of the projected steady-state enrollment in exchanges of about 25 million (CBO 2013).

rises by $1. In the case of the ACA, the main outside option is uninsurance, whose price is the mandate penalty.[4] To estimate the magnitude of the distortion, we need to estimate how much changes in the mandate penalty affect demand for insurance.[5]

To do so, we study the pre-ACA subsidized insurance exchange in Massachusetts, called Commonwealth Care (or CommCare). A key model for the ACA exchanges, CommCare offered subsidized non-group insurance for eligible individuals earning less than 300% of poverty, a similar group as the ACA's subsidized population of 100-400% of poverty.[6] We use administrative data on enrollment and consumer demographics for all CommCare enrollees from the start of the program in November 2006 until June 2011.

To estimate our key statistic, we use variation from two natural experiments that occurred in 2007-2008. The first experiment is the introduction of the mandate penalty in December 2007 for individuals earning above 150% of poverty. This group shows a clear spike in new enrollments in December and subsequent months, with no concurrent changes for a control group of enrollees below poverty not subject to financial penalties. Using difference-in-differences regression, we find that the number of new enrollees in the cheapest plan exceeded trend by about 23% of its steady-state size. Scaling this by the mandate penalty amounts (which varied by income), these results imply that on average each $1 increase in the monthly penalty increased enrollment by 0.97%.

[4] Other sources of insurance are unlikely to be an option for the ACA's subsidy-eligible population. Anyone eligible for "affordable" employer-sponsored insurance (with a premium less than 9.5% of income) is not eligible for exchange subsidies, and a similar provision applied in Massachusetts during our study period. Non-group insurance purchased outside of exchanges is likely to be a dominated choice relative to the heavily subsidized exchange plans.

[5] While this is conceptually related to past work estimating the response of insurance demand to price in settings with a fixed price of outside options (e.g. Gruber and Poterba (1994) on non-group insurance for the self-employed and Gruber and Washington (2005) on employer-sponsored insurance), there are no similar estimates of the response of coverage to the mandate penalty in a low-income exchange setting. The closest related work is that of Chandra et al. (2011), who, like us, study the introduction of the mandate penalty in Massachusetts. However, their focus is on the effect of the mandate penalty on adverse selection, so they do not report estimates of the increase in total coverage.

[6] Prior to 2014, CommCare was separate from Massachusetts' unsubsidized exchange for individuals above 300% poverty, which has been studied by Ericson and Starc (2012). Under the ACA, the two Massachusetts exchange populations merged together, and most individuals below 100% poverty are shifting to Medicaid (with the exception of some legal immigrants remaining in the exchange). Individuals earning less than 400% of poverty receive subsidies, while those above 400% poverty can purchase the same plans without subsidies.

The second experiment uses a policy change in July 2007 that decreased all plans' premiums for consumers 100-150% poverty. Because a lower price of all inside options has an equal effect on relative prices as a higher price of the outside option, this experiment can be used to estimate the response of insurance demand to the relative mandate penalty. We again find a visible spike in new enrollment for the 100-150% poverty group in absolute terms and relative to a control group of higher income enrollees whose prices were essentially unchanged. Our preferred estimates imply that new enrollment in the cheapest plan grew by 17% of steady-state size. Given the price changes, these estimates imply a 0.94% increase in enrollment for each $1 increase in the relative mandate penalty. The similarity of results across these two quite different experiments is striking and reassuring.

To estimate the distortion, our model also requires an estimate of the own-price semi-elasticity of demand. For this, we use the estimates of Chan and Gruber (2010), who studied the same market, time period (2007-2008), and income groups (100-300% poverty). Their coefficient implies that each $1 premium increase reduces demand for the cheapest plan among new enrollees by 1.97%. Plugging their and our estimates into the model, we estimate a upward distortion due to price-linked subsidies of $48 per member per month. This amount is substantial -- about 12% of the average CommCare insurance price at the time. Since this distortion applies to the subsidy-pivotal cheapest plan, it implies an equal distortion of the subsidy amount.

Though the ACA exchanges differ from Massachusetts, we expect the incentive distortion will continue to be important. In particular, our theory predicts the distortion will be larger in more concentrated markets. Data for 2014 suggest that many ACA markets will be highly concentrated, with over half of (county-level) markets having just one or two participating insurers (Abelson et al., 2013). However, the distortion may be mitigated by the presence of unsubsidized consumers (projected to be about 20% of the market).

The policy implications of our results for the ACA exchanges depend on balancing several considerations. One option would be to shift to exogenous subsidies, which do not distort prices. But

exogenous subsidies cannot guarantee the availability of "affordable" post-subsidy plan premiums if prices grow without an accompanying increase in subsidies. Adjusting subsidies based on exogenous factors likely to correlate with local plan prices – e.g., local cost growth in fee-for-service Medicare[7] – could mitigate but would not eliminate the problem. Furthermore, the experience with persistently high payments through exogenous subsidies in Medicare Advantage (see MedPAC, 2013) indicate potential political problems with exogenous subsidies.

Our model suggests a simple alternative to remove the distortion while preserving the affordability advantages of price-linked subsidies: apply the subsidy to the outside option as well. While normally, the cost of not purchasing a good is fixed at zero, the ACA's mandate penalty for uninsurance makes such a subsidy for the outside option possible. Specifically, if the second-cheapest silver price exceeds an expected target level, the difference would be applied to reduce the mandate penalty (and vice versa if its price is below the target). Under this system, subsidies would still ensure the "affordability" of at least two silver plans. But a higher subsidy would not affect the relative prices of insurance vs. non-insurance, removing the distortionary incentive.

Another way to see this is to consider the net subsidy for insurance vs. non-insurance, which equals the subsidy plus the mandate penalty, $S + M$. With exogenous subsidies, $S + M$ is exogenous to plan prices, so there is no distortion. Under the ACA's price-linked subsidies, higher prices increase *S* but do not affect *M*, implying a positive effect of prices on $S + M$. Under our alternate proposal, higher prices increase *S* and decrease *M*, leaving $S + M$ unchanged. As with exogenous subsidies, there is no distortion because plan prices do not affect the net subsidy for insurance. However, a potential downside is that unexpectedly high prices could (by reducing the mandate penalty) increase uninsurance, an outcome the current policy forestalls. We discuss this issue further in Section 2.4.

The distortion we identify is relevant to the choice between exogenous and price-linked subsidies in a variety of settings, including the ACA exchanges, Medicare Advantage and Part D, and employer-

[7] While local premiums in employer-sponsored insurance seem appealing as an adjustment factor, they would not be truly exogenous if the same insurers offered both exchange and employer plans.

sponsored plan choices like the Federal Employees Health Benefits Program. Most of these programs set subsidies using plan prices, either based on a pivotal plan (as in the ACA) or based on a weighted average of plan prices (as in Medicare Part D for non-low-income enrollees). Although estimating the distortion in these other markets is beyond the scope of this paper, our theory suggests that it will be relevant wherever firms have market power and there is meaningful substitution between the in-market plans and the outside option. We discuss the implications for other markets in Section 2.4.

We are not aware of past research that has analyzed the distortion we discuss. The closest related work is Decarolis (2013), which highlights a different pricing distortion in Medicare Part D. By increasing its plan prices for higher-income enrollees, insurers can increase their payments for low-income subsidy recipients, who do not pay prices. However the structure in Part D is different, creating different and often more subtle incentives to game the systems. Also, the subsidized consumer share of the Part D market (about 40%) is substantially smaller than their share in the ACA exchanges (about 80%).

The paper proceeds as follows. Section 2.2 sets up a standard choice model and derives the formula for the change in markup due to the endogeneity of the subsidy, first for single-plan insurers (as in Massachusetts) and then for multi-plan insurers (as in the ACA). Section 2.3 uses two natural experiments from Massachusetts to calibrate the relevant semi-elasticity of demand with respect to the mandate penalty. Section 2.4 combines this estimate with the estimate from Chan and Gruber (2010) of the sensitivity of demand to own price to get a quantitative estimate for the distortion in the markup due to endogenous subsidies. Section 2.5 proposes an alternative subsidy framework that would eliminate the distortion while maintaining price-linked subsidies and compares it to the existing systems of exogenous and endogenous subsidies. It also discusses the broader relevance of our findings to markets besides the ACA. Section 2.6 concludes.

2.2 Theory

We analyze the competitive effects of linking subsidies to plan prices in a simple model of consumer choice and insurer bidding. Insurers offer differentiated products and compete by setting prices. The exchange collects price bids and uses them to determine subsidies based on a formula that insurers know before bidding. Subsidy-eligible consumers then choose which (if any) plan to purchase based on post-subsidy prices and plan attributes. If they choose not to purchase a plan, they are subject to the legally applicable mandate penalty. We assume that insurers set prices simultaneously to maximize static profits, knowing the effects of these choices on demand and cost.[8] As in most recent work on insurance (e.g. Einav et al. (2013)), we assume that plan attributes are held fixed, focusing instead on the effect of subsidies on pricing conditional on plan attributes.[9]

We use the necessary first-order conditions for Nash equilibrium to derive sufficient statistics that capture the effect of subsidy rules on insurers' optimal prices. We do so in two institutional settings: a simpler one with single-plan insurers (based on the situation in the Massachusetts exchange before 2014; for additional background, see Section 2.3) and a more complicated setting with multi-plan insurers (based on the ACA exchanges). The first illustrates the intuition more clearly and generates the formula for our analysis of Massachusetts data, but the second is more relevant for future policy. Our analysis suggests reasons that the distortion may be larger in the ACA case.

2.2.1 Theory with Single-Plan Insurers (Massachusetts Case)

We start with a model where each insurer offers a single plan. There are J insurers, each of which chooses a "price bid" P_j, which is submitted to the exchange regulator and is the total amount the insurer receives per enrollee. Using a pre-determined rule that may incorporate the vector of prices P, the regulator sets a

[8] We therefore abstract from pricing dynamics or incomplete information. As noted below, uncertainty would spread the distortion out across multiple plans, but would not eliminate it.

[9] This assumption is less problematic in the insurance industry, where many attributes are severely constrained by regulation. For instance, in the Massachusetts exchange we study empirically, the regulator specifies the services insurers must cover and all of the associated co-payments.

subsidy $S(P)$. This subsidy applies equally to all plans, so consumers face subsidized premiums denoted $P_j^{cons} = P_j - S(P)$. Consumers can also choose the outside option of not buying a plan, in which case they face a mandate penalty, *M*. Consumer demand for plan *j*, $Q_j\left(P^{cons}, M\right)$, is a function of all premiums and the mandate penalty. As in most discrete choice models, we assume that utility is (locally) quasi-linear in price, so consumers only care about prices relative to other prices and to mandate penalty. We assume constant marginal cost and we abstract from adverse selection by assuming ideal risk adjustment; we can therefore specify a constant per-enrollee marginal cost c_j for insurer *j*.[10]

Given this setup, the insurer profit function is:

$$\pi_j = \left(P_j - c_j\right) \cdot Q_j\left(P^{cons}, M\right).$$

In simultaneous-pricing Nash equilibrium, all insurers set prices to maximize profits, given their opponents' strategies. Thus, the necessary conditions for Nash equilibrium are that each firm price according to its first-order conditions for profit maximization:[11]

$$\frac{d\pi_j}{dP_j} = Q_j\left(P^{cons}, M\right) + \left(P_j - c_j\right) \cdot \frac{dQ_j}{dP_j} = 0. \tag{2.1}$$

This differs from standard oligopoly pricing conditions in two respects. Because of the subsidies, the firm's price P_j enters consumer demand indirectly, through the subsidized premiums, P^{cons}. Also, the term dQ_j / dP_j is not the slope of the demand curve, but a composite term that combines the slope of

[10] Risk adjustment works by adjusting the payment insurers receive to be higher than P_j for sick, high-cost enrollees and lower for healthy enrollees. In ideal risk adjustment, the quantity $\left(P_{ij} - c_{ij}\right)$ is constant across enrollees, allowing us to write it in the standard form $\left(P_j - c_j\right)$. Both the Massachusetts and ACA exchanges include sophisticated (though likely still imperfect) risk adjustment. Nonetheless, the effects we identify all carry over to a model with adverse selection, with more complicated formulas for the sufficient statistics.

[11] These first-order conditions would be necessary conditions for Nash equilibrium even in a more complicated model in which insurers simultaneously chose a set of non-price characteristics like copays and provider network. Thus, the theoretical point we make about price-linked subsidies holds when quality is endogenous, though there may also be effects on quality and cost levels, which we don't capture.

demand and the effect on demand via the effect of the price on the subsidy and mandate penalty. The total effect of the premium on demand is:

$$\frac{dQ_j}{dP_j} = \frac{\partial Q_j}{\partial P_j^{cons}} - \left(\sum_k \frac{\partial Q_j}{\partial P_k^{cons}} \right) \frac{\partial S}{\partial P_j} + \frac{\partial Q_j}{\partial M} \frac{\partial M}{\partial P_j}, \tag{2.2}$$

which depends on the specific policy for determining subsidies and the mandate penalty.

Exogenous Subsidies

Exchanges could set flat, exogenous (i.e., based on factors not controlled by market actors) subsidies and mandate penalties.

$$\begin{aligned} &\text{Exogenous Subsidy: } P_j^{cons} = P_j - S, \\ &\text{with } \frac{\partial S}{\partial P_j} = \frac{\partial M}{\partial P_j} = 0 \qquad \forall j. \end{aligned} \tag{2.3}$$

As a result of subsidies and the mandate penalty being unaffected by any plan's price, dQ_j / dP_j in (2.2) simplifies to the demand slope $\partial Q_j / \partial P_j^{cons}$. Even though there are subsidies, the equilibrium pricing conditions are not altered relative to the standard form for differentiated product Bertrand competition. Under this exogenous subsidies benchmark, firms set markups as:

$$Mkup_j^{Exog} \equiv P_j - c_j = \frac{1}{\eta_j} \qquad \forall j,$$

where $\eta_j \equiv -\frac{1}{Q_j} \frac{\partial Q_j}{\partial P_j^{cons}}$ is the own-price semi-elasticity of demand.

Price-Linked Subsidies

Alternatively, exchanges could link subsidies to prices (but again set *M* exogenously). This is the approach taken in Massachusetts and in ACA exchanges. A key advantage of this approach is that a state can ensure affordability to consumers even if prices are higher than expected and avoid windfalls to consumers when prices are lower than expected. This transfer of health care cost risk from individuals to the state is a key motivation for price-linked subsidies.

However, linking subsidies to prices distorts insurers' pricing incentives. To see this distortion, consider a subsidy rule (similar to that used in Massachusetts) that sets $S(P)$ so that the post-subsidy premium for the cheapest plan equals a pre-determined "affordable amount" (based on income). With this policy, the subsidy rises with the price of the cheapest plan, but the mandate penalty is still exogenous to plan prices. Formally, the subsidy for a given income group would be set so that:

$$\text{Subsidy Linked to Min Price: } S(P) = \min_j P_j - \text{AffAmt}$$
$$\Rightarrow \frac{\partial S(P)}{\partial P_{jmin}} = 1, \qquad \frac{\partial M}{\partial P_j} = 0 \ \forall j, \tag{2.4}$$

where *jmin* is the index of the cheapest plan. For example, in 2007, the affordable amount for a consumer earning between 150-200% of poverty was $40 per month, and the cheapest plan in the "Outside of Boston" region bid a price of $295.84 per month. As a result, the subsidy for this income group was set at $255.84.[12] Equation (2.4) shows that the subsidy rises one-for-one with this cheapest plan's price. Had that plan bid $1 higher, the subsidy would increased by $1 so that the consumer premium was still $40.

This price-subsidy link changes optimal pricing for an insurer that believes it will have the cheapest plan. Raising this plan's price by $1 does not affect its own consumer premium (which is offset by the higher subsidy) but lowers the premium of all other plans by $1. Initially, this may sound the same as the standard case -- raising own price by $1 affects demand identically to lowering all other options' prices by $1. But critically, the price increase for the cheapest plan *does not lower the price of being uninsured*, i.e. the mandate penalty. A price increase for the cheapest plan does not increase its price relative to being uninsured. This creates an extra incentive for the cheapest plan to markup its price, which in turn increases subsidy costs to the state.

[12] In later years in the Massachusetts exchange, there were "incremental" subsidies to pricing higher, so more expensive plans received larger subsidies. We do not model this feature of the subsidy scheme, which will not be replicated in the ACA and was only present for enrollees earning less than 150% poverty during the 2007-08 period we study.

To see the distortion analytically, consider how the endogenous subsidy affects the total derivative of demand with respect to price for the cheapest plan:

$$\frac{dQ_{jmin}}{dP_{jmin}} = \frac{\partial Q_{jmin}}{\partial P_{jmin}^{cons}} - \left(\sum_k \frac{\partial Q_{jmin}}{\partial P_k^{cons}} \right) \cdot 1 + \frac{\partial Q_{jmin}}{\partial M} \cdot 0 = \sum_{k \neq jmin} \left(-\frac{\partial Q_{jmin}}{\partial P_k^{cons}} \right).$$

Since the affordable amount is fixed, raising P_{jmin} is equivalent to lowering consumer premiums for all other plans. Next, we use the fact that with utility linear in price, raising the prices of all options (including uninsurance) equally does not affect demand, that is $\sum_k \partial Q_j / \partial P_k^{cons} + \partial Q_j / \partial M = 0 \; \forall j$. Combining these two equations implies that

$$\frac{dQ_{jmin}}{dP_{jmin}} = \frac{\partial Q_{jmin}}{\partial P_{jmin}^{cons}} + \frac{\partial Q_{jmin}}{\partial M}. \tag{2.5}$$

However, if the firm raises its price so much that it is no longer the cheapest plan, then it is no longer pivotal for the subsidy, and $\frac{dQ_{jmin}}{dP_{jmin}} = \frac{\partial Q_{jmin}}{\partial P_{jmin}^{cons}}$.

Plugging Equation (2.5) into Equation (2.1) and rearranging yields the following markup condition for the cheapest plan when subsidies are endogenous

$$Mkup_{jmin}^{Endog} \equiv P_{jmin} - c_{jmin} = \frac{1}{\eta_{jmin} - \eta_{jmin,M}}$$

where $\eta_{jmin,M} \equiv \frac{1}{Q_{jmin}} \frac{\partial Q_{jmin}}{\partial M}$ is the semi-elasticity of demand with respect to the mandate penalty. Price-linked subsidies lower the effective price sensitivity faced by the cheapest plan, leading to a higher equilibrium markup than under exogenous subsidies. In our model without uncertainty, this distortion only applies to the cheapest plan, though there may be strategic responses by other firms.[13] If the distortion is large enough that the cheapest plan would want to price above the second cheapest plan, it instead sets a price equal to the second cheapest plan.

[13] Like much of the related literature, we make the simplifying assumption that firms know what equilibrium they are in, so there is no uncertainty about which plan will be cheapest. In a more realistic model with uncertainty about others' prices (perhaps due to uncertainty about others' costs) then the distortionary term would be weighted by the probability of being the lowest price plan. The (ex-post) cheapest plan would have a smaller distortion, but there would also be distortions to other plans' prices. Strategic complementarity in prices could further exacerbate these.

If the semi-elasticities of demand are constant across the relevant range of prices (own-cost pass-through equals 1), the increase in markup for the cheapest plan is:

$$Mkup_{jmin}^{Endog} - Mkup_{jmin}^{Exog} = \frac{\eta_{jmin,M}}{\eta_{jmin} \cdot \left(\eta_{jmin} - \eta_{jmin,M}\right)} \tag{2.6}$$

though the distortion cannot cause the cheapest plan to raise its price above that of the second cheapest plan. Constant semi-elasticities may be a good approximation in some markets; alternatively, the estimated increase in markups can be thought of as an estimate of how much marginal costs would have had to decrease to offset the incentive distortion generated by endogenous subsidies.[14]

The distortion will tend to be larger when markets are less competitive. When there are fewer firms, on average we expect η_{jmin} to be smaller because consumers have fewer other options and $\eta_{jmin,M}$ to be higher because a given firm will get a larger share of however many consumers enter the market when the penalty goes up. Both of these effects increase the distortion. Also, with fewer firms, it is less likely that the second cheapest plan will be close in price to the cheapest plan and thereby act as an effective cap on the distortion.

This pricing distortion can have an important effect on social welfare. While the price distortion is on the cheapest plan, it also drives up the subsidy, which raises the government's costs for all enrolled individuals, not just those that chose the cheapest plan. With a marginal cost of government funds above one, this transfer from the state to insurers is socially costly.[15]

[14] This is similar to the idea from Werden (1996) that, without assumptions about elasticities away from the equilibrium, one can calculate the marginal cost efficiencies needed to offset the price-increase incentives of a merger.

[15] In addition, the price distortion affects relative premiums and changes the allocation of consumers across plans. But because relative prices may not equal relative marginal costs with imperfect competition (they differ by the difference in markups), it is not clear whether this has a positive or negative effect.

2.2.2 Theory with Multi-Plan Insurers (ACA Case)

The distortion described above is exacerbated when firms offer multiple plans in a market. In the ACA insurers must offer a plan in each of multiple tiers -- bronze, silver, gold, and platinum.[16] Subsidies are set equal to the second-cheapest silver plan minus a pre-specified "affordable" share of a consumer's income.[17] The fact that insurers are providing additional plans provides an even greater incentive for an insurer to increase the price of its silver plan because the higher subsidy increases demand for the insurer's non-silver plans as well -- again by inducing more customers to enter the market.

Suppose each firm j offers plans in tiers $l = 1,...,L$. The insurer maximizes profits:

$$\max_{P_{j1}...P_{jL}} \sum_{l} \left(P_{jl} - c_{jl}\right) Q_{jl}\left(P^{cons}, M\right),$$

where $P_{jl}^{cons} = P_{jl} - S(P)$ with $S(P) = P_{2nd,Silv} - \text{AffAmt}$. Following the same steps as above, the first-order condition for the silver plan is:

$$\frac{\partial \pi_j}{\partial P_{jsilv}} = Q_{jsilv}(\cdot) + \sum_{l} \left(P_{jl} - c_{jl}\right) \frac{dQ_{jl}}{dP_{jsilv}} = 0.$$

The markup with exogenous subsidies is:

$$Mkup_{jsilv}^{\text{Exog}} = \frac{1}{\eta_{jsilv}} + \frac{1}{-\frac{\partial Q_{jsilv}}{\partial P_{jsilv}^{cons}}} \sum_{l \neq silv} \left(P_{jl} - c_{jl}\right) \frac{\partial Q_{jl}}{\partial P_{jsilv}^{cons}}$$

The second term reflects the fact that the insurer captures revenue from consumers who switch to its other plans when it raises the price of its silver plan. This is a standard effect in settings with multi-product firms. However, with the subsidy set based on the second cheapest silver plan, the markup for that subsidy-pivotal plan is:

[16] Platinum plans cover 90% of medical costs (comparable to a generous employer plan today); gold covers 80% of costs; silver covers 70% of costs; and bronze covers 60% of costs. Consumers with incomes below 250% of poverty also receive so called "cost-sharing subsidies" that raise the generosity of silver plans.

[17] This subsidy is applied equally to all plans (with a cap ensuring that no premium is pushed below $0), ensuring that at least two silver plans (and likely some bronze plans) cost less than the affordable amount for low-income consumers.

$$Mkup_{jsilv}^{\text{Endog}} = \frac{1}{\eta_{jsilv} - \eta_{jsilv,M}} + \frac{\sum_{l \neq Silv} \left(P_{jl} - c_{jl}\right) \left(\frac{\partial Q_{jl}}{\partial P_{jsilv}^{cons}} + \underbrace{\frac{\partial Q_{jl}}{\partial M}}_{\text{AdditionalDistortion}} \right)}{-\left(\frac{\partial Q_{jsilv}}{\partial P_{jsilv}^{cons}} + \underbrace{\frac{\partial Q_{jsilv}}{\partial M}}_{\text{AdditionalDistortion}} \right)}. \quad (2.7)$$

The fact that other plans offered by the firms also gain some of the consumers driven into the market by the additional subsidy generates an additional distortion.

How much larger the distortion is in the multi-product ACA case is not certain, and we do not have data to credibly estimate its size. We discuss some of the issues in translating our estimates for Massachusetts to the ACA case in Section 2.4.

2.3 Data and Estimation

While the theory of the distortion from price-linked subsidies is clear, the question remains whether the distortion is large enough to be of practical importance. To estimate the size of the distortion, we turn to data from the Massachusetts subsidized health insurance exchange (called Commonwealth Care, or "CommCare"). Created in the Massachusetts's 2006 health reform, CommCare facilitates and subsidizes private insurance coverage for individuals earning less than 300% of poverty and without access to insurance through an employer or another government program. The market is quite concentrated, with just four insurers offering plans during the period we study, making it an appropriate setting to study imperfect competition. Importantly, there have been several changes in plan premiums and the mandate penalty over time, allowing us to identify the relevant demand elasticities.

We use administrative data on plan choices and consumer demographics for all enrollees in the CommCare program from 2006 through June 2011.[18] For each month, we observe the set of participating members, their demographics, the plan they chose, and their available plans and premiums. Our main

[18] This data was obtained under a data use agreement with the Massachusetts Health Connector, the agency that runs CommCare.

analysis focuses on trends in the number of consumers entering the market ("new enrollees"). Every month, some individuals become newly eligible for CommCare, for instance due to job loss or other income changes. These individuals then choose whether to enroll and which plan to sign up for if they enroll. We study whether changes in the relative cost of not enrolling (the mandate penalty) affect monthly new enrollments in the cheapest plan. (This is the effect needed for the distortion equation; the semi-elasticity of new enrollments in all plans is similar.)

In theory, a higher mandate penalty could also reduce the number of consumers exiting the market, meaning our estimates would understate the true elasticities. However, studying exits was complicated by an income verification program introduced around the same time as our other reforms (late 2007 and early 2008). This forced many enrollees to leave the market, leading to substantial noise and potentially bias in an analysis of exits. Thus, we chose to focus only on new enrollments.

Table 2.1 shows summary statistics for our sample. The data include 490,368 unique consumers. The population is quite poor, with over half having family income less than the poverty line. There were an average of 11,365 new enrollments per month, giving us a substantial sample to study changes over time in this statistic. Since consumers below poverty do not pay premiums (all plans cost $0) or a mandate penalty, they act as a control group for our analysis of the population between 100-300% of poverty.

Recall that in the case of single-plan insurers, our formula for the increase in markups is:

$$Mkup_{jmin}^{Endog} - Mkup_{jmin}^{Exog} = \frac{\eta_{jmin,M}}{\eta_{jmin} \cdot \left(\eta_{jmin} - \eta_{jmin,M}\right)}$$

An estimate of the own price elasticity, η_{jmin}, is available in the literature (see Section 2.4), so we focus on estimating $\eta_{jmin,M}$, the effect on demand for the cheapest plan when the mandate penalty is raised. In what follows, we use two natural experiments in CommCare to estimate this key statistic.

Table 2.1 Summary Statistics

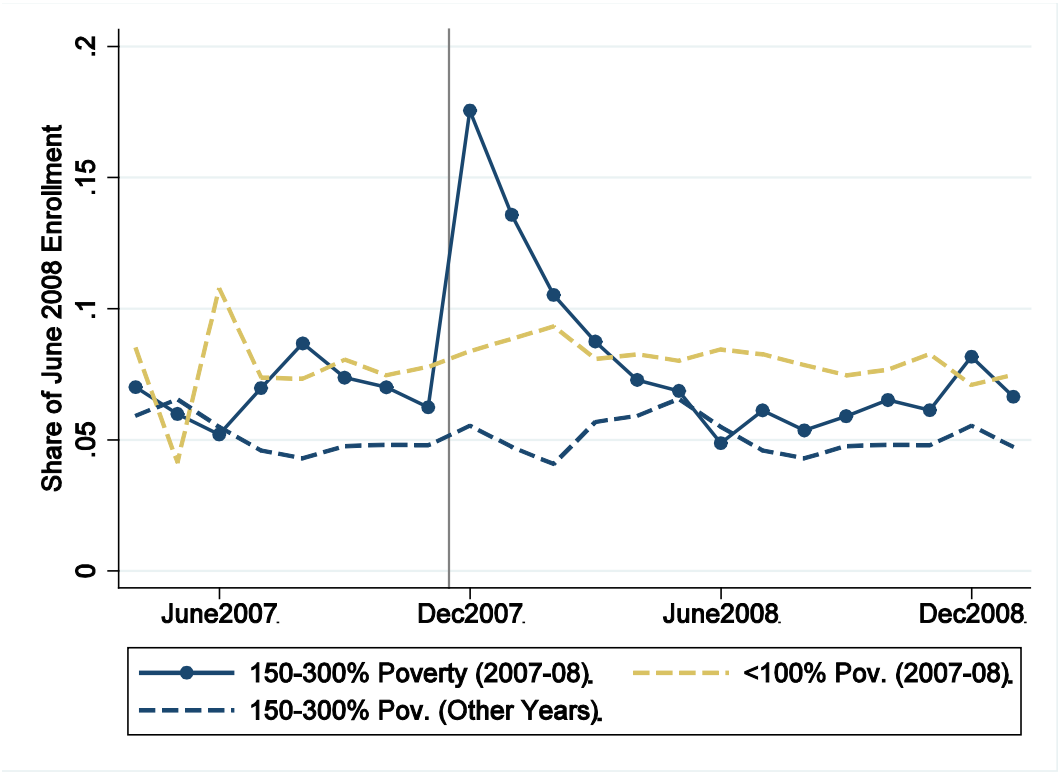
Variable	Full Sample	Income Group (% of Federal Poverty Line)				
		<100%	100-150%	150-200%	200-250%	250-300%
Enrollment						
Total Unique Enrollees	490,368	267,642	95,567	72,657	35,966	18,536
Monthly Enrollment	141,803	72,728	31,173	23,250	12,225	6,338
	[43,968]	[16,922]	[10,596]	[7,741]	[4,457]	[2,375]
Monthly New Enrollees	11,365	6,031	2,346	1,863	942	485
	[3,525]	[2,587]	[1,773]	[635]	[324]	[162]
Enrollment Spell Length (months)	14.5	14.1	15.2	14.6	14.6	14.8
	[12.0]	[11.6]	[12.3]	[12.7]	[12.6]	[12.5]
Share of Enrollment Months	100.0%	51.3%	20.8%	15.5%	8.2%	4.2%
Monthly Prices						
Insurer Price Bid	$390.22	$385.54	$384.18	$392.59	$415.22	$419.71
	[$85.31]	[$83.23]	[$82.13]	[$76.95]	[$101.90]	[$102.89]
Consumer Premium	$22.28	$0.00	$4.90	$48.72	$96.30	$137.83
	[$39.32]	[$0.00]	[$7.80]	[$14.97]	[$21.58]	[$23.19]
Mandate Penalty (starting Jan 2008)	$8.66	$0.00	$0.00	$18.01	$36.37	$54.78
	[$15.25]	[$0.00]	[$0.00]	[$0.88]	[$1.49]	[$2.99]
Demographics						
Age	39.9	37.0	41.4	43.7	44.7	45.7
	[14.2]	[14.2]	[14.1]	[13.1]	[12.9]	[12.7]
Male	47.2%	52.2%	42.0%	40.8%	42.9%	44.4%

NOTE: This table shows means and standard deviations (in brackets) of statistics about CommCare enrollees from November 2006 to June 2011. "Insurer Price Bid" shows the enrollment-weighted average price paid to firms for an enrollee. Premiums are the average (enrollment-weighted) post-subsidy monthly prices consumers pay for plans. The mandate penalty -- which is set by law as half of each income group's affordable amount -- is its average monthly value from January 2008 (the month the regular penalty started) until June 2011.

2.3.1 Mandate Penalty Introduction Experiment

Our first strategy uses the mandate penalty's introduction. Under the Massachusetts health reform, a requirement to obtain insurance took effect in July 2007. However, through November 2007 the requirement was not enforced by any financial penalties. Financial penalties began in December 2007. Those earning more than 150% of poverty who were uninsured in December forfeited their 2007 personal

Figure 2.1 New Enrollees in Cheapest Plan by Month



NOTE: This figure shows for two income groups the monthly number of new enrollees into CommCare who get the cheapest plan. The vertical line is drawn just before the introduction of the mandate penalty, which applied to people 150-300% poverty but not those below 100% poverty. The "150-300% Poverty (Other Years)" series shows average new enrollments in each calendar month in all years in our data except July 2007--June 2008. Each income group's numbers are normalized by the group's total enrollment (in the same plan) in June 2008, so units can be interpreted as fractional changes in enrollment. "New enrollments" include both individuals enrolling in CommCare for the first time and individuals re-enrolling after a break in coverage, since both groups select a new plan. For the 150-300% poverty group, the cheapest plan is defined as the lowest-premium plan (or plans if there is a tie) in each individual's choice set (which can vary by region and income group). For the below 100% poverty group for whom all plans are free, the cheapest plan is defined as the lowest-premium plan for 150-200% poverty enrollees in the same region.

exemption on state taxes – a penalty of $219 (see Care, 2008). Starting in January 2008, the mandate penalty was assessed based on monthly uninsurance. The monthly penalties depended on income, ranging from $17.50 for a person with household income of 150-200% of poverty to $52.50 for someone between 250-300% of poverty.

There was a spike in new enrollees into CommCare for people above 150% of poverty (the group actually subject to the penalties) exactly concurrent to the introduction of the financial penalties in December 2007 and early 2008. Figure 2.1 shows this enrollment spike for the cheapest plan, which is

proportional to the overall spike; (a fairly constant 60% share of new enrollees chose the cheapest plan.) To make magnitudes comparable for income groups of different size, the figure shows new enrollments as a share of the same plan's total enrollment in that income group in June 2008.[19]

We argue that this enrollment spike was caused by the financial penalties for three reasons. First, there were no changes in plan prices or other obvious demand factors for people above 150% of poverty that occurred at the same time. Second, as Figure 2.1 shows, there was no concurrent spike for people earning less than poverty (who were not subject to the penalties),[20] and there was no enrollment spike for individuals above 150% of poverty in December-March of years other than 2007-08. Finally, Chandra et al. (2011) show evidence that the new enrollees after the penalties were differentially likely to be healthy, consistent with the expected effect of a mandate penalty in reducing adverse selection.

We estimate the semi-elasticity associated with this response using a difference-in-differences specification, analogous to the graph in Figure 2.1. We estimate excess new enrollments in December 2007-March 2008 relative to the trend in nearby months, using enrollment trends for people earning less than poverty as a control group. We estimate the effect through March 2008 for two reasons. First, the application process for the market takes some time, so people who decided to sign up in January may not have enrolled until March. Second, the mandate rules exempted from penalties individuals with three or fewer months of uninsurance during the year, meaning that individuals who enrolled in March avoided any penalties for 2008. However most of the effect is in December and January, so focusing on those months does not substantially affect our estimates.

[19] We use June 2008 as a baseline because enrollment, which had been steadily growing since the start of CommCare, stabilizes around June 2008. Therefore, we treat June 2008 enrollment as an estimate of equilibrium market size.

[20] People earning 100-150% of poverty are omitted from this analysis because a large auto-enrollment took place for this group in December 2007, creating a huge spike in new enrollment. But the spike occurred only in December and was completely gone by January, unlike the pattern for the 150-300% poverty groups. This auto-enrollment did not apply to individuals above 150% of poverty (Commonwealth Care 2008) so it cannot explain the patterns shown in Figure 2.1.

We collapse the data to the income group-month level and calculate the new enrollees in the cheapest plan for each group and month, normalized by the same plan's total enrollment for the income group in June 2008 (as discussed above). The difference-in-differences model we estimate is:

$$NewEnroll_{g,t} = \sum_g \alpha_g \cdot 1_g + \beta \cdot Treat_t + \sum_g \gamma_g \cdot 1_g \cdot Treat_t + \sum_{g,t} \delta_{g,t} \cdot 1_g \cdot X_t + \varepsilon_{g,t}, \quad (2.8)$$

where 1_g is an indicator for income group g, $Treat_t$ is an indicator for t being in December 2007 through March 2008, and X_t is a vector of time polynomials and CommCare-year dummies, the coefficients of which we allow to vary across groups. The difference-in-difference coefficient of interest is γ_g for the groups above 150% of poverty subject to the penalties.

Table 2.2 presents the regression results. Column (1) starts with a baseline single-difference specification that estimates based only on enrollment for the 150-300% poverty group in December 2007-March 2008 relative to the surrounding months. Column (2) then adds the <100% poverty group as a control group, to form the difference-in-difference estimates. Finally, Column (3) adds dummies for December-March in all years, forming the triple difference that nets out general trends for those months in other years. All of these specifications use data from March 2007 to June 2011 and also include CommCare-year dummies and time polynomials separately for the treatment and control groups.[21] Despite the small number of group-month observations, all the relevant coefficients are highly significant. Summing across treatment months, the total excess enrollment after the mandate penalty introduction was between 22-25% of equilibrium market size.

[21] CommCare plan years run from July to June, so the year dummies are stable over the treatment period from December 2007 to March 2008. We use data only up to June 2011 because of significant changes in the prices and availability of the cheapest plans that took effect in July 2011.

Table 2.2 Introduction of the Mandate Penalty

Dependent Var: New Enrollees in Cheapest Plan / June 2008 Enrollment

Variable	(1)	(2)	(3)
Sum of Dec2007 - Mar2008 coefficients (below)	**0.253*** (0.017)**	**0.237*** (0.023)**	**0.225*** (0.024)**
150-300% Poverty x Dec2007	0.112***	0.110***	0.103***
	(0.003)	(0.006)	(0.007)
x Jan2008	0.073***	0.067***	0.069***
	(0.004)	(0.006)	(0.006)
x Feb2008	0.043***	0.033***	0.033***
	(0.005)	(0.006)	(0.006)
x Mar2008	0.025***	0.027***	0.020***
	(0.006)	(0.006)	(0.007)
Control Group (<100% Poverty)		X	X
Triple Difference (dummies for Dec-March)			X
Observations	51	102	102
R-Squared	0.969	0.923	0.925

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

NOTE: This table performs the difference-in-difference regressions analogous to the graphs in Figure 2.1. The dependent variable is the number of new CommCare enrollees who choose the cheapest plan in each month in an income group, scaled by total group enrollment in that plan in June 2008. There is one observation per income group (the 150-300% poverty treatment group, plus the <100% poverty control group in columns (2) and (3)) and month (from April 2007 to June 2011). All specifications include CommCare-year dummy variables and fifth-order time polynomials, separately for the treatment and control group, to control for underlying enrollment trends. (The CommCare-year starts in July, so these dummies will not conflict with the treatment months of December to March.) Specification (3) also includes dummy variables for all calendar months of December-March for the treatment group, to perform the triple-difference. See the note to Figure 2.1 for the definition of new enrollees and the cheapest plan.

Table 2.3 takes the final, triple-difference specification from Table 2.2 and breaks the analysis down into narrower income groups (by 50% of poverty interval, the narrowest we have). The coefficients are a little larger for the higher income groups -- about 25% instead of 21% -- who faced higher mandate penalties. Given the mandate penalties of $17.50-$52.50, the coefficients imply that each $1 increase in the mandate penalty raised demand by between 1.2% for the 150-200% of poverty group to 0.48% for the 250-300% of poverty group, with a weighted average of 0.97%.

Table 2.3 Introduction of the Mandate Penalty, by Income Group

Dependent Var: New Enrollees in Cheapest Plan / June 2008 Enrollment

Variable	Income Group (% of Poverty Line)		
	150-200%	200-250%	250-300%
Sum of Dec2007 - Mar2008 coefficients (below)	**0.208*** (0.026)**	**0.263*** (0.022)**	**0.250*** (0.024)**
150-300% Poverty x Dec2007	0.101***	0.113***	0.091***
	(0.007)	(0.007)	(0.008)
x Jan2008	0.061***	0.085***	0.086***
	(0.007)	(0.007)	(0.007)
x Feb2008	0.026***	0.045***	0.051***
	(0.007)	(0.006)	(0.006)
x Mar2008	0.020**	0.020***	0.022***
	(0.008)	(0.006)	(0.007)
Observations	102	102	102
R-Squared	0.927	0.922	0.920

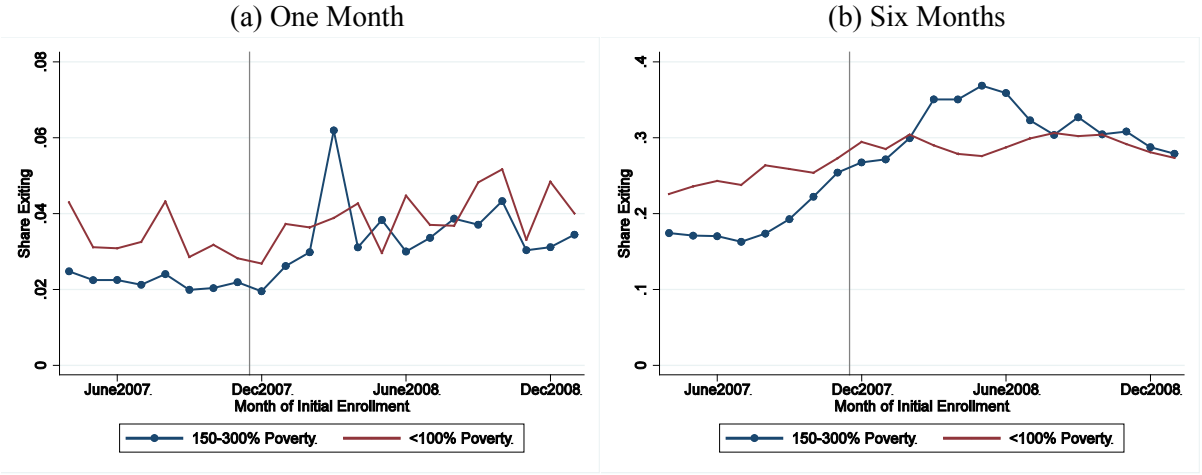
Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

NOTE: This table performs the triple-difference regression specifications in Column (3) of Table 2.2 separately for each 50% of poverty income group (the level at which the mandate penalty varies). The dependent variable is the number of new CommCare enrollees who chose the cheapest plan in each month in an income group (the 50% of poverty group specified in the column headings), scaled by total group enrollment in that plan in June 2008, with the below 100% poverty as the control group. See the note to Table 2.2 for further discussion of variable definitions.

We would like to interpret the increases in enrollment as being the result of the $17.50-$52.50 monthly mandate penalty that went into effect in January 2008. However, the 2007 uninsurance penalty of $219 was assessed based on coverage status in December 2007,[22] making that month's effective penalty much larger. If consumers were only buying insurance because of the larger December penalty, we would expect many of them to leave the market soon after the monthly penalty dropped to the lower level. To assess this story, Figure 2.2 plots the probability of market exit within 1 and 6 months for new enrollees. Each point represents a distinct group of new enrollees in the corresponding month shown on the x-axis, and the y-axis value is the share who exited within 1 or 6 months. While there is a general

[22] In addition, an exemption was given for individuals who applied for CommCare in 2007 and enrolled on January 1, 2008. However, December 31, 2007, was the main advertised date for assessing coverage status.

Figure 2.2 Share of New Enrollees Exiting Within the Specified Number of Months



NOTE: These graphs show the rate of exiting CommCare coverage within one month (left figure) and six months (right) of initial enrollment among people newly enrolling CommCare in a given month. If individuals had enrolled at the end of 2007 to avoid the one-time $219 penalty and quickly dropped coverage thereafter, we would expect to see a spike for the 150-300% poverty series in December 2007 and/or January 2008. The absence of such a spike in exits suggests that the monthly penalties starting in January 2008 were sufficient to induce individuals to remain in the program. The spike that does occur among new enrollees in March 2008 reflects the start of an income-verification program for the 150-300% poverty group in April 2008. See the note to Figure 2.1 for the definition of new enrollees and the cheapest plan.

upward trend over time, there is no jump in either series for people who joined in December 2007. The large spike for people enrolling in March 2008 is due to an unrelated income verification program.[23] This analysis suggests that consumers were not enrolling for just December to avoid the $219 penalty and leaving soon after because of the lower penalty.

2.3.2 Affordable Amount Decrease Experiment

Our second strategy for identifying the effect of the price of the outside option on the cheapest plan's demand uses changes in the "affordable amount." Recall that CommCare sets subsidies so that the post-subsidy premium for the cheapest plan equals the affordable amount. Therefore, for a fixed set of pre-subsidy prices, a $1 decrease in the affordable amount raises the subsidy and lowers the premium of all

[23] The income verification program took effect in April 2008 for individuals above 150% of poverty (but earlier on for people <100% poverty). Prior to April, income group at enrollment was based partly on self-reporting, and changes in income over time were also supposed to be self-reported. The verification program uncovered a large number of ineligible people, who were dis-enrolled in April 2008 and subsequent months. This event can also explain the upward trend in exits within 6 months leading up to April 2008.

plans by \$1. In our model, this has an equivalent effect as a \$1 increase in the mandate penalty (holding plan premiums fixed), so we can use changes in the affordable amount to estimate $\eta_{jmin,M}$.

This approach addresses a concern with our first method: that the introduction of a mandate penalty may have a larger effect (per dollar of penalty) than a marginal increase in penalties.[24] Some individuals may obtain coverage to avoid the stigma of paying a penalty, but this stigma might not change when mandate penalties increase.[25] Our second approach using the affordable amount avoids this problem and also allows us to obtain estimates for the 100-150% poverty group, who faced a \$0 mandate penalty.

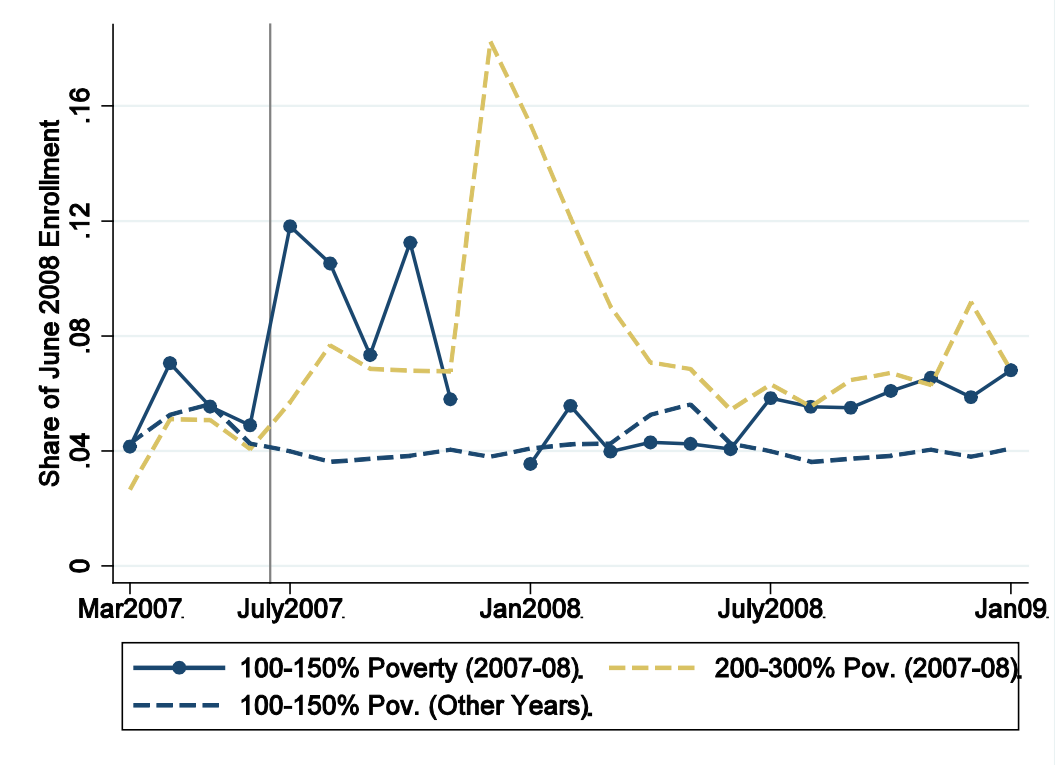
The most significant changes in the affordable amount occurred in July 2007 for consumers between 100-150% of poverty.[26] For the first half of 2007, their affordable amount was \$18 and premiums ranged from \$18 for the cheapest plan to \$74.22 for the most expensive. In July 2007, CommCare eliminated premiums for this group, so all plans became free. We can think of this as the combination of two effects: (1) The affordable amount was lowered from \$18 to zero, and (2) the premium of all plans besides the cheapest one were differentially lowered to equal the cheapest premium (now \$0). The second change should unambiguously lower enrollment in what was the cheapest plan, since the relative price of all other plans falls. Therefore, if we use the actual policy change to estimate the effect on new enrollment in the (formerly) cheapest plan, this will be a lower bound on the effect of just lowering the affordable amount (the effect we want to estimate).

[24] In principle, we could also study several marginal changes in the mandate penalty after 2008. However, these changes were small: a \$0.50 decrease for 150-200% poverty in January 2009, and a \$2-6 increase, depending on income, in January 2010. Their small size makes it difficult to distinguish enrollment changes at these times from underlying noise.

[25] An argument against the stigma explanation is that the legal mandate to obtain insurance had been in place since July 2007, though without financial enforcement. In addition, individuals below 100% of poverty were also required to obtain insurance (again without financial enforcement). However to the extent there is a stigma specifically from paying a fine for non-coverage, this concern is valid.

[26] Prices in CommCare usually change in July, but in the first year of the program, prices were held fixed from November 2006 to June 2008, so firms' price-bids did not change at the same time as this change in the affordable amount.

Figure 2.3 New Enrollees in Cheapest Plan by Month, around the Change in the Affordable Amount



NOTE: This figure shows for two income groups the monthly number of new enrollees into CommCare who chose the cheapest plan, scaled by total group enrollment in that plan in June 2008. The vertical line is drawn just before the decrease in the affordable amount (the consumer premium for the cheapest plan) from $18 to $0 for the 100-150% poverty group. During this period, the affordable amounts for the 200-300% poverty groups were essentially unchanged (they were constant at $70 for people 200-250% poverty and fell by $1 from $106 to $105 for people 250-300% poverty). The "100-150% Poverty (Other Years)" series shows average new enrollments in the corresponding calendar month in all other years in our data after June 2008. %check: why not after Mar 2008? We start the graph from March 2007, the first month with significant CommCare enrollment for these income groups. We exclude December 2007 for the 100-150% poverty group because a one-time auto-enrollment caused a sharp spike in new enrollees (to over 30% of June 2008 enrollment), and showing this point makes it difficult to see the other points in the graph. See the note to Figure 2.1 for the definition of new enrollees and the cheapest plan.

As a control group, we use the 200-300% poverty group, whose affordable amounts were essentially unchanged in July 2007.[27] We exclude other incomes from our controls for several reasons. First, we do not use the below 100% poverty group because of its somewhat different enrollment history and trends. Whereas the groups above poverty only started joining CommCare in February 2007, the

[27] The affordable amount for consumers 200-250% poverty was unchanged at $70, while the amount for 250-300% poverty was lowered by just $1 from $106 to $105. To the extent this slightly increased enrollment for the control group, it would only bias our estimates downward.

below 100% poverty group became eligible in November 2006 and had a large influx in early 2007 due to an auto-enrollment. Second, we exclude the 150-200% of poverty group from the controls because their affordable amount also fell non-trivially (from $40 to $35) in July 2007. While this smaller change does not produce as dramatic of an enrollment spike, we show in Appendix B that their enrollment increase was consistent with semi-elasticity calculated from the mandate penalty introduction results presented in Table 2.3.

Figure 2.3 shows monthly new enrollments in the cheapest plan (again normalized by each group's enrollment in June 2008) for the 100-150% poverty treatment group and 200-300% poverty control group around the July 2007 change, denoted by the vertical gray line. As an additional control, the figure shows the 100-150% of poverty group in the same calendar months in other years. Though the series are noisy there is a clear jump in the new enrollment for the 100-150% of poverty group in July 2007 and subsequent months relative to control groups. The large spike in 200-300% poverty enrollment in December 2007 reflects the mandate penalty introduction, which we used for our first identification strategy.

Table 2.4 presents the regression results corresponding to Figure 2.3. Again, in Column (1) we just look at the enrollment difference for the 100-150% poverty treatment group relative to trend, captured by CommCare-year dummies and time polynomials. Column (2) then adds the 200-300% poverty control group to form the difference-in-difference estimates. The last column does a triple-difference, further netting out changes in July-October of other years. The coefficients change a bit more between specifications, but they all imply significant enrollment increases of at least 15 percentage points. The triple-difference, our preferred specification, suggests that the decrease in the affordable amount increased enrollment in the cheapest plan by 16.9% points, implying a 0.94% effect for each $1 decrease in the affordable amount. This semi-elasticity is very similar the one we estimated from the mandate penalty introduction.

Table 2.4 Decrease in the Affordable Amount

Dependent Var: New Enrollees in Cheapest Plan / June 2008 Enrollment

Variable	(1)	(2)	(3)
Sum of July - Oct 2007 coefficients (below)	**0.200*** (0.021)**	**0.156*** (0.031)**	**0.169*** (0.032)**
100-150% Poverty x July2007	0.064***	0.060***	0.063***
	(0.005)	(0.008)	(0.008)
x Aug2007	0.052***	0.031***	0.035***
	(0.005)	(0.008)	(0.009)
x Sep2007	0.022***	0.011	0.016*
	(0.005)	(0.008)	(0.008)
x Oct2007	0.062***	0.054***	0.055***
	(0.005)	(0.008)	(0.008)
Control Group (200-300% poverty)		X	X
Triple Difference (dummies for July - October)			X
Observations	52	104	104
R-Squared	0.988	0.981	0.981

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

NOTE: This table performs the difference-in-difference regressions analogous to the graphs in Figure 2.3, with specifications analogous to those in Table 2.2. The dependent variable is the number of new CommCare enrollees who chose the cheapest plan in each month in an income group, scaled by total group enrollment in that plan in June 2008. There is one observation per income group (the 100-150% poverty treatment group, and the 200-300% poverty control group in columns (2) and (3)) and month (from March 2007 to June 2011). All specifications include CommCare-year dummy variables and fifth-order time polynomials, separately for the treatment and control group, to control for underlying enrollment trends. (The first CommCare year ends in June 2008, so there is no conflict between the CommCare-year dummies and the treatment months of July to October 2007.) Specification (3) also includes dummy variables for all calendar months of July-October for the treatment group, to perform the triple-difference. Where applicable, specifications also include dummy variables to control for two unrelated enrollment changes: (a) for 100-150% poverty in December 2007, when there was a large auto-enrollment spike, and (b) for 200-300% poverty in each month from December 2007 to March 2008, when there was a spike due to the introduction of the mandate penalty. See the note to Figure 2.1 for the definition of new enrollees and the cheapest plan.

2.4 Pricing Distortion Calculation

We use the coefficients we have estimated to calculate the semi-elasticity with respect to the mandate penalty for each income group and for the market overall. These, combined with the own-price elasticity,

Table 2.5 Calculation of Semi-Elasticities and Distortion

Statistic	Income Group (% of Poverty)			
	Aff. Amt. Change	*Mandate Penalty Introduction*		
	100-150%	150-200%	200-250%	250-300%
Enrollment % Increase	16.9%	20.8%	26.3%	25.0%
Mandate/Affordable Amount Change	$18.00	$17.50	$35.00	$52.50
Mandate Semi-Elasticity	0.0094	0.0119	0.0075	0.0048
Enrollment Shares (June 2008)	*46.1%*	*31.3%*	*15.1%*	*7.5%*

Pooled Population Calculation	
Average Mandate Semi-Elasticity	0.0095
Own Price Semi-Elasticity*	0.0197
Increase in Markup ($/month)	$47.54
(% of avg. price = $383)	(12.4%)

* Based on coefficient from Chan and Gruber (2010).

NOTE: This table puts together the estimated semi-elasticities for different groups into one overall "mandate semi-elasticity." We combine this with the estimate for in-market own price semi-elasticity from Chan and Gruber (2010) to calculate the dollar change in markup due to the endogenous subsidy, as illustrated in Equation (3.9).

allow us to calculate the price distortion due to the price-linked subsidy. Table 2.5 combines the estimated coefficients from Table 2.3 and Table 2.4 with the corresponding changes in the mandate penalty or affordable amount. Since the dependent variables in our regressions are normalized by the plan's steady-state enrollment (taken to be June 2008), a coefficient of 0.20 implies an increase of 20% of the steady-state enrollment. To find semi-elasticities, we divide this by the dollar amount of the change in the mandate penalty or affordable amount. This results in semi-elasticities that are generally decreasing in income, as one would expect. The exception is the semi-elasticity for 100-150% poverty (0.0094), which is lower than that for 150-200% poverty (0.0119), but we consider the 100-150% poverty estimate to be a lower bound because of the other price changes that happened at the same time (see discussion above).

The statistic entering our pricing distortion formula (2.6) is the overall mandate penalty semi-elasticity across all consumers, which equals a share-weighted average of each income group's semi-elasticity. Using June 2008 enrollment shares, we calculate an average semi-elasticity of 0.0095. Thus,

each $1 increase in the monthly mandate penalty increases overall enrollment in the cheapest plan by 0.95%.

The final statistic in our distortion formula is the own-price semi-elasticity of demand, η_{jmin}. Because they have already analyzed this market, we use the estimate from Chan and Gruber (2010). They report an own-price semi-elasticity of demand (of 0.0154), but because their model does not account for the outside option of uninsurance, we need to adjust for that. For new enrollees, they find a price coefficient of -0.027. Using the average market share (including the outside option) for the cheapest plan of 27%,[28] the logit model implies an own-price semi-elasticity of

$$\eta_{jmin} = -\alpha \cdot (1 - s_{jmin}) = 0.027 \cdot (1 - 0.27) = 0.0197.$$

Allowing for out-of-market substitution mechanically makes this elasticity larger than that reported by Chan and Gruber (2010), though the estimated distortion would be larger if we used their number. Plugging the elasticities into Equation (2.6) gives us

$$Mkup_{jmin}^{Endog} - Mkup_{jmin}^{Exog} = \frac{\eta_{jmin,M}}{\eta_{jmin} \cdot \left(\eta_{jmin} - \eta_{jmin,M}\right)} = \frac{0.0095}{0.0197(0.0197 - 0.0095)} \approx \$48. \qquad (2.9)$$

Thus, based on demand parameters from the Massachusetts exchange, the upward distortion in the pivotal plan's markup (and therefore the subsidy) due to price-linked subsidies is about $48 per member-month. This is substantial: it is 12% of the average price of $383 in 2007-2009 for people 100-300% poverty.

Modeling counter-factual pricing without the distortion requires making assumptions about demand away from the observed equilibrium – hence our assumption of a constant semi-elasticity of demand. This assumption is not necessary if we formulate the distortion in terms of the equivalent increase in per-member cost, instead of an increase in price. Our estimation indicates that if the distortionary incentive were removed and per-member costs for the cheapest plan were $48 per month higher, there would be no change in the pricing equilibrium.

[28] During this time period, the in-market share for the cheapest plan was 47.3%. Using the 2008 American Community Survey, we estimate that there were 66,219 Massachusetts uninsured with incomes between 100-300% poverty and citizenship status making them eligible for CommCare, relative to 2008 average monthly CommCare enrollment of 87,491 among those 100-300% poverty. This implies that 43% of potential the market do not buy insurance, so the overall market share for the cheapest plan is 0.473*(1-0.43)=0.27.

It is important to apply several caveats to this estimate. First, our results are not intended to be a perfect estimate of the historical distortion in Massachusetts. The Massachusetts market was relatively new in the period we study, so prices had probably not converged to equilibrium, and there was likely uncertainty about which plan would be subsidy pivotal. Also, we have made several conservative assumptions that may understate the size of the distortion in Massachusetts. And while our model assumes a single plan price and subsidy (as is true in the ACA), the actual pricing rules in Massachusetts at the time were a bit more complicated.[29]

Applying the estimates out-of-sample to the ACA exchanges requires even more caution. The ACA has different institutions in several dimensions: it includes higher income groups (including unsubsidized enrollees above 400% poverty); it has plans across four generosity tiers; and its subsidies are based on the second-lowest price silver plan, not the cheapest plan as in Massachusetts. Also, some ACA exchange consumers will be unsubsidized (about 20%, according to CBO projections). These consumers may help mitigate the distortion, though analysis of the Massachusetts experience (Ericson and Starc, 2012) suggests that most of these consumers will enroll in bronze plans, providing little discipline to silver plan prices. The ACA will also have medical loss ratio (MLR) requirements, which prohibit a plan from increasing its administrative costs and profits beyond an allowed share of revenues (20% in the exchanges). If MLR restrictions are effective and binding, they could limit the distortion.

Working in the other direction is the fact that the ACA has multi-plan insurers (which theoretically should exacerbate the distortion; see Section 2.2) and initial data suggests that many markets may be uncompetitive. According to a New York Times analysis of states served by the federal exchange, 58% of markets (which are counties) have two or fewer insurers, and about 20% of markets have just one insurer (Abelson et al. (2013)). These areas are disproportionately small and rural, but the distortions

[29] Specifically, the complications in plan years 2007-08 were: (1) It allowed for separate pre-subsidy prices for the 100-200% poverty and 200-300% poverty groups, so these were separate markets with different distortions; (2) the exchange had minimum prices (imposed under federal Medicaid rules), which appear to have been binding in some cases, meaning that the cheapest plan could not have been priced lower even without the distortion. Neither of these complications apply under the ACA, so we have not included them in our model.

there are potentially sizable. While normally the distortion is capped by the price of the third-cheapest silver plan, counties with one or two insurers may not have a third-cheapest plan (or it may be controlled by the same insurer as the second-cheapest). Along with the usual benefits of competition, our model suggests that having at least three (and preferably more) insurers in ACA exchanges is important for mitigating the distortions from price-linked subsidies.[30] While, we cannot estimate what the size of the distortion will be in the ACA exchanges, our estimates suggest that it has the potential to be substantial and thus is an important consideration for policy makers designing the exchanges.

2.5 Discussion

In this section, we discuss the policy implications of our results. So far, we have discussed two alternatives for subsidy design in an insurance exchange: exogenous and price-linked subsidies. We have argued that, relative to exogenous subsidies, linking subsidies to prices increase the incentive for insurers to raise prices. We now propose a third subsidy alternative that would eliminate the distortion while maintaining price-linked subsidies with their advantages. However, this alternate policy has its own weaknesses, and we next compare the advantages and disadvantages under all three possible subsidy designs. Finally, we discuss the implications of our findings for insurance programs other than the ACA.

2.5.1 Correcting the Distortion while Maintaining Price-Linked Subsidies

Our model suggests a simple alternative to exogenous subsidies that would fully correct the distortion while keeping price-linked subsidies: apply the same subsidy to the mandate penalty. Specifically, set an exogenous base mandate penalty amount M_0 and reduce this by the subsidy, for a final penalty of $M = M_0 - S(P)$. The base amount could be set so that, based on the government's expectations about

[30] Single-insurer markets are even more problematic. In theory, the lone insurer could turn the subsidy rules into a money machine by raising price arbitrarily without losing subsidized consumers. Although regulations like minimum loss ratio requirements will certainly mitigate this extreme outcome, single-insurer markets are particularly worrying in light of our analysis.

prices and the subsidy, the final penalty would be similar to the penalty specified by current law.[31] This adjustment would remove the pricing distortions discussed above. If the cheapest plan raised its price bid by $1, this would leave its premium unaffected (since the subsidy would rise) but lower by $1 the price of all other plans and the cost of uninsurance. The effect on own demand would be

$$\frac{dQ_{jmin}}{dP_{jmin}} = \sum_{k \neq jmin} \left(-\frac{\partial Q_{jmin}}{\partial P_k^{cons}} \right) - \frac{\partial Q_{jmin}}{\partial M} = \frac{\partial Q_{jmin}}{\partial P_{jmin}^{cons}}$$

The price has only its direct effect on the demand for each plan, so even if insurers offer multiple plans (the ACA case), the subsidy does not distort incentives. Therefore, optimal prices would be identical to the benchmark condition with exogenous subsidies.

A concern with this approach is that it creates uncertainty about the mandate penalty. If prices (and therefore the subsidy) are higher than expected, the final mandate penalty would be lower than the target, likely leading to an increase in the uninsurance rate. This could be desirable: since the cost of insurance has risen, it may be optimal for fewer people to be insured. Alternatively, it would be undesirable if the price increase also signified an increase in the social cost of uninsurance -- for instance, through a general health care cost increase that also increases the externality of uncompensated care (Mahoney, 2012). While we have abstracted away from adverse selection, it is a potential concern in these markets and the endogenous mandate penalty could exacerbate it: unexpectedly higher prices would lower the mandate penalty, which could lead healthier consumers to exit, further driving up prices. Spatial variation in prediction error would also imply variation in the final mandate penalty, which could be seen as inequitable.

2.5.2 Comparing Subsidy Structures

We have so far discussed three subsidy alternatives: (1) exogenous subsidies, (2) price-linked subsidies with an exogenous mandate penalty, and (3) price-linked subsidies also applied to the mandate penalty. If

[31] In particular, policy makers would want to increase M_0 over time with their estimate of medical cost growth, to avoid having the mandate penalty decline as costs and therefore prices and subsidies increase.

insurers always priced at cost, option (2) (the current ACA policy) would have desirable properties: it would allow the government to take on price risk, guarantee affordability for consumers, and ensure a certain mandate penalty. Unfortunately, in imperfectly competitive markets, this policy distorts pricing incentives – and our results suggest the distortion is substantial.

The other two alternatives do not distort firm pricing, but each has its own disadvantage. Exogenous subsidies shift price risk onto poor enrollees and may, if they grow too large, push post-subsidy prices to their lower bound of zero. Price-linked subsidies applied to the mandate penalty create certainty about consumer prices but shift the volatility onto the mandate penalty. Both of these policies involve the government making a prediction for prices at the market (county) level to properly set exogenous subsidies or the base mandate penalty (M_0). If such predictions can be made reasonably well, then these disadvantages will be mitigated. Unfortunately, experience with Medicare Advantage's exogenous subsidies suggests that this process can be challenging.

Thus, none of the three policies is perfect. The optimal policy could involve a mixture of two or more of them. For instance, a 50/50 mixture between exogenous subsidies and our policy suggestion would increase subsidies by 50 cents for each $1 increase in prices (passing 50 cents onto consumers) and reduce the mandate penalty by 50 cents. Because the net insurance subsidy, $S + M$, remains constant, the distortion would be fully corrected. Alternatively, a 50/50 mixture between ACA policy and our policy would raise subsidies by the full $1 and reduce the mandate penalty by 50 cents for each $1 increase in prices. This would not eliminate the distortion, but would reduce it.

2.5.3 Broader Relevance for Health Insurance Programs

Although our empirical estimates are from a specific setting (the subsidized Massachusetts exchange in 2007-2008), we believe that our theoretical point about distortions with price-linked subsidies is relevant more broadly.

The applicability of our theory to a market depends on two factors: (1) There must be some substitution to an unsubsidized outside option, and the distortion is larger the greater is this

substitutability. (2) Insurers must have some market power,[32] and the distortion is larger with greater market power. Price-linked subsidies work by exacerbating existing market power – by effectively removing the competition of the outside option – so are more severe where baseline market power is greater.

We have discussed the close link with the ACA exchanges, but the theory also applies to Medicare Advantage, Medicare Part D, and employer-sponsored insurance programs. In the Medicare Advantage market, which uses exogenous baseline subsidies,[33] our results suggest that switching to endogenous subsidies would create a pricing distortion unless the subsidy was also applied to traditional Medicare (as in "premium support" proposals). As long as the government does not have a preference for whether consumers use traditional Medicare or private plans, the volatility of the price of the outside option is not an issue in this case. In order to avoid the potential for negative prices for traditional Medicare (if the Medicare Advantage bids are high), a premium support system would need to count the cost of traditional Medicare as another bid in the market.

Medicare Part D uses flat, price-linked subsidies (as in the ACA), but the distortion for the main subsidy is probably relatively small (though the low-income subsidy is likely larger; see Decarolis (2013)). The subsidy is based on a national enrollment-weighted average of plan price bids. Because all plans' prices affect the subsidy through this average, our theoretical distortion applies to all plans -- not just a subset of potentially pivotal silver plans as in the ACA -- but the distortion for each plan is smaller. It is approximately proportional to the national market share of the plan's parent insurer, the largest of which is United Health Group with 28% in 2011 (see Decarolis, 2013).

Employers who subsidize insurance also need to consider the trade-off between making prices predictable for employees and distorting pricing incentives. Employers typically pick a small menu of options for their employees and set subsidies based on prices (either implicitly or explicitly). To the

[32] The presence of market power is not a restrictive condition; it merely requires that a $1 price increase does not cause a plan's demand to fall to zero (as would be the case in perfect competition).

[33] However, after applying this baseline subsidy, Medicare reduces the subsidy when a plan reduces its price below the benchmark. This creates a different kind of pricing distortion, which may be significant.

extent that an employer's chosen insurer(s) have market power (for evidence of insurer market power, see Dafny (2010)), this can lead to the same pricing distortions we discuss. Because under tax rules employers cannot subsidize employees' outside options to the same extent (e.g., obtaining coverage through a spouse), they would need to use exogenous subsidies to avoid this distortion.

2.6 Conclusion

This paper considers the distortion of pricing incentives generated by price-linked subsidies in health insurance exchanges, an important topic for economists analyzing these markets and policy makers designing and regulating them. We highlight this distortion in a simple theoretical model and derive sufficient statistics for estimating its size. We then use two natural experiments in the Massachusetts exchange to confirm that insurance demand responds to the relative price of the outside option. Using the Massachusetts estimates to calibrate our model and assuming constant semi-elasticities, we find an upward distortion of the subsidy-pivotal cheapest plan's price of $48 per member-month, or 12% of the average price of insurance. The potential budgetary effect is substantial: Massachusetts had about 2 million member-months of subsidized coverage in fiscal 2009, so a $48 increase in monthly subsidies would translate to $96 million per year in government costs. For the ACA, using the CBO projection of 20 million subsidized enrollees annually, a $48 per month subsidy increase would translate to $11.5 billion per year in federal spending.

This estimate suggests the importance of price-linked subsidies in affecting equilibrium pricing incentives and subsidies in a health insurance exchange. Nonetheless, we do not view it as a perfect estimate of either the historical distortion in Massachusetts or the expected distortion in the ACA exchanges. Rather, we think that our calibration implies that the pricing distortion we identify in theory is of practical concern to policy makers deciding on subsidy rules.

How policy should respond depends on balancing several goals: preventing the pricing distortion, guaranteeing “affordable” post-subsidy premiums, and ensuring a target mandate penalty. The current

ACA price-linked subsidies guarantee affordability and a fixed mandate penalty but involve pricing distortions. Exogenous subsidies prevent distortions and allow for a fixed mandate penalty but shift risk to poor consumers and cannot guarantee that post-subsidy premiums will be affordable. We present an alternate policy that guarantees affordability and eliminates the pricing distortion but does so by reducing the mandate penalty when prices are higher than expected. This policy carries the risk that unexpectedly high prices in an area would increase uninsurance. Thus, none of the alternatives is perfect, and optimal policy could involve a mixture between two or more of them. Nonetheless, the potentially large cost of the pricing distortion under current ACA policy makes reform of subsidy design worthy of consideration.

While our theory and empirics are focused on the ACA, we think the point that price-linked subsidies distort pricing incentives is applicable more generally. It applies to any program that uses price-linked subsidies and has an outside option. In addition to analyzing the ACA markets as data becomes available, future research could seek to measure the relevant elasticities in order to assess the quantitative importance of pricing distortions in Medicare Advantage, Medicare Part D, and employer-sponsored insurance programs.

Chapter 3

Social Security Claiming and the Annuity Puzzle

3.1 Introduction

Economists have long known that annuities provide an efficient means of funding retirement needs over an uncertain lifespan because they pool longevity risk across many individuals. As a result of this advantage, a classic prediction from life cycle models is that retirees will annuitize all of their wealth, aside from any portion designated for bequests (Yaari 1965). In reality, most U.S. retirees hold no annuities outside of mandatory Social Security, which represents about half of retirement wealth. Further, defined benefit pensions – the only other significant source of annuitized wealth – have been declining in favor of defined contribution pensions that typically pay out a lump sum (Poterba, Venti, and Wise 2009a, 2009b).

Understanding why annuitization is low is critical for evaluating these trends and modeling life cycle behavior. One possibility is that the annuities provided by Social Security and pensions are already sufficient, and retirees are rationally declining additional annuitization (Bernheim 1991). In the life cycle model, this argument requires that retirees have high discount rates (Gustman and Steinmeier 2005; Warner and Pleeter 2001), strong bequest motives or intra-family risk sharing (Jousten 2001; Kotlikoff and Spivak 1981), or precautionary motives due to uncertain health expenditures (Ameriks et al. 2011; Turra and Mitchell 2004). A second possibility is that low annuitization could reflect retiree misunderstanding of annuities or other non-standard, behavioral factors absent from the life cycle model (Brown 2007; Brown et al. 2008, 2013). Distinguishing these explanations has been challenging because aside from restrictive cases, the optimal level of annuitization is theoretically ambiguous. Further, because

so few people buy annuities outside of pensions, empirical studies of annuity demand have been limited by data availability.

To make progress on these issues, I study an annuitization decision that is well known but has received little attention in the literature: the timing at which retirees claim Social Security benefits.[1] As the U.S. social insurance program for the elderly, Social Security is the largest source of annuity income, dispensing $509 billion in benefits to 41.7 million people in 2008. Importantly, Social Security offers beneficiaries flexibility in the size of their annuity benefits. If individuals take up or "claim" benefits at the earliest eligibility age (age 62), they receive the smallest allowed annuity. By delaying past 62 (up to a maximum age of 70), an individual forgoes a lump sum of benefits in return for a higher annuity at take-up. Although beneficiaries usually cannot receive benefits before retiring, they are free to delay claiming past retirement, by which they make a purely financial decision to purchase a larger annuity. Nonetheless, Coile et al. (2002) find that the vast majority of men claim benefits almost immediately upon becoming eligible, a fact that is also true in the more recent data I study. As with commercial annuities, retirees typically obtain as little as possible of the Social Security annuity. This is notable because Social Security delay represents a particularly attractive annuity. It is inflation-adjusted, government-guaranteed, and actuarially fair or better over a variety of ages (Shoven and Slavov 2013) – versus commercial annuities which are typically 15-20% worse than fair for an average retiree (Mitchell et al. 1999).

To evaluate retirees' annuitization decisions, I take a different approach than much of the past literature. This literature typically calibrates a life cycle model to study whether factors like actuarial unfairness or bequest motives can explain retirees' low levels of annuitization. But the conclusion of these analyses can be challenging to interpret because they depend on a host of parameter assumptions. To sidestep this issue, I focus on whether life cycle theory can explain retirees' *marginal* decisions not to annuitize further. Studying marginal annuitization lets me evaluate its optimality through perturbations

[1] Despite voluminous research on Social Security (see Feldstein and Liebman (2002) for a review), the claiming decision has received little study, other than to document the fact that beneficiaries claim soon after retirement and show that this is hard to explain given the substantial incentives to delay claiming (Coile et al. 2002; Sun and Webb 2011).

around the observed levels of assets and annuity income. These perturbation arguments both clarify the logic for annuitization and obviate the need for a complete model of preferences.

Social Security claiming is an ideal setting for this approach because early claiming is an active choice not to purchase a small, marginal annuity. Delay is rewarded in monthly increments, so a person who claims n months after retirement indicates that the annuity available from the n+1 month of delay was not worth the cost. I show that life cycle theory delivers testable sufficient conditions for the optimality of a marginal annuity based on the annuity return, subsequent asset decumulation behavior, and risk probabilities. Because these variables are at least partly observable, I can evaluate early claiming with simple tests that do not rely on detailed assumptions on individuals' utility parameters.

I start by analyzing early claiming in the standard life cycle model without bequest motives or risky liquidity shocks. In this setting, the life cycle model delivers the strong predictions about the optimality of additional annuitization for liquidity unconstrained agents. These predictions are based on the logic, shown by Davidoff, Brown, and Diamond (2005), that annuities have an "arbitrage-like dominance" over non-annuitized assets. A retiree can delay claiming – while selling other assets to replace the lost benefits and avoid reducing consumption – and increase their permanent benefit by 7-8%. Effectively, the retiree sells conventional assets (returning perhaps 3%) to "buy" Social Security benefits returning 7-8%. As long as the retiree is not liquidity constrained, this transaction generates a gain that allows retirees a pure increase in consumption.[2] The only barrier to this arbitrage opportunity is liquidity constraints that prevent retirees from reallocating resources intertemporally. This explanation is easy to test, since retirees with plenty of assets cannot be liquidity constrained.

For the minimum marginal delay of one month, the liquidity requirements are quite low – never more than a single month's benefits, or about $1,000 for a typical beneficiary. Using panel asset data from the Health and Retirement Study, I show that at least 70% of non-disabled retirees had sufficient non-

[2] Another way to understand this arbitrage logic is that it increases lifetime consumption by reducing incidental bequests – which by the assumption of no bequest motives are costless to cut. This transfer (called the "mortality premium") allows annuities to deliver super-normal returns like the 7-8% available from delaying Social Security.

housing assets at all observed ages to carry out this arbitrage and delay Social Security past their observed claiming date. Most retirees had sufficient assets for much longer additional delays: about 50% had assets high enough to delay three years longer and about 40% had enough to delay all the way to age 70. Under standard life cycle assumptions without bequest motives, this forgone arbitrage opportunity is quite valuable. For instance, delaying from 62 to 63 would be worth at least $5,600 immediately at age 62 for a typical retiree holding positive assets through age 90, while delaying from 62 to 65 would be worth at least $13,000 in pure gain.[3] These calculations indicate the magnitude of the puzzle of early claiming in the basic life cycle model without bequest or precautionary motives.

Behind this result are two basic facts about retiree behavior that are in tension in the life cycle model. Retirees spend down wealth extremely slowly[4] – as if worried about outliving their resources – but simultaneously fail to purchase annuities, including through Social Security delay. A life cycle model without bequest or precautionary motives has no way of reconciling these facts with intertemporal optimization. Therefore, I next consider adding bequest motives or precautionary motives due to uninsured health care costs. Recent research suggests that these factors may be able to explain slow asset decumulation.[5] Intuitively, they could explain low annuitization because they are preferences for two features that annuities lack: preservation after death and liquidity in an emergency.

However, I find that both theoretically and empirically, these factors have limited ability to explain early Social Security claiming. The theoretical reasoning again highlights retirees' choice to turn down the marginal annuity available from an incremental delay. A key fact underlying the argument is that the marginal Social Security delay was substantially better than actuarially fair for most people in the 1930s birth cohort I study. While the benefit schedule was close to fair when it was implemented in 1961, it has been left largely unchanged since then, despite substantial gains in elderly longevity. As a result, the

[3] This calculation assumes a $10,000 annual benefit (close to the sample median) and Social Security rules for the cohort born before 1938. See Section 3.2.2 for details about the calculation method.

[4] For evidence on this, see Dynan, Skinner, and Zeldes (2004) and Love, Palumbo, and Smith (2009).

[5] See Dynan, Skinner, and Zeldes (2002); Lockwood (2014); De Nardi, French, and Jones (2010).

real return of 8.3% for delaying from 62 to 63 (the relevant margin for most people) was 15% better than actuarially fair for an average person born in 1930.[6] Put another way, an individual would need to have an annual mortality risk that is 47% higher than in the actuarial life tables for the delay from 62 to 63 to have been exactly fair. Above-fair returns are even larger for sub-populations like women and the highly educated who have longer life expectancies but face the same Social Security schedule as everyone else.

I show that when an annuity has above-fair returns, a standard expected utility model of the bequest motive predicts that additional annuitization is optimal, regardless of the preference for bequests relative to consumption. The logic is that the annuity both increases the expected present value of the bequest and provides insurance value against the bequest declining with age. Thus, retirees can increase their bequest utility in both a first-order and second-order sense (again without having to reduce consumption). I test this theory empirically using both actuarial and self-reported longevity probabilities and show that at least half of beneficiaries would have gained in this unambiguous sense by delaying a year longer. However, a significant caveat to this conclusion is its reliance on an expected utility bequest motive with a discount rate equal to the interest rate. Retirees might claim early due to bequest motives if they myopically overweight near-term bequests relative to bequests at older ages.

I show that a similar theory carries over to precautionary motives in the presence of risky liquidity shocks. If the risk is sufficiently concentrated at older ages, retirees can derive insurance value by annuitizing assets and saving out of the annuity payments. Intuitively, annuitization provides a valuable and otherwise missing tool for life cycle planning: the ability to transform assets early in retirement into assets later in retirement without sacrificing consumption should death occur early. The conditions under which a retiree would get insurance value by using this tool are similar to those for bequest motives, with the liquidity risk probabilities replacing the mortality probabilities in the theory. I test this theory on the prime example of elderly liquidity shocks: medical expenses from long-term

[6] This calculation follows the Social Security Administration in using a 3% real interest rate and its cohort life table projections (Bell and Miller 2005). See Shoven and Slavov (2013) for additional calculations showing that delaying Social Security is better than fair on average, particularly given today's low interest rates.

nursing home care costs not covered by Medicare.[7] The distribution of nursing home stays in the HRS is even more concentrated at advanced ages than is mortality (presumably because people who die at younger ages are less likely to spend a long period in a nursing home). Therefore, nursing home risks have even less ability than bequests to explain early claiming. An open question that I plan to address in future revisions is whether there is another unspecified risk that could explain reluctance to annuitize. To do so, it would have to be both concentrated early in retirement and severe enough to completely exhaust assets.

These results cast doubt on the ability of standard life cycle forces to explain early Social Security claiming. I also argue that simple explanations such as lack of information or political risk to Social Security benefits are unlikely to explain the puzzle. Rather, an understanding of early claiming is more likely to rest in a non-standard behavioral explanation. In the final section I discuss three such explanations and how I can test them in future work on this project.

The paper is organized as follows. Section 3.2 provides background on Social Security claiming and tests whether a life cycle model without bequest or precautionary motives can explain claiming patterns. Section 3.3 considers the theory of claiming with bequest and precautionary motives and tests their explanatory power. Section 3.4 concludes and discusses the next steps in this project.

3.2 Social Security Claiming in the Basic Life Cycle Model

The largest U.S. social insurance program, Social Security provided annual benefits of $509 billion for retirees, their dependents, and surviving spouses in 2008 (SSA 2009).[8] Because benefits continue at a constant real level until death, Social Security is an inflation-protected life annuity. Importantly, retirees partially choose the size of this annuity by their timing of benefit take-up, or "claiming." Beneficiaries

[7] I focus on long-term nursing home stays (longer than 60 days), since most other medical expenses are covered by Medicare or by supplemental insurance held by about 90% of the elderly. See De Nardi, French, and Jones (2010); Marshall, McGarry, and Skinner (2011).

[8] Social Security also has a $106 billion Disability Insurance component, which I will not study because it does not involve an annuitization decision.

who claim at the earliest age allowed – 62 for most people[9] – choose the minimum annuity. Each month of delay past 62 permanently increases the benefit size at take-up. It is easy to see that delaying executes a transaction equivalent to purchasing an incremental, deferred annuity. A retiree "pays" one month's benefits in exchange for a larger stream of benefits from the new claiming date until death.

While past work has often assumed simultaneous retirement and benefit claiming, the two need not occur together. Labor force exit effectively sets a lower bound on the claiming age.[10] But nothing prevents claiming *after* labor force exit. Retirees can continue to raise their annuity benefits by delaying through age 70, beyond which further delay is not rewarded. Delaying past retirement may require financial adjustment to fund consumption in the interim but does not have any other real costs. In this paper, I will focus on understanding how retirees make this financial annuitization decision, holding the retirement decision fixed.

The benefit schedule is intended to be actuarially fair, so that claiming early or late does not affect the total expected present value of benefits received. While, for reasons discussed in Section 3.3, delayed claiming is often better than actuarially fair, it is important to realize that actuarial fairness does not imply indifference over claiming ages. Rather, actuarially fair delay is nearly always optimal in the life cycle model because it purchases an annuity. Like other forms of insurance, annuities are valuable because they transfer resources from low to high marginal utility states. Here, the low marginal utility state is death and the high marginal utility state is life. When the marginal utility of income in death is zero – the case without bequest motives – Davidoff, Brown, and Diamond (2005) show that the correct benchmark for annuities is whether their return exceeds the interest rate on bonds of comparable risk. For

[9] The major exceptions are widow(er)s claiming benefits through their deceased spouse's earnings record, who can claim benefits at 60, and disabled workers, who can start benefits five months after the start of their disability at any age.

[10] This lower bound is enforced through the "earnings test" rules, which prevent beneficiaries younger than the "full retirement age" (formerly 65 but now rising gradually to 67) from collecting full benefits while working. Current benefits are reduced by 50 cents for each dollar of earnings above a modest limit ($14,160 per year in 2011). These withheld benefits are refunded in an actuarially equivalent increase in benefits after the full retirement age, after which simultaneous work and benefit receipt is allowed.

Social Security, a riskless real annuity, delay is optimal if its return (the benefit increase as a percentage of benefits forgone while delaying) exceeds the real interest rate on inflation-protected Treasury bonds.[11]

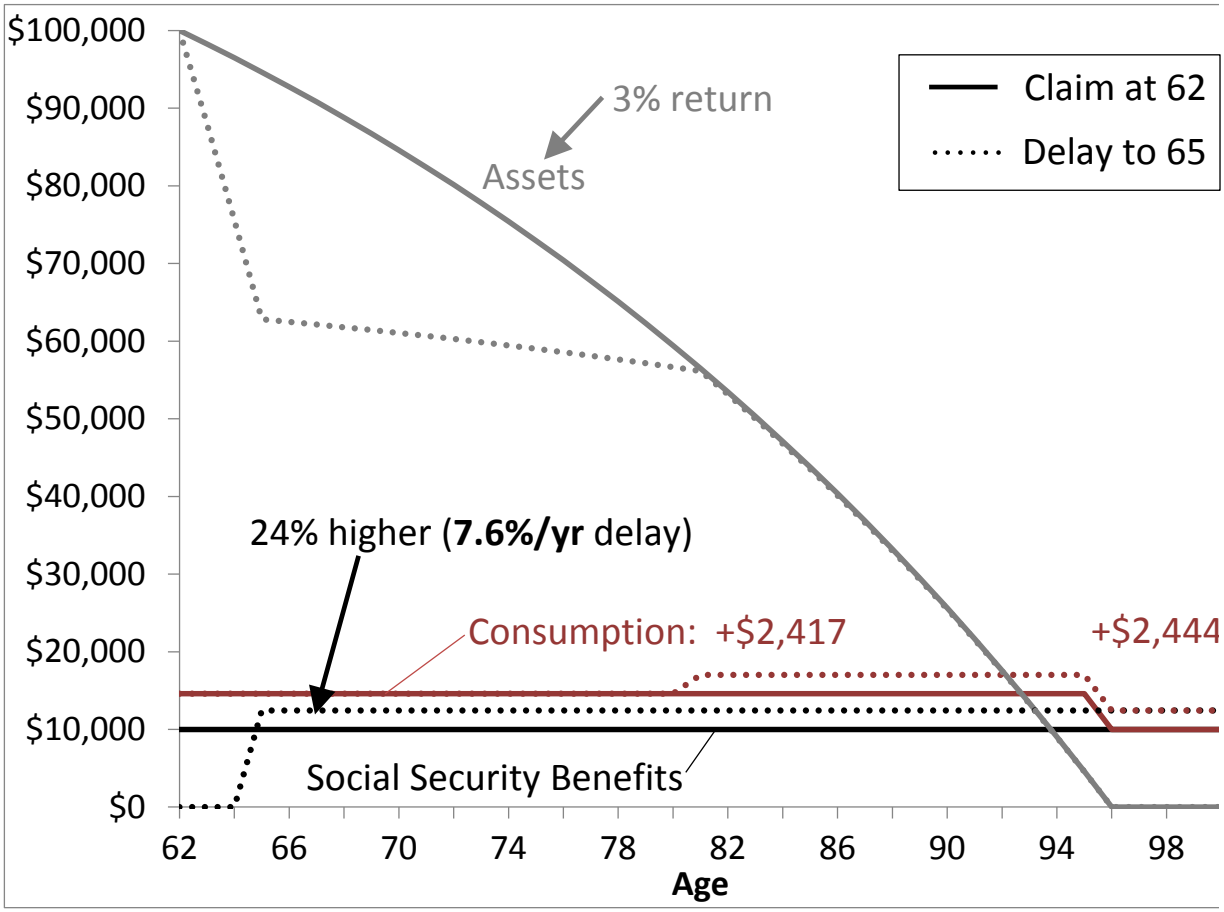
To see why, consider the example shown in Figure 3.1 of a 62-year-old single retiree who would receive a Social Security benefit of $10,000 per year conditional on claiming at 62. The retiree starts with $100,000 in assets that return 3% per year and has no other pension or annuity income. The solid lines show the benefit path and a particular consumption/asset plan the retiree could choose by claiming at 62. If the retiree dies early, the plan is truncated and any remaining assets are bequeathed. If instead, the retiree waits until 65 to start benefits, his real annual benefit would be 24% higher – on average 7.6% per year of delay.[12] Critically, the 7.6% annual increase in Social Security far exceeds the 3% return on assets. In the language of Davidoff, Brown, and Diamond (2005), the annuity from delaying has an "arbitrage-like dominance" over non-annuitized assets. The retiree can exploit this dominance by delaying to 65 and spending down assets more rapidly to maintain consumption, as shown in the dotted lines. The higher benefit at 65 allows the retiree to rebuild assets and achieve a riskless increase in consumption by $2,417 (or 17%) per year starting at 81, and $2,444 (24%) after assets are exhausted at 96. Alternatively, if the retiree spent down assets more quickly, consumption could increase immediately by $779 (5%) per year from age 62 through 95 and by $2,444 after asset exhaustion.

This illustration shows how delaying Social Security facilitates pure increases in life cycle consumption. Consumption increases can occur immediately as well as later in life, so there is no intertemporal tradeoff. Rather, the cost of delay is a reduction in liquid assets early in retirement. Critically, the basic life cycle model without bequest motives or stochastic liquidity shocks takes a clear stand on this tradeoff. As long as the asset reduction is *feasible*, the life cycle model considers it *costless*

[11] These rates have varied but have generally fallen in the range of 1.5-3.5% since their inception in the late 1990s, with lower rates in the recent past and higher rates for longer maturity bonds. To be conservative, I assume a real interest rate of 3.0%, which is what the Social Security Administration assumes.

[12] This follows the Social Security rules for an individual born between 1943-1954. See Table 3.1 for returns of other cohorts. Because benefits are also increased for any inflation between ages 62 and 65, the 24% figure is a real increase.

Figure 3.1 Social Security Delay Example #1: Pure Consumption Increase



NOTE: This graph illustrates how delaying Social Security from can result in pure gains in consumption, which makes it unambiguously optimal in the basic life cycle model. The solid lines show a life cycle plan for a hypothetical single retiree who starts with $100,000 in assets, claims a Social Security benefit of $10,000 per year at 62, and chooses consumption to amortize his assets to age 95. Early death truncates the plan, with any remaining assets left to heirs. The dotted lines show a feasible plan if the retiree instead delays Social Security to 65 without delaying retirement, using the Social Security rules for an individual born between 1943-1954 (which for delaying from 62 to 65 are quite similar to the rules for earlier cohorts). Because she has sufficient assets to smooth consumption, consumption never falls and rises by a substantial $2,417 per year after 81. The source of this gain is the 7.6% real return on Social Security delay, which exceeds the 3% real return on assets.

for utility. Therefore, the proposed Social Security delay will be unambiguously optimal for a retiree with sufficient assets or access to credit.

3.2.1 Test of Basic Life Cycle Theory

I now formalize this intuition that delay is optimal for retirees with sufficient assets and derive an empirical test of the basic theory. Consider a life cycle model with general utility function

$U(c_t, c_{t+1}, c_{t+2}, ...)$ that is increasing in each consumption argument. The standard time-separable, exponentially discounted utility is a special case of this more general specification. Let assets earn gross riskless return $R \geq 1$ and be constrained to exceed liquidity constraint L_t at age *t*. Let b_s be the benefit level conditional on claiming at age *s,* and define the "return" on delaying Social Security as the resulting increase in benefits: $R_{SS}^{s,s+k} \equiv b_{s+k} / b_s$. The two key assumptions are as follows:

Assumption 1: No utility *directly* from assets: $\left.\frac{\partial U}{\partial a_t}\right|_{\{c_\tau\}_{\tau=0}^{\infty}} = 0$ for all *t*

Assumption 2: There are no uninsured stochastic liquidity shocks.

Under these assumptions, the following result lays out sufficient conditions for delay to be optimal.

Proposition 1: Let Assumptions 1 and 2 hold. Consider a retiree planning to claim Social Security at age *s* and consume and spend down assets according to the feasible plan $\{c_t, a_{t+1}\}_{t=s}^{\infty}$. Delaying to age $s+k$ is optimal whenever the following are true:

(a) Excess return on Social Security delay: $R_{SS}^{s,s+k} > R^k$

(b) Unconstrained assets:

$$a_{t+1} - L_{t+1} \geq \begin{cases} \sum_{i=s}^{t} R^{i-s} \cdot b_s & \text{for } t = s, ..., s+k-1 \\ \max\left\{0, \sum_{i=s}^{t} R^{i-s} \cdot b_s - \sum_{i=s+k}^{t} R^{i-(s+k)} \cdot b_{s+k}\right\} & \text{for } t \geq s+k \end{cases}$$

The proof of this proposition (which is shown in the appendix) is by construction. I demonstrate a feasible consumption plan from claiming at $s+k$ that is higher than $\{c_t\}_{t=s}^{\infty}$, as in Figure 3.1.

This result is analogous to that of Davidoff, Brown, and Diamond (2005) and is equally general. When empirically testable conditions (a) and (b) hold and when liquidity shocks are ruled out, Social Security delay is optimal for any utility function satisfying Assumption 1. This rules out bequest motives

but places no restrictions on discount rates, mortality probabilities, or intertemporal separability. Indeed, the baseline consumption plan need not even be optimal (because the result does not invoke any envelope conditions). Regardless of what the retiree was planning, she could consume more by delaying to $s+k$ because of the arbitrage-like transaction of selling assets that return *R* and buying an annuity that returns $R_{SS} > R$.

Table 3.1 tests the Social Security return condition in (a), showing the real benefit increases a single retiree earns from delay. The rates vary across cohorts because of legal changes and across ages because Social Security uses a linear, rather than a log-linear, schedule. Nonetheless, for every cohort the return per year's delay before age 65 exceeds 6.7% – well above conventional risk-free interest rates and comparable to the average return on much riskier equities. Delaying past 65 was historically less generous. But for more recent cohorts born after 1930, delaying through age 70 always returned at least 4.1% per year.

These returns are calculated for a single retiree. However, benefit increases for couples are usually even larger for two reasons. First, the legal benefit increases for low-earning spouses (usually women) who claim 50% of their husband's benefit are higher -- often exceeding 10% -- as shown in the bottom panel of Table 3.1. Second, a higher-earning spouse, by delaying claiming, increases his partner's survivor benefit should she outlive him. For simplicity, I do not consider couples' incentives in this draft of the paper, though I plan to address them in a future draft.

A retiree who delays Social Security must offset the foregone benefits by adjusting one or more of consumption, income, and/or assets. The levels in condition (b) are how much lower assets would be at each age if the offset came entirely from asset reductions. Even without access to credit, an arbitrage transaction like the one shown in Figure 3.1 would be feasible for a retiree planning to hold assets above these levels. While asset plans are unobservable, in a model without stochastic liquidity shocks, planned assets equal realized assets. This motivates comparing the asset levels in (b) to observed assets for older Americans.

Table 3.1 Real Benefit Increase from Delaying Social Security

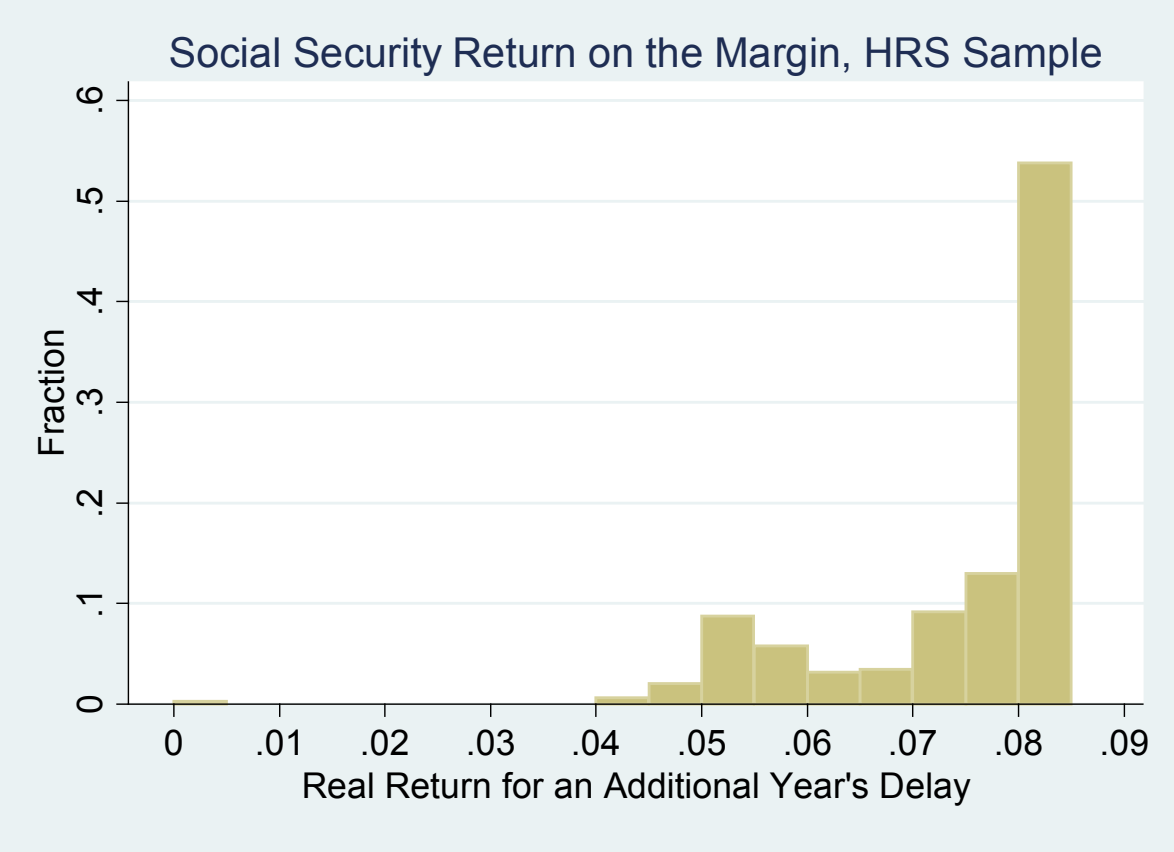
Real Benefit Increase from Delaying Social Security

Birth Year	Retiree Delaying between ages:				
	62 to 63	63 to 64	64 to 65	65 to 66	66 to 70*
1925-26	8.3%	7.7%	7.1%	3.5%	3.2%
1927-28	8.3	7.7	7.1	4.0	3.6
1929-30	8.3	7.7	7.1	4.5	4.1
1931-32	8.3	7.7	7.1	5.0	4.5
1933-34	8.3	7.7	7.1	5.5	4.8
1935-36	8.3	7.7	7.1	6.0	5.2
1937	8.3	7.7	7.1	6.5	5.6
1938	8.1	7.8	7.2	6.6	5.7
1939	7.8	7.9	7.3	7.0	6.1
1940	7.5	8.0	7.4	7.1	6.2
1941	7.2	8.1	7.5	7.3	6.6
1942	7.0	8.2	7.6	7.2	6.7
1943-54	6.7	8.3	7.7	7.1	7.2
Spouse's Birth Year	**Low-Earning Spouse Claiming Delay between ages:**				
	62 to 63	63 to 64	64 to 65	65 to 66	66 to 70*
1937 or Earlier	11.1%	10.0%	9.1%	0.0%	0.0%
1938	10.5	10.2	9.2	1.4	0.0
1939	9.8	10.3	9.4	2.9	0.0
1940	9.2	10.5	9.5	4.3	0.0
1941	8.5	10.7	9.7	5.9	0.0
1942	7.8	10.9	9.8	7.5	0.0
1943-54	7.1	11.1	10.0	9.1	0.0

* Annualized average increase *Source: Social Security Administration*

NOTE: This table shows the real percent increase (above inflation) in Social Security benefits per year of delay at various ages. The final column shows the average annualized increase over the four-year delay from 66 to 70. The top panel shows the increase for a single retiree, assuming no additional earnings while delaying. The bottom panel shows returns for an individual with low lifetime earnings claiming based on his or her spouse's earnings record. I report percent increases (as in a rate of return calculation), rather than increases as a percentage of the (fixed) Primary Insurance Amount, because percent returns are directly connected to the theory in Section 3.2. When these rates of returns from delay exceed the risk-free interest rate, the basic life cycle model predicts delay is optimal for all retirees who are not liquidity constrained.

Figure 3.2 Social Security Delay Marginal Return for HRS Sample



NOTE: The graph shows the distribution of real increases in Social Security benefits (above inflation) for the HRS sample (described in Section 3.2 and Appendix B) had they delayed an additional year beyond their observed claiming ages. As shown in Section 3.2.1, these values are equivalent to real returns on Social Security delay, which are directly comparable to real interest rates on assets. The calculation considers only the effect on own worker benefits, for simplicity ignoring spousal incentives which usually raise the return on delaying. Because most people claimed at 62 -- when the marginal returns are highest for this cohort born in the 1930s -- more than half of the sample has marginal real returns above 8%. Ninety-five percent of the sample has real returns of 5.0% or higher. Only the very few individuals (0.2% of the sample) who claimed at age 70 or later, when the marginal returns are zero, had returns below a conventional real interest rate of 3%.

To implement such a test, I use data from the Health and Retirement Study (HRS), a nationally representative panel survey of older Americans. Data on assets, income, health, and many other household variables are available biennially since 1992, and I use the first nine waves through 2008. Starting from the full sample, I exclude Social Security Disability Insurance recipients and others (mostly widows) who start Social Security before 62. To ensure a sufficient period of observation, I focus on those born between 1931 and 1938. I also exclude those who enter the sample after turning 62, who exit

before 62, or whose claiming age is unavailable. The final sample contains 4,179 individuals, with a mean Social Security claiming age of 63.1. Slightly over half the sample claims at age 62, and 94% claim at 65 or earlier. Additional details on sample construction are in Appendix C.1.

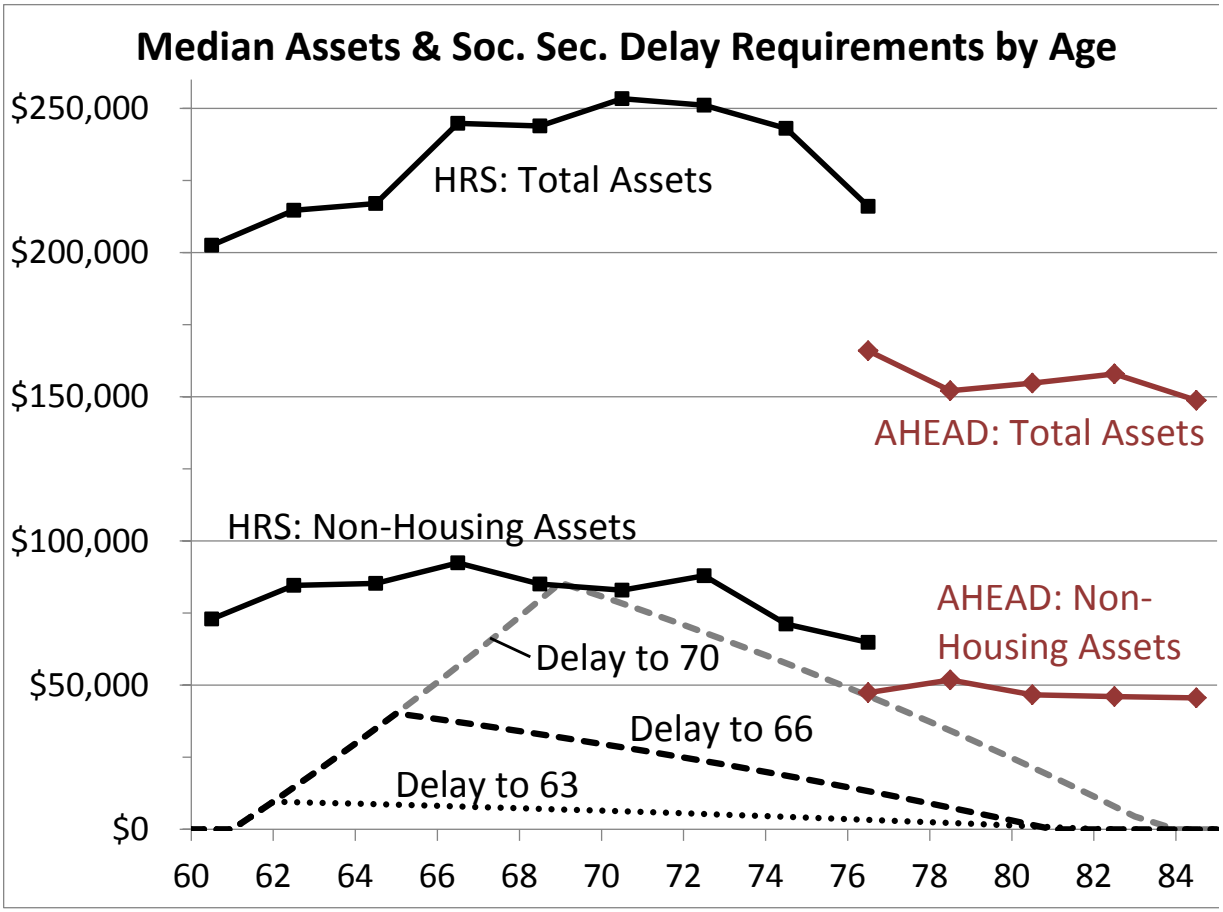
Figure 3.2 plots the sample's actual distribution of marginal Social Security returns had they delayed an additional year.[13] Because most people claimed at 62, the median real return is 8.1%, and even the 5th percentile is 5.0%. Less than 0.5% of people -- all of whom delayed until the maximum age of 70 -- had returns below a conventional interest rate of 3%. Therefore, the basic life cycle model can only reconcile the early observed claiming if the vast majority of retirees are liquidity constrained.

Figure 3.3 provides evidence that this is not the case. The dashed and dotted lines plot the asset requirements in condition (b) for delaying from 62 to various ages, using the HRS sample's average annual benefit of $9,340. The solid black lines show median assets by age (in year 2000 dollars) for the HRS sample. To examine assets at older ages than available from the HRS sample, the red lines plot data from the AHEAD survey (a close cousin of the HRS) for people born in 1917-1923. Here and elsewhere, I consider two measures of assets: (1) "total assets," a relatively comprehensive measure including all non-annuitized wealth except defined contribution pension balances,[14] and (2) "non-housing assets," a more conservative measure excluding housing and vehicle wealth. Economists have debated which measure is more appropriate for analyzing retirement savings, so I will show results for both.

[13] This calculation again treats individuals as single, ignoring couples' incentives. A future draft will take these incentives into account.

[14] Balances in defined contribution pension accounts are not available in the RAND version of the HRS, though any 401(k) balances that are rolled over into an IRA are measured. I am in the process of adding pension wealth measured in the core HRS survey to my data. Survey evidence from the HRS indicates that about a third of people separating from their jobs at age 55 or later leave 401(k) balances in place, while the remaining two-thirds roll them over, withdraw, or annuitize them (Johnson, Burman, and Kobes 2004).

Figure 3.3 Median Assets in HRS Sample vs. Assets Needed for Social Security Delay



NOTE: The graph compares median observed assets by age for two cohorts of older Americans against the minimum asset requirements in Proposition 1, calculated for delaying from 62 to various ages using the HRS sample's average Social Security benefit of $9,340 and rules for the cohort born from 1943-1954. All figures are inflation-adjusted to 2000 dollars. The key thing to note is that median assets (even excluding housing) are much higher than the levels needed for an average beneficiary to delay Social Security several years past 62. Each asset point represents the median real assets over pairs of ages (60-61, 62-63, etc.) for the associated sample. The HRS sample is the main sample (see Appendix C.1 for more information) comprised of individuals born 1931-1938. The AHEAD sample is comprised of all survey members born 1917-1923, excluding those who received Disability Insurance or claimed Social Security before 62 (to match the HRS sample exclusions). No corrections are made for differential attrition due to mortality, so the asset levels are representative only of surviving cohort members.

Figure 3.3 reveals two facts. First, the minimum asset levels implied by condition (b) are modest relative to elderly wealth – even excluding housing, which represents over half of a typical retiree's assets. Delaying for the minimum period of one month (not shown) would never require more than $800 (one month's benefits) in assets, and delaying to 63 would only require $9,600. Even delaying from 62 to

66 would reduce assets by a maximum of $40,200 at age 66. Non-housing assets in the HRS sample were more than twice as high for a typical 66-67 year old, and total assets were six times as high. Second, the asset requirements rapidly decline to zero after peaking just before the delayed claiming age. By contrast, observed assets decline more slowly with age.[15] Any theory which attempts to explain early Social Security claiming with impatience and liquidity constraints will have difficulty accounting for this post-retirement asset profile.

While these median asset levels are suggestive, Proposition 1 can be tested directly at the individual level using the panel dimension of the HRS. For an individual observed to claim at s (e.g., 62) and receive benefit b_s, I test whether *every available asset observation* falls above the individual-specific levels in condition (b) for delaying to age $s+k$. An individual for whom this is true could literally have executed the arbitrage-like transaction shown in Figure 3.1. This test is conservative because measurement error will cause more individuals to fail the test spuriously because of a single low observation than to pass the test incorrectly.

Table 3.2 shows the results for additional delays of 1 month (the minimum allowed), 1 year, 3 years, and all the way to age 70. Results are quite similar in the full sample and the sample that claimed at age 62. Regardless of the asset measure, a large majority had enough to delay at least an additional month – 90% using total assets and 70% using non-housing assets. The remaining 10-30% of people are liquidity constrained at some point in the years after retirement, and while delaying may still be optimal, the theory is less unambiguous. Many individuals appear able to afford much longer additional delays. Using total (non-housing) assets, 82% (52%) have enough assets to delay an additional three years, and 73% (42%) can afford to delay all the way to 70. One shortcoming of this test is that asset histories may be truncated due to early death or the final survey wave in 2008. To ensure sufficient observations, I have excluded individuals if assets are not observed at least once within two years around both the actual and alternate claiming age. In addition, the row labeled "Data Available through Age 75" further restricts the sample to

[15] Though Figure 3.3 does not correct for bias due to higher mortality among lower-asset individuals, restricting the sample to those who survive through age 76-77 shows an asset decline that is similarly gradual.

Table 3.2 Test of Proposition 1: People with Sufficient Assets for Delayed Claiming

Fraction of HRS Sample with Sufficient Assets for Additional Social Security Delays

Sample	Additional Delay Past Observed Claiming Age:			
	1 month	**1 Year**	**3 Years**	**To Age 70**
	Measure: Total Assets			
Full Sample	90.4% (0.5%)	86.8% (0.6%)	82.3% (0.7%)	73.1% (0.8%)
Claim at 62	90.6% (0.7%)	87.2% (0.8%)	83.3% (0.9%)	71.1% (1.2%)
Data Available through Age 75	90.5% (1.0%)	86.9% (1.2%)	81.2% (1.4%)	71.5% (1.6%)
	Measure: Non-Housing Assets			
Full Sample	71.2% (0.8%)	61.0% (0.9%)	52.4% (1.0%)	42.3% (1.0%)
Claim at 62	70.2% (1.1%)	61.4% (1.2%)	53.0% (1.3%)	38.3% (1.3%)
Data Available through Age 75	68.3% (1.7%)	60.1% (1.8%)	49.5% (1.9%)	40.2% (1.9%)

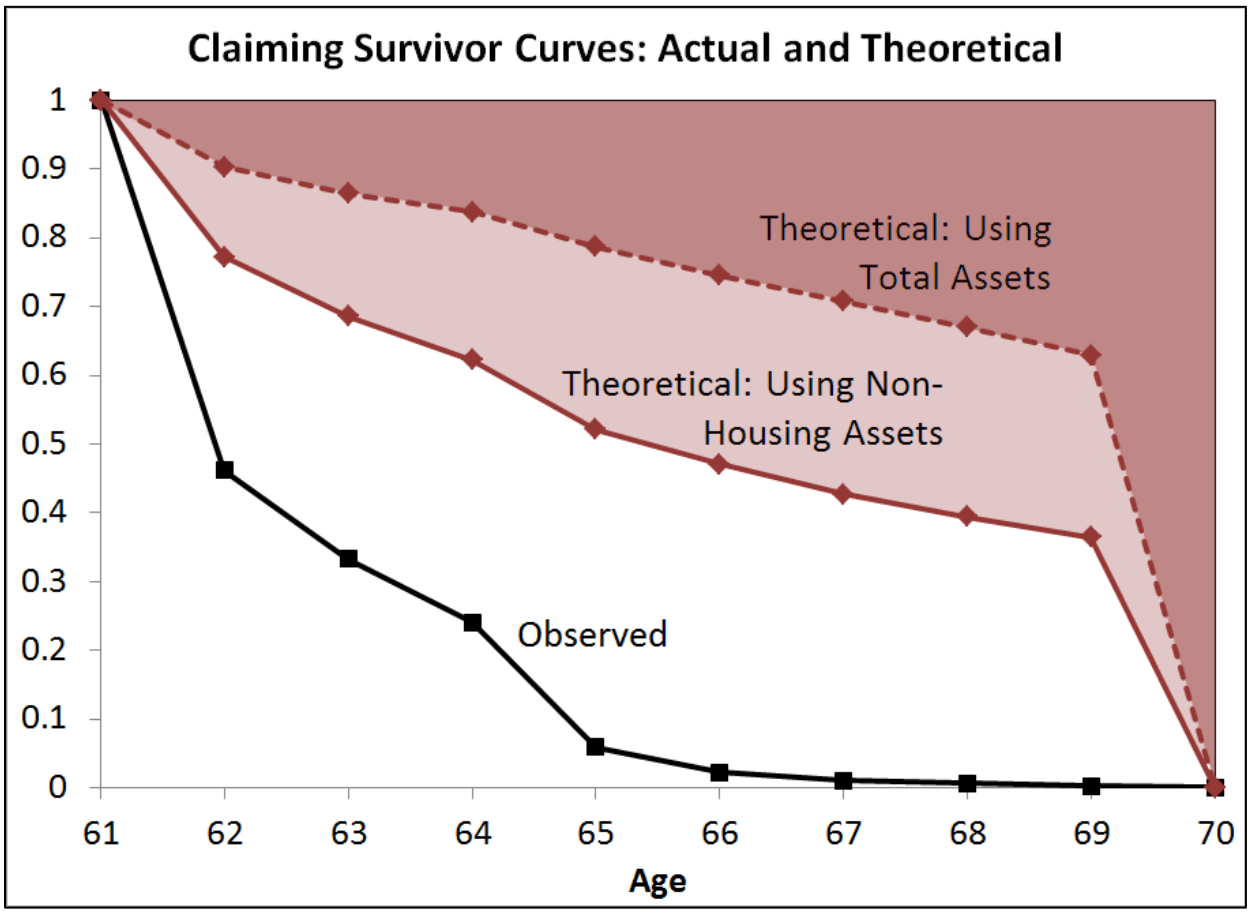
NOTE: Standard errors, clustered at the household level, are in parentheses. This table displays the fraction of people in the HRS sample who had sufficient assets, based on the conditions in Proposition 1, to delay claiming Social Security beyond the age actually observed – by 1month, 1 year, 3 years, and up to age 70. I use two measures of assets: total assets, a comprehensive measure of all non-annuitized wealth, and non-housing assets, which excludes housing and vehicle wealth. An individual satisfies the conditions of Proposition 1 if she has assets above the age-specific threshold for *every observation* in the data. Individuals are excluded from the test if assets are not observed at least once within two years around both the actual and alternate claiming age, or if their Social Security benefit level is missing. The "Full Sample" line includes all remaining individuals in the sample, while the "Claim at 62" line only includes those who claimed Social Security at age 62. To address the possibility that truncation in asset histories creates bias, the final line restricts the sample to 917 individuals for whom assets are available through at least age 75. The results are quite similar, suggesting there is little bias.

those whose asset histories extend to at least 75. The nearly identical results suggest that truncated histories do not create substantial bias.

Figure 3.4 illustrates the magnitude by which observed claiming violates the theory underlying Proposition 1. The black line plots the sample's observed survivor curve for Social Security claiming (the

fraction who delay past a given age). The red lines show the survivor curves after reassigning individuals to the highest age for which they pass the test of Proposition 1 using non-housing and total assets. Because Proposition 1 provides sufficient conditions for the optimality of delay, the theory predicts that the claiming curve should lie above the red lines. Instead, the sample's average claiming age is 63.1, while the theoretical average age using total (non-housing) assets is 68.6 (66.6). The largest discrepancies are for delaying past 65, which theory suggests should be common but is rare in reality.

Figure 3.4 Social Security Delayed Claiming: Actual and Theoretical



NOTE: This figure shows actual and theoretical survivor curves for Social Security claiming, defined as the fraction of individuals who delay past a given age. The black line plots the HRS sample's observed claiming patterns. The red lines show the survivor curves after reassigning individuals to the higher of their observed claiming age and the oldest age for which they pass the delay optimality test in Proposition 1. The solid red lines use non-housing assets for the test, and the dashed lines use total assets. If there is not enough asset data to conduct the test for a given delayed claiming age (usually due to early death), the individual is treated as failing the test. Therefore, the fractions who can delay past a given year in this graph are slightly lower than the corresponding numbers in Table 3.2.

3.2.2 Welfare Implications of Early Claiming

If individuals had delayed until the theoretical lower bounds in Figure 3.4 for total (non-housing) assets, the average annual Social Security benefit would have been $12,732 ($11,503), an increase of 36% (23%) over the observed $9,340 annual benefit. The value of these foregone arbitrage opportunities under the life cycle model are substantial. Past work has found that retirees could achieve welfare gains equal to 20-30% of total wealth by fully annuitizing at actuarially fair rates (Mitchell et al. 1999). These valuations typically use a fully specified life cycle model with a host of functional form and parameter assumptions. However, if the assumptions of no bequest motives or stochastic liquidity shocks (Assumptions 1 and 2 above) are maintained, it is possible to place a lower bound this valuation under weaker assumptions.

Specifically, assume that retirees have chosen intertemporal consumption to maximize utility, with indirect utility $V = \max_{\{c_t\}} U(c_s, c_{s+1}, ...)$. Define T as the first age at which savings fall to zero (so $a_{T+1} = 0$ and $a_t > 0 \ \ \forall t \le T$). Bernheim (1987) shows that T is a key parameter for valuing annuities. Up to T, intertemporal optimization implies that the expected discounted marginal utility of wealth is equal at every age: $U_{c_t} = R^{s-t} \cdot U_{c_s}$ for any t and s.[16] Therefore, optimization ensures that incremental annuity payments up to T are worth their present discounted value, regardless of discount or mortality rates. Applying this result, it is easy to show that the immediate money value of a (marginal) Social Security delay from age s to age $s+k$ equals to a first-order approximation:

$$\frac{\Delta V}{U_{c_s}} \approx \left(\sum_{t=s+k}^{T} R^{s-t} \cdot R_{SS}^{s,s+k} - \sum_{t=s}^{T} R^{s-t} \right) \cdot b_s + \sum_{t=T+1}^{\infty} \left(R_{SS}^{s,s+k} - 1 \right) \left(\frac{U_{c_t}}{U_{c_s}} \right) \cdot b_s \qquad (3.1)$$

The first term is the change in present value of benefits up to the age of wealth exhaustion. Because the Social Security return $R_{SS}^{s,s+k}$ exceeds the compounded interest rate R^k, it is simple to show that this term will be positive for T sufficiently large. The minimum T for which this is true is sometimes called the

[16] This is equivalent to stating that Euler equations hold as long as individuals are unconstrained. In the standard model with $U(c_s, c_{s+1}, ...) = \sum_t \beta^{t-s} u(c_t)$, the equation $U_{c_t} = R^{s-t} \cdot U_{c_s}$ can be written as $u'(c_t) = (\beta R)^{s-t} u'(c_s)$, which is the standard formulation of the Euler equation.

"break-even age" of an annuity. In Figure 3.3, the break-even ages are the first ages at which the asset requirements (from condition (b) of Proposition 1) fall to zero – typically around age 80, with some variation depending on the specific interval of delay. Equation (3.1) provides further intuition for the result in Proposition 1. For a small enough Social Security delay, the asset exhaustion age T will be unaffected. As long as T exceeds the break-even age, the first term in (3.1) will be positive. Because the second term – the utility value of additional benefits after T – is nonnegative, the total value of delaying Social Security is guaranteed to be positive. Based on this logic, the first term is a lower bound on the value of delaying Social Security, which can be used to analyze welfare.

Everything in the first term of (3.1) is in theory observable. A key statistic for annuity valuation on which there is little empirical evidence is the age of wealth exhaustion, T. The population distribution of T could be estimated by examining wealth trajectories (making sure to account for selective mortality) or by surveying retirees. I will not perform that exercise in this draft but will simply calibrate this value to show the approximate magnitude. For instance, let $T = 90$ and $R = 1.03$. Consider an individual with a typical benefit of $b_{62} = \$10,000$. Using the first term of (3.1) as a lower bound, the value of delay from 62 to 63 is at least \$5,600 for the 1937 and earlier cohorts (for whom $R^{SS}_{62,63} = 1.083$). Delaying from 62 to 67 – the average delay age-62 claimers could have afforded using non-housing assets – would have been worth \$14,900 for the 1937 cohort $(R^{SS}_{62,67} = 1.413)$. If assets were positive only to age 80, these lower bounds would be \$1,500 for 62 to 63 and -\$5,800 for 62 to 67 (i.e., not guaranteed to be worth it), while if assets were exhausted at 100, the values would be \$8,700 and \$30,200. These lower bounds are measures of the substantial value being "left on the table" from early claiming if the standard life cycle model describes preferences and welfare.

3.2.3 Information and Beliefs about Social Security

The prevalence of early claiming raises the question of how well retirees understand Social Security. The nature of claiming rules out the possibility, common in other contexts, that early claimers are passively

adopting the default choice. To start receiving Social Security, beneficiaries must actively apply for benefits. Retirees who procrastinate delay claiming by default.

Whether this timing decision is well-informed is less clear. Most people appear to know the basic facts about how delayed claiming increases benefits. A recent survey of non-elderly Americans found that 75% of respondents understood that claiming can be delayed past retirement (Greenwald, et al. 2010). Another survey found that near-retirees estimate fairly well the return to working longer and delaying benefit claiming (Liebman and Luttmer 2012). There is mixed evidence about whether claiming responds to marginal incentives. Coile et al. (2002) find that cross-sectional variation in benefit claiming broadly conforms with incentives, but the associations are weak. By contrast, Benítez-Silva and Yin (2009) find essentially no increase in delayed benefit claiming past the Full Retirement Age over the 1994-2004 period, despite a more than 50% increase in the incentive to do so. The one rule beneficiaries clearly respond to is the earliest age they can access benefits: claiming shifted rapidly to the earliest age allowed both after the earliest eligibility age was decreased from 65 to 62 in 1961 (Diamond and Orszag 2004) and after workers were allowed to claim full benefits at age 65 in 2000 (Song and Manchester 2007).

Treatments to increase information have also had little effect. People can learn about their claiming incentives through an individualized Social Security statement mailed annually to Social Security-covered workers and retirees. After the introduction of the statement in the late 1990s, near-retiree HRS respondents became much better informed about their Social Security benefit levels, but this information had no effect on retirement plans or claiming behavior (Mastrobuoni 2011). Evidence from a randomized field experiment teaching near-retirees about their Social Security incentives found no significant changes in claiming timing, despite a significant increase in work among females (Liebman and Luttmer 2015).

Even if beneficiaries know the Social Security rules, they may not believe the government will follow through with them. For instance, a 62-year-old who believes Social Security benefits may be cut in the next few years may choose to "get his money out now" rather than risking a benefit reduction if he delays. This sentiment is intuitive and consistent with decades of Social Security Trustee reports that the

program is not actuarially solvent over 75 years. For two sets of reasons, however, worries about benefit reductions are unlikely to be causing early claiming.[17] First, the logic underlying the "get my money out now" sentiment is questionable. For cuts to reduce the return to delay, they would have to apply to *current seniors* over 62, whereas nearly all reform plans exempt anyone over 55 at the time of reform. Further, to fully offset the high returns to delay, the cuts would have to substantially favor seniors who had already claimed over those of the same age who had delayed retirement or claiming. Given the longstanding goal of actuarial neutrality towards the retirement and claiming decisions, it seems unlikely policymakers would design cuts this way.

The second set of reasons is empirical. In 1992 and 1996, HRS retirees were asked to rate "the chances that congress will change Social Security so that it becomes less generous than now." Consistent with the perception of impending cuts, the average rating was about 60% for my sample in both years.[18] However, this rating has little or no ability to explain actual claiming patterns. The raw correlation of the reported cut probability with claiming age is actually significantly positive (though small), suggesting that those expecting a cut claim later than others. But after controlling for education and gender, this correlation disappears. Individuals who report a chance of a cut of 80-100% claim at almost exactly the same age (an insignificant 0.02 years later) as those who report an 0-20% chance, and the confidence interval rules out a difference less than -0.15 years. Early claiming, therefore, cannot be explained empirically by this proxy for retiree beliefs about future Social Security reductions.

3.3 Claiming with Bequest and Precautionary Motives

Consider again the basic intuition for annuitization shown in the life cycle plan of Figure 3.1. Delaying Social Security allows for a substantial increase in consumption with no offsetting decrease. Notice,

[17] For an alternate perspective arguing that benefit expectations can explain early claiming, see Benitez-Silva, Dwyer, and Sanderson (2006).

[18] Similar questions were also asked in 2006 and 2008, well after most of my sample had claimed, with an average answer of 58% in both years. In addition, a clarifying question about the chance that "these Social Security changes might affect your own benefits" was asked, to which the average answer was about 40% in each year.

however, that the most salient aspect of the new plan is how much assets must fall early in retirement. To delay to 65 without cutting consumption, the retiree must spend down three years of benefits -- about $30,000 for a typical person, or more than a third of median non-housing assets at ages 62-63.

It is not hard to imagine that a real-world retiree would find this plan disconcerting. In reality, assets are useful not just for saving for old-age consumption but also as a buffer stock for liquidity needs and as bequests for heirs. Because a larger annuity through Social Security is illiquid and cannot be left to heirs (except to a limited extent discussed below), adding bequest and precautionary savings motives to the basic life cycle model may be able to rationalize early claiming. Though bequest and precautionary motives are conceptually different, their theory shares much in common. Both are stochastic events in which conventional assets are more valuable than an actuarially equivalent annuity. In this section, I analyze their common theory and use its predictions to test whether bequest or precautionary motives can reverse the optimality of delayed claiming.

3.3.1 Theory

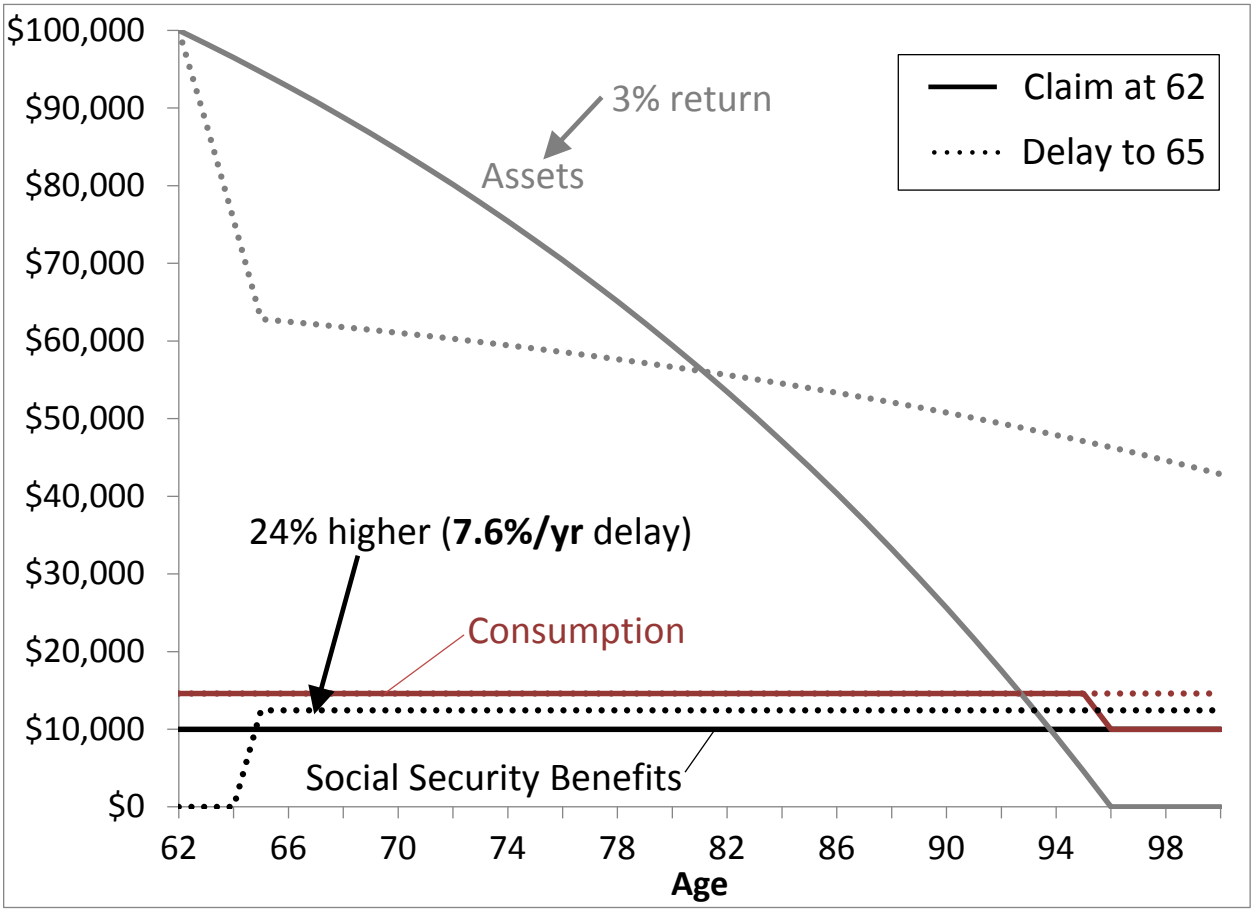
Consider a 62-year-old retiree's choice between claiming Social Security right away versus delaying until a later age. I drop Assumptions 1 and 2 from Section 3.2 by allowing the retiree to have bequest motives and to face stochastic liquidity shocks. Importantly, these liquidity shocks must be severe enough to effectively bankrupt the retiree: if assets will never fall below the levels in Proposition 1, then the same arbitrage logic applies.

I again consider a perturbation argument in which the retiree delays claiming, spends down assets to avoid a fall in consumption, and uses the incremental benefit to rebuild assets thereafter. The intuition for the argument is shown in Figure 3.5. The solid lines show the retiree's path for assets, consumption, and benefits conditional on claiming at 62 and not dying or experiencing a liquidity shock, and the dotted lines show the same for delaying to age 65.[19] In order to maintain consumption while delaying claiming,

[19] The formal perturbation argument below is based on a short delay (e.g., 1 month), so that first-order approximations apply. I depict delaying to 65 in the figure to make the differences more visible, but the same logic applies.

the retiree spends down assets more quickly early in retirement. Because death or a liquidity shock may occur at any time, this fall in assets is risky ex-ante (at age 62), even if it appears feasible ex-post for the majority of retirees who did not die or experience a shock. However, this risk is compensated by a potential benefit: assets available should death or a shock occur at older ages are higher with delayed

Figure 3.5 Social Security Delay Example #2: Insurance against Late-Life Risks



NOTE: This graph illustrates how delaying Social Security can provide insurance for bequests and life cycle risks in practice for a hypothetical retiree with a $10,000 annual benefit at 62, Social Security rules for the 1943-1954 birth cohorts, and a real return on assets of 3%. The solid lines depict the asset, consumption and Social Security benefit path if the retiree claims Social Security at 62 and amortizes assets to age 95. Mortality is uncertain and any remaining assets at death are left as a bequest. The alternate asset, consumption, and benefit path from delaying claiming to 65 are depicted in dotted lines. Instead of increasing consumption at the break-even age of 81, the retiree saves the incremental Social Security benefits, which raises assets substantially in later years. Because this transaction flattens the path of assets with aging, it provides insurance for bequests or other risks. This is particularly valuable because most risks tend to be concentrated in later years when assets are increased by the transaction.

claiming. Effectively, delayed claiming shifts assets from immediately after retirement to ages after 81 (the "break-even age" in this case), without having to reduce consumption. Annuitization therefore provides net insurance against late-life risks, at the expense of reducing insurance against risks more likely to occur earlier on. Whether delay is still optimal depends on the probability distribution of these shocks over time, as well as the relative severity of the shocks that occur earlier and later on.

Formalizing this argument requires moving away from the arbitrary utility function of Section 3.2. It is easy to think of non-standard preferences that would justify early claiming – for instance, a retiree might care only about bequests if he dies before age 80 but not thereafter. But I would like to test a more standard model first. I do so in the following framework. Suppose that an individual receives no utility from assets, except in the event of a single risky state (either death or a liquidity shock). Define q_t as the unconditional probability that the first occurrence of this risk occurs at time *t*. For instance, for mortality, q_t would be the probability that someone will be exactly *t* years old at death. If the risk occurs, the problem is truncated and the individual receives separable utility $v\left(Ra_t - \xi_t, b\right)$ from assets Ra_t net of the liquidity shock ξ_t and potentially also from Social Security benefits *b*. This function can be thought of as a reduced form for either bequest utility (for mortality risk) or for the value of consuming all one's remaining assets (for a liquidity shock). I assume that in both cases, $v(.)$ is increasing and weakly concave in assets.[20] The following condition captures these assumptions:

Assumption 3: (Expected Utility over Asset Needs) Utility as of period t_0 takes the form:

$$U_{t_0} = U\left(c_{t_0}, c_{t_0+1}, \ldots\right) + \sum_{t=t_0}^{\infty} \beta^{t-t_0} q_t v\left(Ra_t - \xi_t, b\right)$$

where $U(.)$ is an arbitrary function of consumption before the shock, q_t is the probability that the first instance of the risk occurs at time *t*, and $v_{a'}(.) > 0$, $v_{a''}(.) \leq 0$.

[20] For liquidity shocks, this assumption follows from concavity of utility of consumption, since $v\left(Ra_t - \xi_t, b\right) = u\left(Ra_t - \xi_t + b\right) + \beta E_t\left[V_{t+1}\left(a_{t+1} = 0, b\right)\right]$, where $V_{t+1}\left(a, b\right)$ is the value function.

Importantly, aside from utility discounting by β, I assume that the function $v(.)$ valuing assets net of the liquidity shock in the risky state is stable over time. This is equivalent to assuming a stable bequest utility function or for liquidity shocks, a stable period utility function $u(.)$ (since $v(.)$ is the utility of consuming all of one's resources immediately) – both standard assumptions. For a non-fatal liquidity shock to fit into this framework, it must be significant enough that the retiree wishes to consume all remaining assets, creating a binding liquidity constraint. If the shock does not reduce assets below the levels in Proposition 1 (which are essentially zero for a one-month delay), the same theory carries over. This restriction essentially rules out asset price fluctuations, whose costs are proportional to asset holdings, and ordinary medical costs, which are insured by Medicare and supplemental policies held by the vast majority of Medicare beneficiaries. The only major uninsured health care costs among the elderly are for nursing homes and other long-term care, which Medicare and supplemental policies do not cover. Therefore, I will focus on long-term care needs as the prime example of a liquidity shock.

Given this expected utility setup, the following result establishes testable conditions on risk probabilities and assets under which Social Security delay is optimal:

Proposition 2: Consider the possibility of delaying Social Security from age *s* to $s+1$, where the period length is small. Suppose that preferences satisfy Assumption 3. Then the following are sufficient for delay to be optimal:

(a) Delay Improves the Correlation of Payouts with Risks:

$$\sum_{t=s}^{\infty} \beta^{t-s} q_t \cdot \Delta Ben_t \geq 0$$

where $\Delta Ben_t = \left(\sum_{k=s+1}^{t} R^{t-k} b_{s+1} - \sum_{k=s}^{t} R^{t-k} b_s \right)$ is the change in the accumulated value of benefits received through period *t* by delaying claiming.

(b) Declining Assets with Age:

$$\bar{A}_{early} \geq \bar{A}_{late}$$

where $\bar{A}_p \equiv R \cdot \bar{a}_p - \bar{\xi}_p$, $p \in \{early, late\}$ are the levels of assets net of liquidity shocks that match the average marginal utility of assets before and after the break-even age (T^{BE}) such that:

$$v'\left(\bar{A}_{early}\right) = \sum_{t<T^{BE}} \left(\frac{\omega_t}{\sum_{\tau<T^{BE}} (\omega_\tau)} \right) E\left[v_{a'}\left(Ra_t - \xi_t, b \right) \right]$$

where the weights are $\omega_t \equiv \beta^{t-s} q_t \Delta Ben_t$, and likewise for $\bar{A}_{late}$ with the summation over $t \geq T^{BE}$.

(c) Unconstrained assets: Planned assets absent a shock exceed the levels of condition (b) in Proposition 1.

Proof: I verify the feasibility and optimality of a transaction like the one in Figure 3.5. See the appendix.

When these conditions are satisfied, delaying Social Security and saving the incremental benefits provides net insurance value for the risk in question. This formulation avoids the need to measure risk aversion or the strength of bequest motives, which would be necessary to evaluate a tradeoff between consumption and risk protection. Instead, the argument is that *regardless* of how much individuals care about the risk, delaying Social Security can help them be more prepared for it without any consumption cost.

Condition (a) formalizes the insight of Davidoff, Brown, and Diamond (2005) that annuities are better than conventional assets for saving for late-life shocks, but worse for early-life shocks. The key technical requirement is that delay increases the correlation of accumulated benefits with the risk timing. Because delay shifts assets downward initially but upward after the break-even age, risks sufficiently concentrated after the break-even age will satisfy this condition. A useful benchmark is mortality (for bequest motives). When the discount rate equals the interest rate, condition (a) is equivalent to a

requirement that the returns be actuarially fair or higher (see the appendix for a proof). Condition (a) therefore can be understood as a generalization of actuarial fairness to an arbitrary risk. For any risk that is more concentrated at older ages than mortality, delay provides insurance value even at lower Social Security returns.[21]

Condition (b) is more difficult to test empirically because $v(.)$ is unknown. If asset decumulation were monotonic and liquidity shocks were increasing with age, (b) would hold for any concave function $v(.)$. But as Figure 3.3 shows, assets decline quite gradually and may even increase early in retirement. Therefore, accurately testing (b) would require observing how much assets subsequently decline in individuals' 70s, 80s, and 90s, necessitating a longer panel of asset observations than is available in the HRS. Further, an accurate test is challenging because HRS asset data often show implausibly large intertemporal fluctuations, potentially due to measurement error, although the exact reason is not known (see Poterba, Venti, and Wise 2011). For these reasons, I have not formally tested condition (b) but will explore testing it in future drafts.

3.3.2 Mortality Risk and Bequest Motives

The test of Proposition 2 is simplified in the case of bequest motives because mortality probabilities are widely available and there is no need to account for a liquidity shock. When the interest rate equals the discount rate, the mortality risk will be concentrated sufficiently late in life for delaying to satisfy condition (a) only if delaying is actuarially fair or better. Because the schedule was designed to be actuarially fair, one might expect that Social Security claiming would be close to neutral with respect to bequest motives.

In practice, delaying Social Security is typically better than actuarially fair on the margin, a fact that appears to have received little, if any, attention. The primary reason for this is historical. The schedule for delays between 62 and 65 was introduced for women in 1956 and men in 1961, and it was

[21] Note that it is only the relative concentration of the risk timing (earlier versus later), not the level of the risk, that matters, since condition (a) is unaffected by scaling all values of q_t by a constant.

close to actuarially fair at the time. But in the 50 years since, the schedule has been left essentially unchanged despite historic gains in elderly longevity.[22] Table 3.3 documents these trends, showing the money's worth of the Social Security delay annuity for both genders in several cohorts and ages. Money's worth is a standard statistic for annuity valuation equal to the expected present value of payouts per dollar of premium. Money's worth also equals the annuity return (in my notation, $R_{SS}-1$) divided by the fair return ($R_{fair}-1$), so a value above 1.0 indicates better than fair. The statutory return of 8.3% for delaying from 62 to 63 had a money's worth of 1.043 for an average individual born in 1900 -- just slightly above fair. But by the 1930 cohort, the same 8.3% rate was almost 15% better than fair. By the 1950 cohort who are retiring today, legal reforms had lowered the return for delaying from 62 to 63, but the now 8.3% return to delaying from 63 to 64 was 18% better than fair on average. Thus, these Social Security money's worth exceed actuarial fairness by about the same amount that commercial annuities fall short of it (Mitchell et al. 1999).

Table 3.3 shows two more reasons delay is often better than fair. First, Social Security delay is effectively a community rated annuity, despite large heterogeneity in expected mortality. As a result, delay was already better than fair for female workers born in 1900 and only became more so over time. A similar point applies to healthier and highly educated retirees, who are also less likely to be prevented from delaying by liquidity constraints. Second, the returns to delay are highest at younger ages (i.e., just after 62) because Social Security uses a linear benefit schedule. Each month of delay raises the benefit by a specified fraction of the "full benefit," the benefit available at the full retirement age. This linear growth implies that the increase *as a proportion of the foregone benefit* – the key statistic for a money's worth calculation – is higher in earlier years. By claiming at age 62, most beneficiaries are turning down a marginal unit of annuity that is the most generous available.

[22] Between 1961 and 1999, the actuarial schedule between 62 and 65 was not adjusted at all. Between 2000 and 2005, the return to delaying between 62 and 63 fell (as part of the increase in the full retirement age to 66), but the return to delaying from 63 to 65 increased, leaving unchanged the average return to delaying from 62 to 65.

Table 3.3 Money's Worth of Social Security Delay Annuity

Money's Worth of Social Security Delay Annuity

Ages of Delay:	Soc. Sec. Return	Money's Worth of Delay Annuity		
		Average Person	Males	Females
1900 Birth Cohort				
62 to 63	8.3%	1.043	0.911	1.154
63 to 64	7.7%	0.936	0.815	1.037
64 to 65	7.1%	0.845	0.733	0.937
65 to 70 (avg)	0.0%	0.000	0.000	0.000
1930 Birth Cohort				
62 to 63	8.3%	1.149	1.065	1.222
63 to 64	7.7%	1.031	0.954	1.097
64 to 65	7.1%	0.929	0.859	0.990
65 to 70 (avg)	4.1%	0.489	0.451	0.522
1940 Birth Cohort				
62 to 63	7.5%	1.074	1.012	1.132
63 to 64	8.0%	1.109	1.043	1.171
64 to 65	7.4%	0.997	0.935	1.053
65 to 70 (avg)	6.3%	0.773	0.721	0.819
1950 Birth Cohort				
62 to 63	6.7%	0.970	0.917	1.017
63 to 64	8.3%	1.179	1.113	1.238
64 to 65	7.7%	1.057	0.996	1.111
65 to 70 (avg)	7.2%	0.897	0.841	0.947

NOTE: This table shows the returns and money's worth of the annuity purchased by delaying Social Security claiming (without delaying retirement) for three cohorts. The Social Security return is the annual real return (or the average annual return for the ages 65-70 interval). The "money's worth" of an annuity equals the expected present value of the payouts, divided by the present value of the premium. In the Social Security context, the "payouts" are the increment in the benefits after the new claiming age, and the "premium" is the benefits foregone while delaying. Money's worth also equals the annuity return divided by the actuarially fair return, so values above 1.0 indicate returns that are better than fair. The expected present value calculation assumes a real interest rate of 3% and uses gender-specific cohort life tables estimated by the SSA (Bell and Miller 2005).

While these facts are suggestive, a formal test of condition (a) of Proposition 2 is needed to confirm whether delay provides insurance value against bequest risk. I implement this assuming a discount rate equal to the interest rate of 3%.[23] Table 3.4 shows the results for two sets of mortality probabilities (as well as for long-term care risk, discussed in the next subsection). The top panel uses gender and birth-year specific mortality from the SSA cohort life tables, as a measure of average mortality risk. The results for the sample who claimed at 62 are illustrative. Delay by an extra year to age 63 provides insurance value (satisfies condition (a)) for 100% of the sample, and delay to 64 is justified for all women and some men in the younger cohorts. Because the returns to delay fall with age, particularly after 65, delays to 66 or later do not satisfy condition (a). For the sample as a whole, these results imply that 64% of people – most of whom claimed at 62 or 63 – would have obtained bequest insurance by delaying by an additional year.

In sum, bequest motives can help explain why individuals do not delay all the way to 70. However, mortality is sufficiently concentrated in later years that the dominant behavior of claiming at 62-63 cannot be justified for the average person. It is also not true that only sicker retirees claim at these early ages. Even among people who report "excellent" or "very good" health in the interview closest to age 62, 52% claim at age 62. While poor health is correlated with earlier claiming, this effect appears to be entirely a mechanical effect due to earlier retirement. Conditioning on being retired before 62, earlier claiming is not significantly correlated with worse health.

One simple deviation from rational expectations that could explain earlier claiming would be if the population as a whole pessimistically believed they were likely to die early in retirement. The middle panel of Table 3.4 provides evidence against this by updating the test of condition (a) using individuals' own self-reported longevity expectations based on questions in the HRS. Specifically, I scale life table mortality rates to best match each individual's self-reported probability of living to 75, 80, and/or 85,

[23] Adjusting both the interest and discount rate upward or downward does not materially affect the results. A discount rate exceeding the interest rate could help explain early claiming but would be difficult to reconcile with the slow path of asset decumulation shown in Figure 3.3.

averaging over all available self-reports in interviews surrounding Social Security claiming to reduce noise (see the note to Table 3.4 for details). Longevity expectations vary widely across the population, but longevity beliefs are almost exactly right for the median male and only a couple years too short for the median female. Therefore, the results in Table 3.4 for self-reported mortality risks are qualitatively similar to those using the objective probabilities. About two-thirds of people claiming at 62 and about half of the full sample report beliefs suggesting that delaying an additional year would provide bequest insurance.

Therefore, although I have not yet implemented a formal test of condition (b) of Proposition 2, I conclude that bequest motives alone are unlikely to explain the dominant behavior of claiming Social Security at ages 62 and 63. I next turn to an analysis of liquidity shocks in retirement.

3.3.3 Stochastic Liquidity Needs in Retirement

For a liquidity risk to reverse the optimality of delayed claiming, it would have to satisfy two conditions. First, it would have to be severe enough to completely exhaust assets, creating a binding liquidity constraint. Second, it would have to be concentrated early in retirement so that condition (a) of Proposition 2 does not hold. As discussed above, the only obvious candidate for a severe shock is uninsured nursing home expenses. Because post-retirement income is relatively steady, most shocks other than medical expenses are either small or – in the case of asset price fluctuations -- proportional to asset holdings and therefore unlikely to cause bankruptcy. And essentially all major medical expenses other than nursing home costs are insured for U.S. retirees through Medicare and supplemental plans. Therefore, in this draft, I test the theory only for nursing home risks. In future work, I will examine the empirical pattern of ages at which post-retirement assets are exhausted and attempt to associate asset exhaustion events with other risks.

Table 3.4 Test of Proposition 2: Fraction for whom Delay Provides Insurance Value

Fraction of HRS Sample for Whom Delayed Claiming Provides Insurance Value

Sample	Change in Claiming Relative to Observed Age:			
	1 Year Later	**2 Years Later**	**3 Years Later**	**Delay to 66***
	Mortality Risk (Life Tables)			
Full Sample	63.7% (0.8%)	35.1% (0.7%)	9.7% (0.5%)	0.0% (0.0%)
Claim at 62	100.0% (0.0%)	59.4% (1.0%)	17.7% (0.8%)	0.0% (0.0%)
	Mortality Risk (Self-Reported)			
Full Sample	47.3% (0.9%)	33.3% (0.8%)	22.0% (0.7%)	11.8% (0.6%)
Claim at 62	64.5% (1.1%)	48.6% (1.1%)	35.7% (1.1%)	11.0% (0.7%)
	Long-Term Nursing Home Risk			
Full Sample	76.2% (0.7%)	67.2% (0.8%)	54.4% (0.9%)	0.0% (0.0%)
Claim at 62	100.0% (0.0%)	100.0% (0.0%)	100.0% (0.0%)	0.0% (0.0%)

* Restricted to individuals who claimed before age 66.

NOTE: The table shows results of a test of condition (a) in Proposition 2 for additional Social Security delays of 1, 2, or 3 years beyond the observed claiming age, or to age 66 (if applicable). When these conditions are satisfied, the theory underlying Proposition 2 implies that retirees obtain insurance value for bequests and long-term care needs by delaying Social Security. All calculations assume an interest and discount rate of 3%. The top panel uses objective cohort-gender-specific mortality probabilities from life tables projected by SSA actuaries (Bell and Miller 2005). The bottom panel uses objective probabilities of long-term nursing home use (exceeding 60 days), estimated from the HRS and AHEAD data (see the note to Figure 3.6 for a description of how these are estimated). The middle panel adjusts life-table survival curves to best match each individual's self-reported longevity expectations. These self-reports are noisy and sometimes inconsistent, so I take several steps to reduce noise. Specifically, for self-reported probabilities of living to 75, 80, and 85, I take an average of all available self-reports in interviews within two years before or after claiming Social Security. I then calculate the factor by which gender-specific life table mortality would have to be proportionally scaled to match these self-reports for 75, 80, and 85, and take a geometric mean of these factors. I use the resulting average factor for each individual to scale life table mortality probabilities, and use this distribution as an estimate of self-reported mortality.

As intuition for the test of condition (a), Figure 3.6 shows the probability distribution for the age at first long-term nursing home stay based on the HRS and AHEAD sample's experience, and for comparison, the distribution of age at death. For reference, the vertical line at age 78 marks the break-even age for delaying from 62 to 63, the relevant margin for most people.[24] The solid black line shows the distribution of age at death (conditional on being alive at 62) for the 1930 cohort, for whom delay from 62 to 63 was 15% better than fair. The solid gray curve shows how much earlier age at death would have to needed to be for delay to be exactly actuarially fair -- and condition (a) to hold with equality. This curve peaks about three years earlier than the true 1930 cohort distribution and corresponds to 47% higher annual mortality. The dashed red curve shows the distribution of the age of first long-term nursing home stay estimated using the actual experience of the HRS and AHEAD samples.[25] Notably, this distribution is much more concentrated in later years than the actuarially fair standard or even age at death. Therefore, nursing home risks are even less likely than mortality to reverse the optimality of delayed claiming.

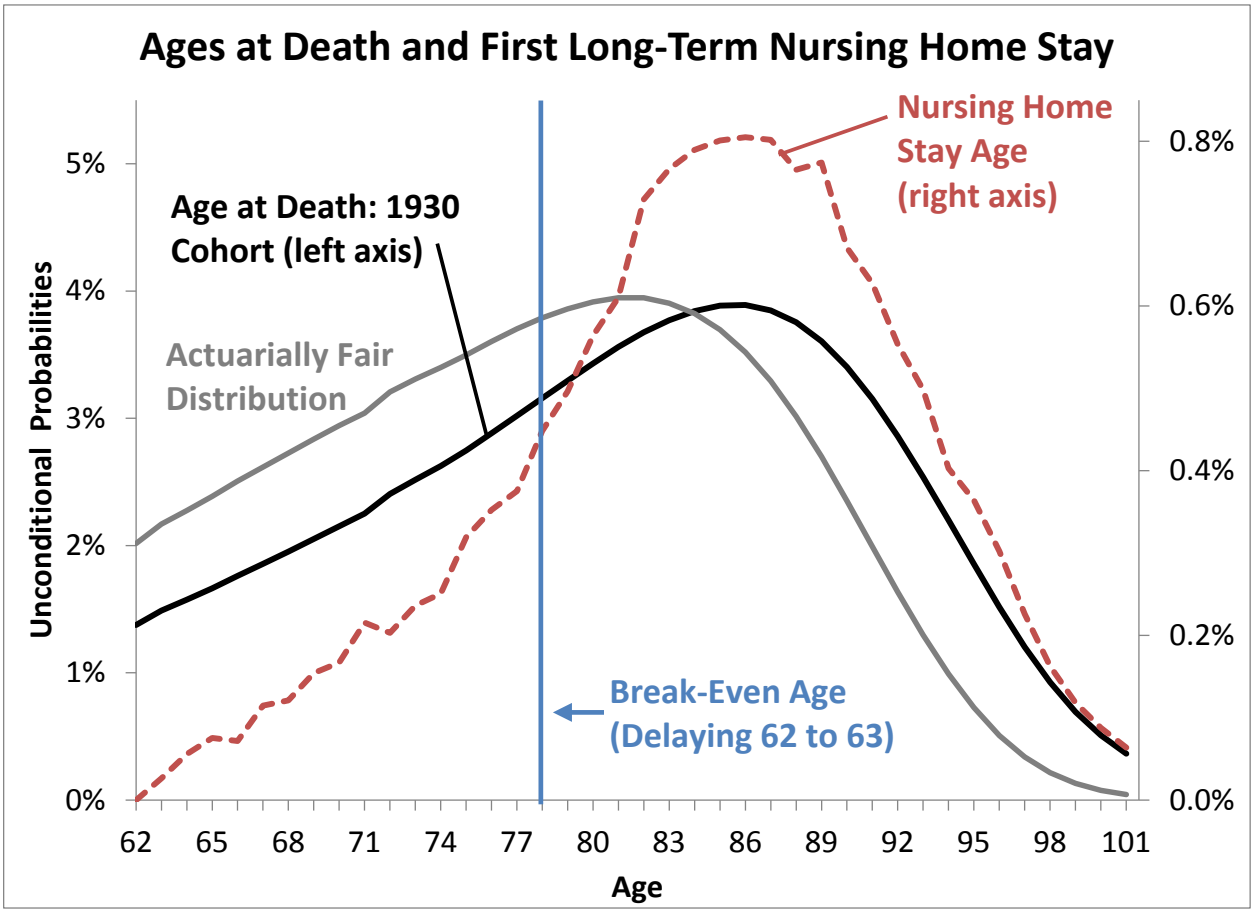
The bottom panel of Table 3.4 verifies this result by implementing the formal test of condition (a) using these nursing home risk probabilities. Because this distribution is concentrated at older ages than mortality, delay all the way to age 65 provides insurance value against the risk. As a result, additional delay is justified for everyone who claimed before 65, or 76% of the sample. Retirees worried about nursing home risks should therefore delay claiming until at least age 65 and save the incremental benefits to build a larger buffer stock at the older ages when nursing home needs are most likely. Alternatively, the higher benefits could be used to help pay for long-term care insurance. Though I do not implement a formal test of condition (b), I note that such a test would require accounting for how the cost of nursing

[24] The break-even age of 78 for delaying from 62 to 63 is lower than for delaying from 62 to 65 (shown in Figure 3.5) because the former increases benefits by 8.3%, while the latter increases benefits on average by only 7.6% per year.

[25] I define "long-term" as a nursing home stay longer than 60 days. Most shorter stays are for inpatient rehabilitation, which is covered by Medicare for up to 100 days, and do not represent the catastrophic expenses for which I am trying to proxy. Because not everyone has a long-term nursing stay, this distribution integrates to less than 1. See the note to Figure 3.6 for details on how I estimate this distribution from the HRS and AHEAD data.

home shocks (ξ_t) rises or falls with age. If the costs rise with age, condition (b) is more likely to be true, since assets are especially valuable later in life when the size of the risk is larger. But if costs fall with

Figure 3.6 Ages at Death and First Long-Term Nursing Home Stay



NOTE: This graph shows estimated probability distributions for age at death and age at first long-term nursing home use (both conditional on being after age 62). The solid black curve shows the distribution of age at death, using life table mortalities for the 1930 birth cohort, as estimated by SSA actuaries (Bell and Miller 2005). As a reference point, the solid gray curve shows this distribution with annual mortality scaled up proportionately (by 47%) until delay from 62 to 63 is exactly actuarially fair. The vertical line at 78 marks the break-even age (see text for the definition) for this delay. The dashed red curve shows the distribution of the age of first long-term nursing home stay (defined as a stay longer than 60 days) estimated using the actual experience of the HRS, AHEAD, and Children of the Depression (CODA) cohorts in the HRS. To estimate this distribution, I denote the first age, if any, that each sample member has a long-term nursing home stay. I input these into a Kaplan-Meier survival model to estimate a hazard rate of first nursing home use at each age, conditional on being alive. I use these hazard rates, along with life table mortality probabilities, to calculate the probability that a non-institutionalized 62 year old will first have a long-term nursing home use at each age. This distribution is somewhat noisy, so the curve shown is smoothed by taking a 5-year symmetric moving average. Since most people never have a long-term nursing home stay, these probabilities integrate to about 15%. Therefore, I have adjusted the right scale to be 1/6.5 (≈15%) as high as the left scale to make the curves visually comparable.

age, the reverse is true. A priori, it is not clear which of these is true. One additional consideration not yet covered is the value of higher Social Security benefits should the individual recover and exit the nursing home after having exhausted assets. This consideration would provide an extra reason for delaying Social Security.

3.4 Conclusions and Next Steps

To shed light on the longstanding annuity puzzle, this paper has studied U.S. retirees' annuity choices implicit in their timing of Social Security benefit claiming. While there has been debate in the literature about whether complexities or imperfections in commercial annuities can explain their unpopularity among the elderly, the nature of Social Security delay as a small actuarially fair real annuity allows me to essentially rule out this line of reasoning. I do this in two steps. First, I derive theoretical predictions about the optimality of a marginal delay in claiming based on future asset holdings. These predictions are robust because they are based on the logic of arbitrage. The prediction that liquidity unconstrained retirees will not claim at 62 are clearly rejected in the HRS data. This result strengthens the findings of Davidoff, Brown, and Diamond (2005) because it implies that the basic life cycle model cannot explain annuitization behavior regardless of the shape of intertemporal utility of consumption.

Second, I consider a richer life cycle model with either bequest motives or uninsured risks that create binding liquidity constraints. Because retirees' marginal delays are often better than actuarially fair, a simple perturbation argument suggests annuitization is still optimal. If retirees delay claiming, hold consumption fixed, and save all of the incremental benefits, they can increase the total amount of assets at advanced ages when death or liquidity shocks are most likely to occur. Thus, the annuity itself may provide insurance for bequest or long-term care risks. I derive conditions on risk timing under which annuitization is optimal by this logic. I show empirically that death and nursing home risk satisfy these conditions for most retirees' marginal decisions using either objective or self-reported risk probabilities, where available. If preferences take a forward-looking, expected utility form, then for these individuals,

delaying Social Security provides insurance value against both bequests and nursing home costs, the two major asset risks that the elderly face.

These two lines of reasoning jointly suggest that standard specifications of intertemporal preferences are inadequate to explain Social Security claiming patterns for many people. In the standard versions of the life cycle model, individuals would not both claim Social Security at 62 and hold onto substantial assets deep into retirement. Yet this is what a large fraction of retirees appear to be doing.

One of several "behavioral" explanations may be more promising for understanding early claiming. In future drafts, I intend to examine three such explanations. Here are their descriptions and how I will attempt to test them:

1. Retirees might be concerned about bequests and late-life risks but feel unable to save out of annuity income due to temptation costs. This story would argue that despite being myopic or hyperbolic discounters, retirees overcame their inability to save during working years by using commitment vehicles like 401(k) pensions, IRAs, and mental accounts (Shefrin and Thaler 1988). Although much of this wealth is liquid in retirement, mental accounting allows retirees to avoid spending it, thus maintaining it for bequests and late-life risks. However, were they to annuitize this wealth, they would have difficulty saving out of the monthly stream of income to re-accumulate a buffer stock for later years. This limitation would make the perturbation argument in Section 3.3 infeasible, potentially reversing its results.

2. Retirees may have non-standard preferences, such as a psychic sense of security or control from holding conventional assets (even if they are not consumed) but not from holding annuities (Brown (2007); see also Carroll (2002), Kaplow (2009)). These preferences are non-standard because they involve valuing assets for reasons other than their payouts. At its heart, this "utility of control" explanation boils down to retirees misunderstanding or mistrusting annuities and instead feeling safe with the liquid assets that they understand. This theory might also explain the failure to buy long-term care insurance if they mistrusted that insurance product as well.

3. Retirees may misunderstand how to value annuities, but otherwise be fully rational. In this case, relatively simple reframing could significantly increase Social Security delays. Recent lab experiments have provided some evidence for this framing explanation for Social Security claiming (Brown, Kapteyn, and Mitchell 2011) and annuities more generally (Brown et al. 2008).

The strategy for testing the first two theories is to examine how exogenous changes in people's annuity holdings or in their current annuity income, holding total wealth constant, affects their consumption and asset decumulation decisions. The third theory is more difficult to test without experimental evidence as in the papers cited above. There are two basic predictions that differ in the first two models:

1. *Effect of current annuity income on consumption*: The mental accounting model predicts a higher marginal propensity to consume out of current income (like annuities) than out of liquid wealth holdings (like conventional assets). Therefore, if current annuity income increases without a change in total wealth, retirees would be expected to increase consumption in the short run. Intuitively, they would spend more because they no longer have to bear the temptation cost of spending out of their mental account. By contrast, the utility of control model would predict a smooth path for consumption.

2. *Effect of annuity share of wealth on asset decumulation*: The mental accounting theory predicts that, controlling for total wealth, retirees with a larger share of their portfolio in annuitized form will consume more (rather than saving from the annuity income) and therefore accumulate a smaller liquid buffer stock for bequests and risks at older ages (e.g., 80+) when the risks are most likely to occur. By contrast, the utility of control model would predict substantial savings out of annuity income by over-annuitized retirees worried about their low level of liquid assets. Thus, retirees with a higher annuity share of wealth would end up with higher late-life non-annuitized asset holdings.

There are several sort of events I can use to generate a plausibly exogenous change in current annuity income or in annuity share of wealth. The most promising are exogenous changes in Social Security rules

that changed retirees' annuity holdings. The first is the introduction of early claiming in 1961, which unexpectedly allowed a cohort beneficiaries to access Social Security at 62 rather than 65. This change raised current annuity income for early-claiming 62-64 year olds in 1961 but lowered the long-term annuity share of their portfolios. I can use the Consumer Expenditure Survey of 1960-1961 and surveys of retirees' asset holdings in the 1960s and 1970s to test how consumption and savings responded. The second policy change was the unexpected removal of the earnings test in 2000 for current workers over 65. Most workers who had not yet claimed shifted their claiming ages forward in response to this change. As with the 1961 experiment, this change raised current annuity income but lowered long-term annuity share of wealth. The third policy change was the exogenously higher benefits given to the "notch generation" of Social Security retirees in the late 1970s. This change increased their total wealth and also their annuity share of wealth relative to cohorts immediately before and after them. I can test how much of this additional annuity wealth was saved by comparing their post-retirement asset decumulation to that of surrounding cohorts.

Bibliography

Abelson, R., Thomas, K., and McGinty, J. C. (2013). Health care law fails to lower prices for rural areas. *New York Times*, October 23, 2013.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, 488-500.

Ameriks, J., Caplin, A., Laufer, S., and Van Nieuwerburgh, S. 2011. "The Joy of Giving or Assisted Living? Using Strategic Surveys to Separate Public Care Aversion from Bequest Motives." *Journal of Finance* 66(2): 519–61.

Association of American Medical Colleges, Center for Workforce Studies. (2013) 2013 State Physician Workforce Data Book. Retrieved from https://www.aamc.org/download/362168/data/2013statephysicianworkforcedatabook.pdf

Bell, F.C., and Miller, M.L.. 2005. "Life Tables for the United States Social Security Area: 1900-2100." SSA Actuarial Study 120.

Benitez-Silva, H., D.S Dwyer, and W.C Sanderson. 2006. "A Dynamic Model of Retirement and Social Security Reform Expectations: A Solution to the New Early Retirement Puzzle." U. Michigan Retirement Research Center Working Paper 2006-134.

Benítez-Silva, H., and N. Yin. 2009. "An Empirical Study of the Effects of Social Security Reforms on Benefit Claiming Behavior and Receipt Using Public-Use Administrative Microdata." *Social Security Bulletin* 69(3): 77–95.

Bernheim, B.D. 1987. "The Economic Effects of Social Security." *Journal of Public Economics* 33(3): 273–304.

_____. 1991. "How Strong Are Bequest Motives? Evidence Based on Estimates of the Demand for Life Insurance and Annuities." *Journal of Political Economy* 99(5): 899.

Berry, S., Levinsohn, J., & Pakes, A. (2004). Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market.*Journal of Political Economy*, *112*(1), 68-105.

Brown, J., Duggan, M., Kuziemko, I., & Woolston, W. (2011). "How does risk selection respond to risk adjustment? Evidence from the Medicare Advantage program." NBER Working Paper No. 16977.

Brown, J.R. 2007. "Rational and Behavioral Perspectives on the Role of Annuities in Retirement Planning." NBER Working Paper 13537.

Brown, J.R, A. Kapteyn, E.F.P. Luttmer, and O.S Mitchell. 2013. "Cognitive Constraints on Valuing Annuities." NBER Working Paper 19168.

Brown, J.R, A. Kapteyn, and O.S Mitchell. 2011. "Framing Effects and Expected Social Security Claiming Behavior." NBER Working Paper 17018.

Brown, J.R., J.R. Kling, S. Mullainathan, and M.V. Wrobel. 2008. "Why Don't People Insure Late-Life Consumption? A Framing Explanation of the under-Annuitization Puzzle." *American Economic Review: Papers & Proceedings* 98(2): 304–9.

Bundorf, M. K., Levin, J., & Mahoney, N. (2012). Pricing and Welfare in Health Plan Choice. *American Economic Review*, *102*(7), 3214-48.

Cardon, J. H., & Hendel, I. (2001). Asymmetric information in health insurance: evidence from the National Medical Expenditure Survey. *RAND Journal of Economics*, 408-427.

Carroll, C.D. 2002. "Why Do the Rich Save So Much ?" In *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, ed. Joel Slemrod. Cambridge, MA: Harvard University Press, 465–84.

Capps, C., Dranove, D., & Satterthwaite, M. (2003). Competition and market power in option demand markets. *RAND Journal of Economics*, 737-763.

Chan, D. and Gruber, J. (2010). How sensitive are low income families to health plan prices? *The American Economic Review*, 100(2):292–296.

Chandra, A., Gruber, J., & McKnight, R. (2010). Patient cost-sharing and hospitalization offsets in the elderly. *The American economic review*, *100*(1), 193.

Chandra, A., Gruber, J., & McKnight, R. (2011). The importance of the individual mandate—evidence from Massachusetts. *New England Journal of Medicine*, *364*(4), 293-295.

Chandra, A., Gruber, J., & McKnight, R. (2014). The impact of patient cost-sharing on low-income populations: Evidence from Massachusetts. *Journal of health economics*, *33*, 57-66.

Chiappori, P. A., & Salanie, B. (2000). Testing for asymmetric information in insurance markets. *Journal of Political Economy*, *108*(1), 56-78.

Coile, C., P. Diamond, J. Gruber, and A. Jousten. 2002. "Delays in Claming Social Security Benefits." *Journal of Public Economics* 84: 357–85.

Commonwealth Care (2008). Report to the Massachusetts Legislature: Implementation of the health care reform law, chapter 58, 2006-2008.

Commonwealth of Massachusetts, Center for Health Information and Analysis (CHIA) (2013). Annual Report on the Massachusetts Health Care Market. Retrieved from http://www.mass.gov/chia/docs/r/pubs/13/ar-ma-health-care-market-2013.pdf

Commonwealth of Massachusetts, Center for Health Information and Analysis (CHIA) (2014). Massachusetts Hospital Profiles, Acute Hospitals Databook. Retrieved from http://www.mass.gov/chia/docs/r/hospital-profiles/2012/massachusetts-hospital-profiles-acute-databook-2.xlsx

Congressional Budget Office (2013). May 2013 estimate of the effects of the affordable care act on health insurance coverage.

Crawford, G. S., & Yurukoglu, A. (2012). The Welfare Effects of Bundling in Multichannel Television Markets. *American Economic Review*, *102*(2), 643-85.

Cutler, D. M., & Reber, S. (1998). Paying for Health Insurance: The Trade-off between Competition and Adverse Selection. *Quarterly Journal of Economics*,*113*(2).

Dafny, L. S. (2010). Are health insurance markets competitive? *American Economic Review*, 100(4):1399–1431.

Davidoff, T., J.R. Brown, and P.A. Diamond. 2005. "Annuities and Individual Welfare." *American Economic Review* 95(5): 1573–90.

Decarolis, F. (2013). Medicare Part D: Are insurers gaming the low income subsidy design? Working Paper.

De Nardi, M., E. French, and J.B. Jones. 2010. "Why Do the Elderly Save? The Role of Medical Expenses." *Journal of Political Economy* 118(1): 39–75.

Diamond, P., and P. Orszag. 2004. *Saving Social Security: A Balanced Approach*. (Washington, DC: Brookings Institution).

Dorfman, R., & Steiner, P. O. (1954). Optimal advertising and optimal quality.*The American Economic Review*, 826-836.

Dynan, K.E., J. Skinner, and S.P. Zeldes. 2002. "The Importance of Bequests and Life-Cycle Saving in Capital Accumulation: A New Answer." *American Economic Review: Papers & Proceedings* 92(2): 274–78.

_____. 2004. "Do the Rich Save More?" *Journal of Political Economy* 112(2): 397–444.

Einav, L., Finkelstein, A., & Cullen, M. R. (2010). Estimating Welfare in Insurance Markets Using Variation in Prices. *The Quarterly Journal of Economics*, *125*(3), 877-921.

Einav, L., Finkelstein, A., Ryan, S. P., Schrimpf, P., & Cullen, M. R. (2013). Selection on Moral Hazard in Health Insurance. *American Economic Review*,*103*(1), 178-219.

Einav, L., Jenkins, M., & Levin, J. (2012). Contract pricing in consumer credit markets. *Econometrica*, *80*(4), 1387-1432.

Ericson, K.M.M. and Starc, A. (2012). Heuristics and heterogeneity in health insurance exchanges: Evidence from the Massachusetts Connector. *American Economic Review, Papers & Proceedings*, 102(3):493–97.

_____ (2013). How Product Standardization Affects Choice: Evidence from the Massachusetts Health Insurance Exchange (Working Paper No. w19527). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w19527

Ericson, K.M.M. (2014). Consumer Inertia and Firm Pricing in the Medicare Part D Prescription Drug Insurance Exchange. *American Economic Journal: Economic Policy*, *6*(1), 38-64.

Feldstein, M., and J.B. Liebman. 2002. "Social Security." In *Handbook of Public Economics*, 2245–2325.

Finkelstein, A., & Poterba, J. (2013). Testing for Asymmetric Information Using "Unused Observables" in Insurance Markets: Evidence from the UK Annuity Market. *Journal of Risk and Insurance*.

Frakt, A. (2014, May 26). When Hospital Systems Buy Health Insurers. *The New York Times*. Retrieved from http://www.nytimes.com/2014/05/26/upshot/when-hospital-systems-buy-health-insurers.html?abt=0002&abg=1

Gaynor, M., & Vogt, W. B. (2003). Competition among Hospitals. *RAND Journal of Economics*, 764-785.

Gowrisankaran, G., Nevo, A., & Town, R. (2013). Mergers when prices are negotiated: Evidence from the hospital industry (Working Paper No. w18875). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w18875

Gruber, J., & McKnight, R. (2014). Controlling Health Care Costs Through Limited Network Insurance Plans: Evidence from Massachusetts State Employees (Working Paper No. w20462). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w20462

Gruber, J. and Poterba, J. (1994). Tax incentives and the decision to purchase health insurance: Evidence from the self-employed. *Quarterly Journal of Economics*, 109(3):701– 33.

Gruber, J. and Washington, E. (2005). Subsidies to employee health insurance premiums and the health insurance market. *Journal of Health Economics*, 24(2):253–76.

Gustman, A.L., and T.L. Steinmeier. 2005. "The Social Security Early Entitlement Age in a Structural Model of Retirement and Wealth." *Journal of Public Economics* 89(2-3): 441–63.

Handel, B. R. (2013). Adverse selection and inertia in health insurance markets: When nudging hurts. *The American Economic Review*, *103*(7), 2643-2682

Handel, B. R., Hendel, I., & Whinston, M. D. (2013). Equilibria in Health Exchanges: Adverse Selection vs. Reclassification Risk (Working Paper No. w19399). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w19399

Health Care Cost Institute Inc. (HCCI). 2014. 2013 Health Care Cost and Utilization Report. Retrieved from http://www.healthcostinstitute.org/files/2013%20HCCUR%2010-28-14.pdf

Health Policy Commission. 2013. Health Care Cost Trends in the Commonwealth, Hearing, September 16, 2013 (#13632). Retrieved from http://www.mass.gov/anf/docs/hpc/celticare.pdf

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

Ho, K. (2009). Insurer-Provider Networks in the Medical Care Market. *The American Economic Review*, *99*(1), 393-430.

Ho, K., & Pakes, A. (2014). Physician Payment Reform and Hospital Referrals.*The American Economic Review*, *104*(5), 200-205.

Johnson, R.W, L.E Burman, and D.I. Kobes. 2004. "Annuitized Wealth at Older Ages: Evidence from the Health and Retirement Study." Final Report to the Employee Benefits Security Administration, U.S. Department of Labor.

Jousten, A. 2001. "Life-Cycle Modeling of Bequests and Their Impact on Annuity Valuation." *Journal of Public Economics* 79(1): 149–77.

Kaplow, L. 2009. "Utility from Accumulation." NBER Working Paper 15595.

Kotlikoff, L. J., and A. Spivak. 1981. "The Family as an Incomplete Annuities Market." *Journal of Political Economy* 89(2): 372.

Lee, R. S. (2013). Vertical integration and exclusivity in platform and two-sided markets. *The American Economic Review*, *103*(7), 2960-3000.

Liebman, J.B., and E.F.P. Luttmer. 2012. "The Perception of Social Security Incentives for Labor Supply and Retirement: The Median Voter Knows More Than You'd Think." *Tax Policy and the Economy* 26(1): 1–42.

———. 2015. "Would People Behave Differently If They Better Understood Social Security? Evidence from a Field Experiment." *American Economic Journal: Economic Policy* 7(1): 275–99.

Lockwood, L.M. 2014. "Incidental Bequests : Bequest Motives and the Choice to Self-Insure Late-Life Risks." *mimeo*.

Love, D.A., M.G. Palumbo, and P.A. Smith. 2009. "The Trajectory of Wealth in Retirement." *Journal of Public Economics* 93(1-2): 191–208.

Mahoney, N. (2012). Bankruptcy as implicit health insurance. *NBER Working Paper 18105*.

Mahoney, N., & Weyl, E. G. (2013). Imperfect Competition in Selection Markets. *Unpublished Manuscript*.

Marshall, S., K. McGarry, and J.S Skinner. 2011. "The Risk of Out-of-Pocket Health Care Expenditure at the End of Life." In *Explorations in the Economics of Aging*, ed. David A Wise. University of Chicago Press, 101–28.

Mastrobuoni, G. 2011. "The Role of Information for Retirement Behavior: Evidence Based on the Stepwise Introduction of the Social Security Statement." *Journal of Public Economics* 95(7-8): 913–25.

McKinsey Center for U.S. Health System Reform (2013). Exchanges go live: Early trends in exchange dynamics.

McKinsey for U.S. Health System Reform. (2014). Hospital networks: Updated national view of configurations on the exchanges. Retrieved from http://healthcare.mckinsey.com/hospital-networks-updated-national-view-configurations-exchanges

McWilliams, J. M., Hsu, J., & Newhouse, J. P. (2012). New risk-adjustment system was associated with reduced favorable selection in Medicare Advantage. *Health Affairs*, *31*(12), 2630-2640.

MedPAC (2013). Report to the Congress (March): Medicare payment policy. Chapter 13: The Medicare Advantage Program.

Miller, R. H., & Luft, H. S. (1997). Does managed care lead to better or worse quality of care?. *Health affairs*, *16*(5), 7-25

Mitchell, O.S, J.M Poterba, M.J Warshawsky, and J.R. Brown. 1999. "New Evidence on the Money's Worth of Individual Annuities." *American Economic Review* 89(5): 1299–1319.

Newhouse, J. P., Garber, A. M., Graham, R. P., McCoy, M. A., Mancher, M., & Kibria, A. (Eds.). (2013). *Variation in health care spending: target decision making, not geography*. National Academies Press.

Office of Attorney General Martha Coakley. (2013). Examination of Health Care Cost Trends and Cost Drivers. Retrieved from http://www.mass.gov/ago/docs/healthcare/2013-hcctd.pdf

Poterba, J.M, S.F Venti, and D.A. Wise. 2009a. "The Decline of Defined Benefit Retirement Plans and Asset Flows." In *Social Security Policy in a Changing Environment*, eds. Jeffrey R. Brown, Jeffrey B. Liebman, and David A Wise. University of Chicago Press.

_____. 2009b. "The Changing Landscape of Pensions in the United States." In *Overcoming the Saving Slump: How to Increase the Effectiveness of Financial Education and Saving Programs*, ed. Annamaria Lusardi. University of Chicago Press, 17–46.

_____. 2011. "Family Status Transitions, Latent Health, and the Post-Retirement Evolution of Assets." In *Explorations in the Economics of Aging*, ed. David A Wise. , 23–69.

Richards, T. (October 30, 2013) "Top Hospitals Opt Out of Obamacare." *U.S. News & World Report*. Retrieved from: http://health.usnews.com/health-news/hospital-of-tomorrow/articles/2013/10/30/top-hospitals-opt-out-of-obamacare.

Rothschild, M., & Stiglitz, J. (1976). Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. *The Quarterly Journal of Economics*, *90*(4), 629-649.

Shefrin, H.M, and R.H. Thaler. 1988. "The Behavioral Life-Cycle Hypothesis." *Economic Inquiry* 26: 609–43.

Shoven, J.B., and S.N. Slavov. 2013. "Does It Pay to Delay Social Security?" *Journal of Pension Economics and Finance* (October): 1–24.

Social Security Administration, (SSA). 2009. *Annual Statistical Supplement to the Social Security Bulletin*.

Song, J.G., and J. Manchester. 2007. "New Evidence on Earnings and Benefit Claims Following Changes in the Retirement Earnings Test in 2000." *Journal of Public Economics* 91(3-4): 669–700.

Song, Z., Safran, D. G., Landon, B. E., He, Y., Ellis, R. P., Mechanic, R. E., Day, M.P. & Chernew, M. E. (2011). Health care spending and quality in year 1 of the alternative quality contract. *New England Journal of Medicine*, 365(10), 909-918.

Starc, A. (2014). Insurer pricing and consumer welfare: Evidence from medigap. *The RAND Journal of Economics*, *45*(1), 198-220.

Sun, W., and A. Webb. 2011. "Valuing the Longevity Insurance Acquired by Delayed Claiming of Social Security." *Journal of Risk and Insurance* 78(4): 907–30.

Town, R., & Vistnes, G. (2001). Hospital competition in HMO networks. *Journal of Health Economics*, *20*(5), 733-753.

Turra, C.M., and O.S. Mitchell. 2004. "The Impact of Health Status and Out-of-Pocket Medical Expenditures on Annuity Valuation." PARC Working Paper 04-02.

Warner, J.T., and S. Pleeter. 2001. "The Personal Discount Rate: Evidence from Military Downsizing Programs." *American Economic Review* 91(1): 33–53.

Werden, G. J. (1996). A robust test for consumer welfare enhancing mergers among sellers of differentiated products. *Journal of Industrial Economics*, 44(4):409–413.

Yaari, M.E. 1965. "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer." *Review of Economic Studies* 32(2): 137–50.

Appendix A Appendices for Chapter 1

A.1 Sample Summary Statistics

Hospital Choice Sample

Patient Characteristics

Variable		Mean
No. of Hospitalizations		74,383
Age		44.6
Male		49%
Emergency Department		56%
Diagnoses	Mental Illness	16.7%
	Digestive	13.5%
	Circulatory	12.6%
	Injury / Poisoning	7.1%
	Respiratory	7.0%
	Cancer	6.4%
	Endocrine / Metabolic	6.0%
	Musculoskeletal	5.6%
	Genitourinary	5.1%
	Pregnancy / Childbirth	5.0%
	All Others	14.9%

Chosen Hospital Statistics

Variable	Mean	Std. Dev.
Distance: Chosen Hosp. (miles)	14.1	16.3
All Hospitals (miles)	48.4	25.9
Hospital Category		
Academic Med. Ctr.	29%	---
Teaching Hospital	19%	---
All Others	52%	---
Partners Hospital	14%	---
Out-of-Network	8%	---
Past Used Hospital (>60 days before)		
Any Use	54%	---
Inpatient Use	19%	---
Outpatient Use	51%	---
Total Cost to Insurer	$11,369	$15,711
Price (estimated)	$10,981	$4,112
Patient Severity (estimated)	1.000	0.310

Plan Choice Sample

Enrollee Characteristics

Variable		Mean	Std. Dev.
No. of Enrollees		611,455	---
Age		39.6	13.8
Male		46.5%	---
Income:	<100% Poverty	47.1%	---
	100-200% Poverty	39.6%	---
	200-300% Poverty	13.3%	---
Past Hospital User		44.3%	---
	Partners Hospitals	7.4%	---
	Other Hospitals	40.3%	---
Risk Adjustment Score		0.99	0.90
Choice Type:	New Enrollee	29.5%	---
	Re-Enrollee	13.5%	---
	Current Enrollee	57.1%	---

Plan Choice Statistics

Variable	Mean	Std. Dev.
No. of Choice Instances	1,588,889	---
Insurer Price	$380.7	$69.5
Cons. Premium: Below Poverty	$0.0	$0.0
Above Poverty	$47.3	$45.7
Costs per Month: Total	$371.5	$1,480
Hospital Inpatient	$81.5	$1,048
Non-Inpatient	$290.0	$873
Current Enr: Non-Switching	95.8%	---
Market Shares: BMC	35.5%	---
Network Health	34.7%	---
NHP	19.2%	---
CeltiCare	6.9%	---
Fallon	3.8%	---

A.2 Demand and Cost Model Estimation Details

A.2.1 Insurance Plan Demand Estimation Details

I estimate the plan demand model parameters by matching moments that fall into two categories. First, for plan dummies, I match market shares for the appropriate region/year/income group *g*. These market shares uniquely identify plan mean utilities, which in my case are equivalent to the plan dummies.[1] The formula for these market share moments is:

$$G_{j,g}^{(1)}(\theta) = \frac{1}{N}\sum_{i,j,t} 1\{i,t \in g\} \cdot \left[1\{y_{it} = j\} - Pr(y_{it} = j \mid \theta)\right]$$

where θ is the parameter vector, $1\{y_{it} = j\}$ is an indicator for whether individual *i* chose plan *j* at time *t*, and $Pr(y_{it} = j \mid \theta)$ is the predicted choice share from the logit model.

Second, for the coefficients for premium, network utility, and other observed characteristics (which are interacted with observed enrollee attributes), I match the average values for chosen plans in the data to those in the model. Specifically, the moments for characteristic $X^{(k)}$ (e.g., premium) interacted with enrollee attribute $Z^{(r)}$ (e.g., income) are:

$$G_{k,r}^{(2)}(\theta) = \frac{1}{N}\sum_{i,j,t} X_{ijt}^{(k)} Z_{it}^{(r)} \cdot \left[1\{y_{it} = j\} - Pr(y_{it} = j \mid \theta)\right]$$

Another way of interpreting these is as matching the *covariance* between plan characteristics and household attributes. In the case of observing the full market, these moments are equivalent to the micro BLP covariance moments. These moments are also equivalent to first-order conditions from the associated maximum likelihood problem.

Stacking all of the moments into a vector $G(\theta)$, the MSM estimator searches for the parameter θ that minimizes the weighted sum of squared moments, $G(\theta)' \cdot W \cdot G(\theta)$. Because the system is just-

[1] A difference in my setting from the standard BLP approach is that I treat the plan dummies as parameters, with associated standard errors, since both they and the characteristics coefficients are estimated from a dataset of the same size (the full market data). In previous applications including Berry, Levinsohn, and Pakes (2004), the micro data came from a sample, while the market shares came from aggregate data on the whole market.

identified (equal number of parameters and moments), I am able to match the moments exactly, making the solution invariant to the choice of W. I calculate standard errors using the standard GMM sandwich formula. To account for the fact that network utility variable is derived from the hospital demand estimates, I am planning to implement an adjustment following the lecture notes of Pakes (2013). However, I have not yet implemented this adjustment in the current draft.

A.2.2 Inattention Interpretation of Plan Inertia Coefficients

For current enrollees, I included in the logit demand model a dummy variable for their current plan, so their full demand utility was:

$$U_{ijt}^{Curr} = \hat{U}_{ijt} + \underbrace{\chi(Z_i) \cdot 1\{j = CurrPlan\}}_{\text{Switching Cost / Inertia}} + \varepsilon_{ijt}^{Plan} \tag{4.1}$$

where $\hat{U}_{ijt}$ is the plan utility for new enrollees (defined in Section 1.5.3), excluding the ε_{ijt}^{Plan}. In this equation, $\chi(Z_i)$ is interpreted as a switching cost – an extra utility for the current plan needed to rationalize the low level of plan switching. The plan demand estimates in Table 1.7 reports these switching costs but also an alternate interpretation based on an inattention model. I show here how I derive the inattention/passive probability reported in Table 1.7.

Consider a two-step model in which the first step models whether enrollees make an active choice, and the second step models plan choice conditional on being active. The second step is standard and follows the logit model for new enrollees (or current enrollees excluding switching cost):

$$Pr(y_{it} = j \mid Active) = \frac{\exp(\hat{U}_{ijt})}{\sum_k \exp(\hat{U}_{ikt})}$$

The first step is a reduced form model of being passive:

$$Pr_{it}\left(Passive\right)=\frac{\exp\left(\hat{U}_{i,j_{curr},t}+\tilde{\chi}_i\right)}{\exp\left(\hat{U}_{i,j_{curr},t}+\tilde{\chi}_i\right)+\exp\left(I_{i,Active,t}\right)} \quad \text{where } I_{i,Active,t}=\log\left(\sum_k \exp\left(\hat{U}_{ikt}\right)\right)$$

Notice that it is the choice probability from a two-choice logit model, where the utility of being passive is the current plan utility plus a reduced-form inertia coefficient $\tilde{\chi}_i$ (which is different from the switching cost χ). The utility of being active is $I_{i,Active,t}$, which is the inclusive value (or expected utility) from the second-stage active choice model.

I claim that if $\tilde{\chi}_i=\log\left(\exp\left(\chi(Z_i)\right)-1\right)$, the switching cost and inattention models have identical predictions for choice probabilities. For the current plan, the inattention model predicts a probability that it is chosen of $Pr_{it}\left(Passive\right)+\left(1-Pr_{it}\left(Passive\right)\right)\cdot Pr\left(y_{it}=j_{curr}\mid Active\right)$, which simplifies to:

$$Pr\left(y_{it}=j_{curr}\right)=\frac{\exp\left(\hat{U}_{i,j_{curr},t}+\chi\left(Z_i\right)\right)}{\exp\left(\hat{U}_{i,j_{curr},t}+\chi\left(Z_i\right)\right)+\sum_{k\neq j_{curr}}\exp\left(\hat{U}_{ikt}\right)}$$

This equals the current plan's choice probability in the switching cost model in (4.1). Further, the inattention model's probability of switching to another plan j is $\left(1-Pr_{it}\left(Passive\right)\right)\cdot Pr\left(y_{it}=j\mid Active\right)$, which simplifies to:

$$Pr\left(y_{it}=j\right)=\frac{\exp\left(\hat{U}_{i,j,t}\right)}{\exp\left(\hat{U}_{i,j_{curr},t}+\chi\left(Z_i\right)\right)+\sum_{k\neq j_{curr}}\exp\left(\hat{U}_{ikt}\right)}$$

which is again equivalent to the choice probability from the switching cost model in (4.1).

Hence, these two models have equivalent predictions for choice probabilities. The plan demand results in Table 1.7 report both the average switching costs $\chi\left(Z_i\right)$ and the passive probability $Pr_{it}\left(Passive\right)$, as defined by the equation above.

A.2.3 Details of Hospital Price Model

As discussed in Section 1.6.1, I estimate a risk-adjusted hospital price model. Recall that I estimate a Poisson regression (also known as a generalized linear model with a log link) of the form:

$$E\left[Payment_{i,j,h,t,a} \mid Diag_{ita}, Z_{ita} \right] = \exp\left(\rho_{j,h,t} + Diag_{ita}\lambda + Z_{ita}\gamma \right)$$

For the principal diagnosis ($Diag_{ita}$), I use the Clinical Classification Software (CCS) dummies defined by the U.S. government's Agency for Healthcare Research and Quality. The additional covariates (Z_{ita}) include age, gender, income, and Elixhauser comorbidity dummies for the secondary diagnoses.

I specify a restricted model for $\rho_{j,h,t}$ to avoid over-fitting for hospital-insurer-year cells with small samples. Specifically, I start from the model:

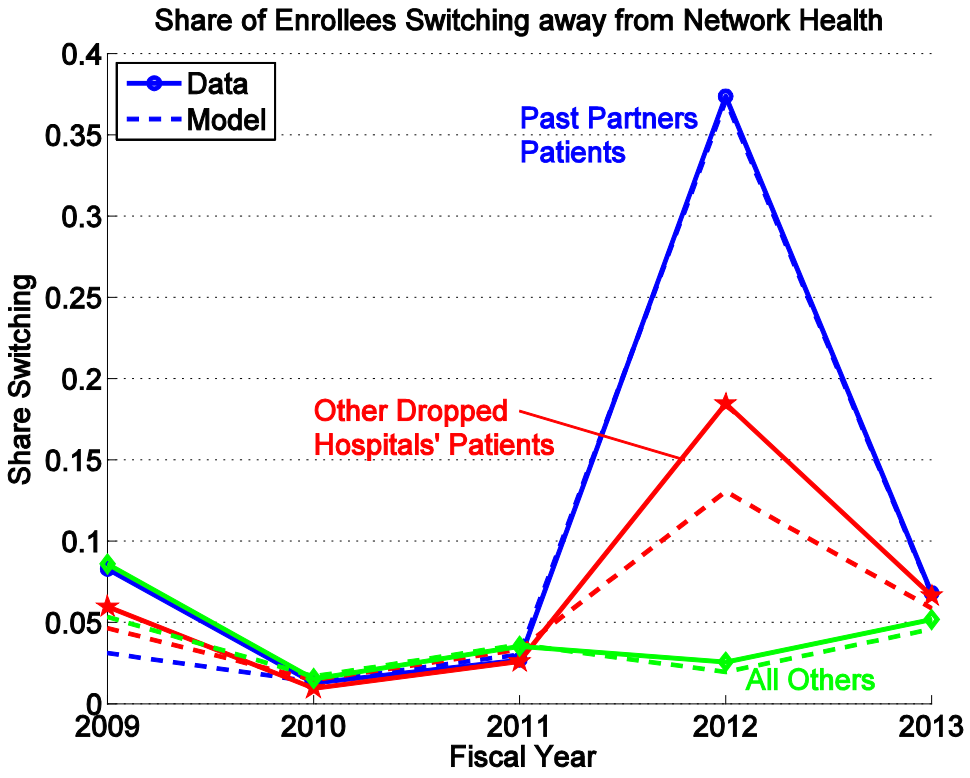
$$\rho_{j,h,t} = \rho_{j,h,NetwStat(h,t)} + \rho_{j,Sys(h),t} + \rho_{j,t,NetwStat(h,t)}$$

The first term, $\rho_{j,h,NetwStat(h,t)}$, is a coefficient on hospital-insurer-network status (i.e., in or out of network) dummies that is constant across years. I include this term for all cells with at least 50 observations; otherwise, I set it to zero. The second term, $\rho_{j,Sys(h),t}$, is a coefficient on insurer-hospital system-year dummies for the top six hospital systems. This allows for a separate hospital price paths over time for each of the largest systems (including Partners). I do not include this term for hospitals in smaller systems or when the large system is out-of-network, with the exception that I always include these dummies for Partners regardless of whether it is in-network. The final term, $\rho_{j,t,NetwStat(h,t)}$, is a residual that allows for a separate effect for each plan, year, and network status. This captures the average insurer-specific price path for all smaller hospitals not included in one of the six largest systems.

A.3 Model Fit Tables and Figures

This appendix shows the model's ability to match the reduced form patterns around Network Health's dropping of the Partners hospitals in 2012, as discussed in Section 1.6.4.

Appendix Table A.3.1. Plan Switching Patterns

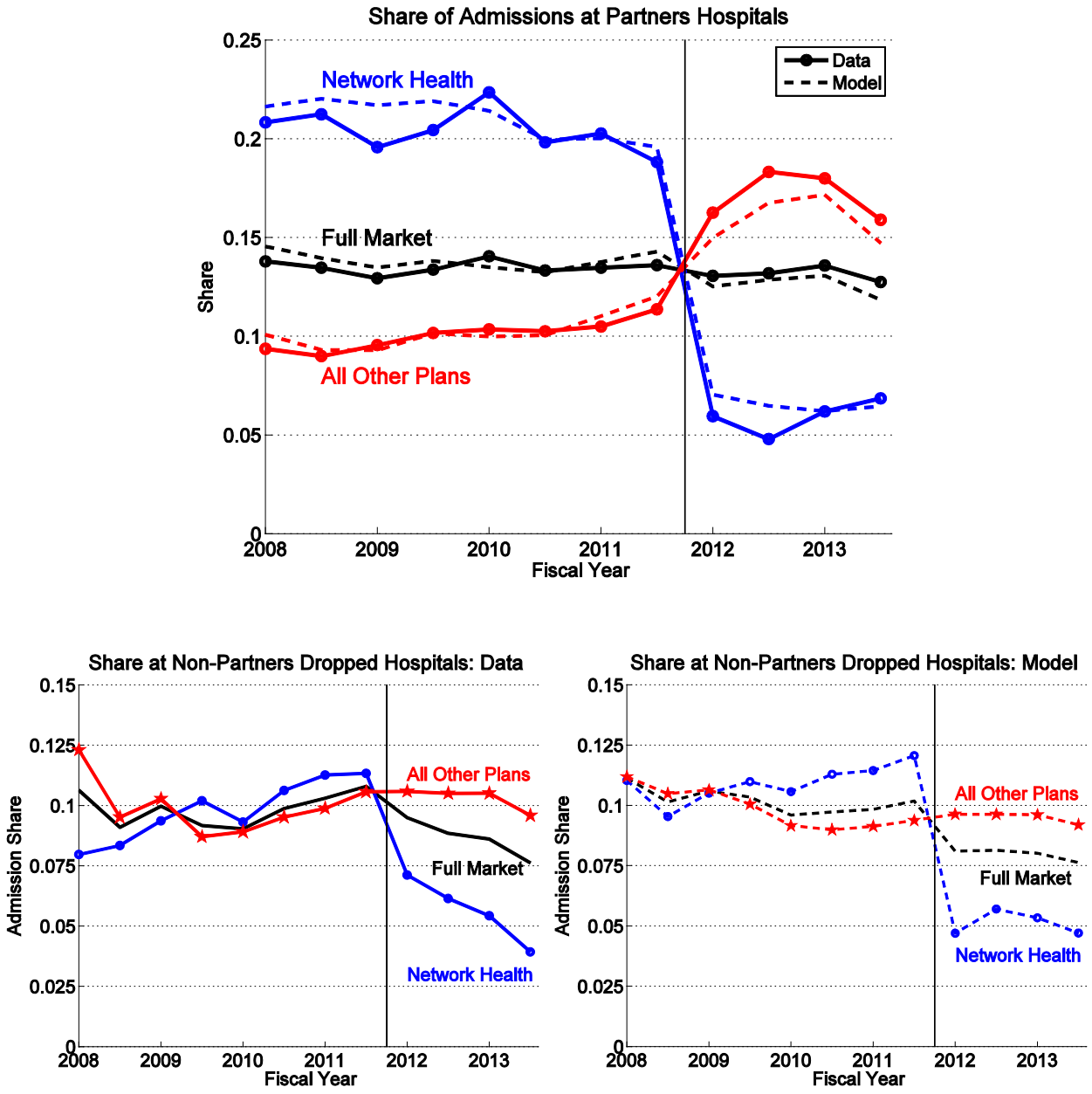


Appendix Table A.3.2. Cost Changes for Network Health

Network Health: Average Costs 2011-12

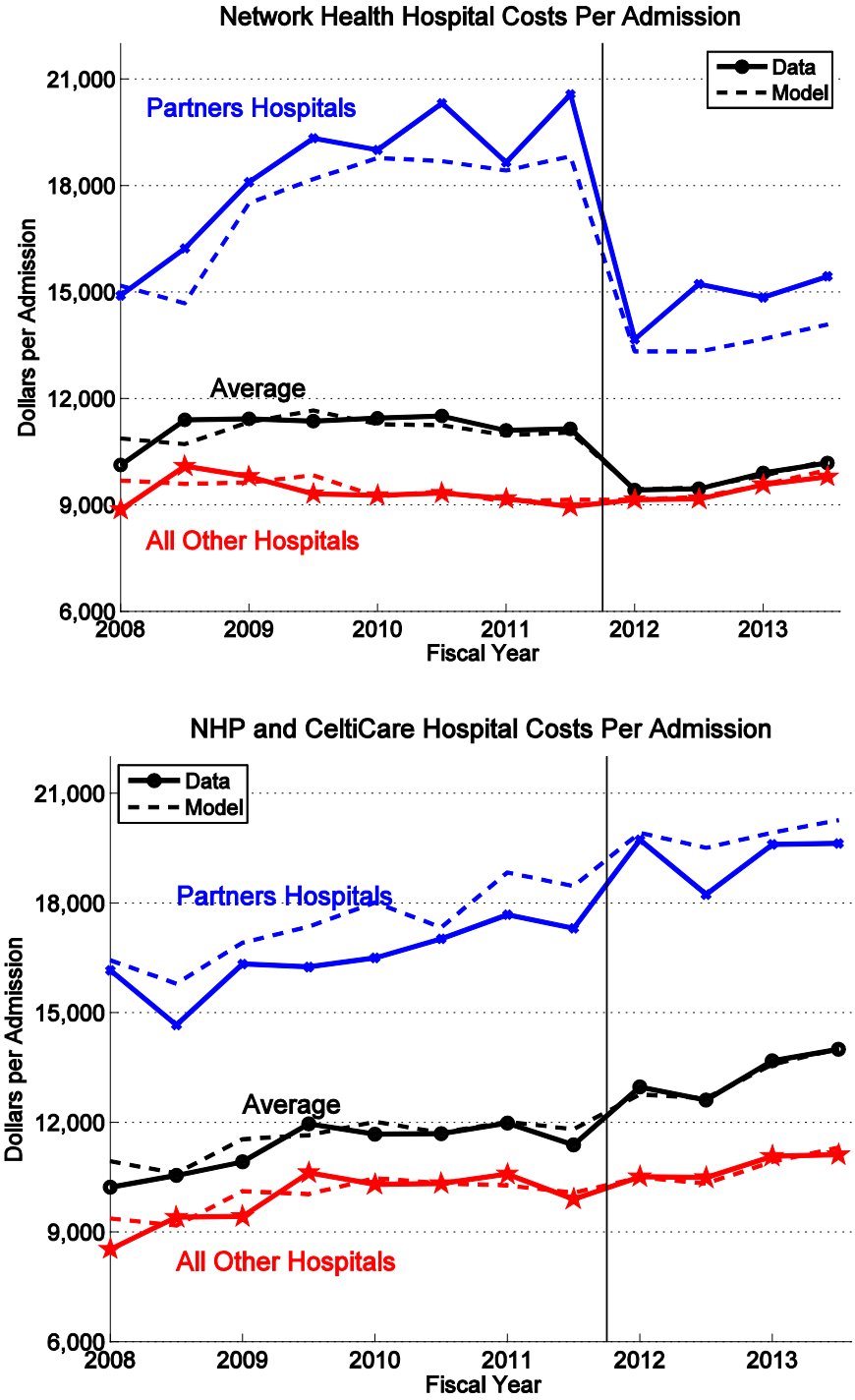
Enrollee Group	Data				Model			
	2011	2012	%Δ	Risk Adj. %Δ	2011	2012	%Δ	Risk Adj. %Δ
All Enrollees	**$378**	**$313**	**-17%**	**-15%**	**$374**	**$310**	**-17%**	**-16%**
Stayers (in plan both years)	$317	$305	-4%	-5%	$334	$312	-7%	-9%
2011 Only Enrollees	$476	---	---		$435	---	---	
2012 Only Enrollees	---	$310	---		---	$302	---	

Appendix Table A.3.3. Admission Shares at Hospitals Dropped by Network Health in 2012



NOTE: These figures show the share of hospital admissions at hospitals that Network Health plan dropped from its network in 2012. See Figure 1.2 for a more detailed description of the values in the data. The dashed lines show the model's prediction for the same statistics. These are calculated holding fixed each individual's observed plan, not reassigning plan choices using the plan demand model.

Appendix Table A.3.4. Changes in Cost per Hospital Admission around 2012 Network Changes



NOTE: These figures show average costs per hospital admission for two sets of plans: Network Health (top figure), which dropped the star Partners hospitals in 2012, and NHP and CeltiCare (bottom figure), which continued to cover them. See Figure 1.3 for a more detailed description of the values in the data. The dashed lines show the model's prediction for the same statistics. These are calculated holding fixed each individual's observed plan, not reassigning plan choices using the plan demand model.

A.4 Simulation Method Details

This appendix details the simple approach I use to incorporate a future profit effect in a static pricing model for my simulations in Section 1.8. Note that in a dynamic model, an insurer's pricing FOC includes a term capturing the effect of changing today's price on future profits on consumer *i*. I model this "future profit effect" as the product of the change in future demand $(\partial D_{ij}^{Fut} / \partial P_j)$ times an expected profit margin M_{ij}^{Fut}, which is unaffected by today's price. For the change in future demand, a lower price increases demand today and therefore increases the number of inertial enrollees in the future. To simplify, I take future market enrollment $(nMon_{i,t+k})$ as given and assume an exogenous, constant inertia probability η at each year's switching choice, which I set at 89%.[2] Given these assumptions:

$$\frac{\partial D_{ij}^{Fut}}{\partial P_j} = \frac{\partial S_{ij}}{\partial P_j} \cdot \left(\sum_{k \geq 1} \eta^k \cdot nMon_{i,t+k} \right) \tag{4.2}$$

where $\partial S_{ij} / \partial P_j$ is the effect of price on current year's choice share.

Finally, I need to specify insurers' future profit margins. Although imperfect, I simply assume that insurers expect M_{ij}^{Fut} to equal current margins at the enrollee level – which assumes that prices and costs grow in parallel for each enrollee. Notice that I still treat M_{ij}^{Fut} as a constant in the pricing FOC but plug in the equilibrium margin $(= \varphi_i P_j^* - c_{ij}\left(N_j\right))$ for it at the end.

Combining these assumptions and defining the term in parentheses in (4.2) as $nFutMon_i$, the pricing FOC for insurer *j* is:

[2] I use 89% rather than the 95% inertia probability reported in the plan demand estimates based on a rough correction for unobserved heterogeneity. Looking at re-enrollees (people who leave the market and return later), people tend to actively choose the same plan as during their prior spell about 55% of the time. For an inertia probability of ρ the overall non-switching probability is $\rho + (1-\rho) \cdot Pr_j^{Active}$. Plugging in $Pr_i^{Active} = 55\%$, $\rho = 89\%$ is required to rationalize a 95% overall non-switching probability.

$$
\begin{aligned}
0 &= \frac{\partial \pi_j}{\partial P_j} + \sum_i M_{ij}^{Fut} \cdot \frac{\partial D_{ij}^{Fut}}{\partial P_j} \\
&= \sum_i \varphi_i \cdot nMon_i \cdot S_{ij}(.) + \sum_i \left(\varphi_i P_j - c_{ij}\right)\left(nMon_i + nFutMon_i\right) \cdot \frac{\partial S_{ij}}{\partial P_j}
\end{aligned}
\tag{4.3}
$$

Accounting for future profits adds the $nFutMon_i$ term to the FOC, which increases the incentive to lower prices (just like a steeper demand curve). This effect is likely to have a significant impact. Months in the current year $(nMon_i)$ average 6.2, and future months $(nFutMon_i)$ average 6.8. So the future profit effect works like a more than doubling of the demand slope.

Appendix B Appendix for Chapter 2

We use the smaller changes in the affordable amount that occurred for the 150-200% poverty group in July 2007 as a check on the semi-elasticity estimated for this group. Appendix Table B shows the results. Our preferred triple-difference estimate from column (3), indicates that the $5 affordable amount decrease translated into a 6.3% increase in demand for the cheapest plan. This implies a semi-elasticity of 0.0126, quite similar to the semi-elasticity of 0.0119 that we found for this group using the mandate penalty introduction (see Table 2.5).

Appendix Table B. Decrease in the Affordable Amount: 150-200% Poverty

Dependent Var: New Enrollees in Cheapest Plan / June 2008 Enrollment

Variable	(1)	(2)	(3)
Sum of July - Oct 2007 coefficients (below)	**0.063*** (0.019)**	**0.052** (0.025)**	**0.063** (0.027)**
150-200% Poverty x July2007	0.013***	0.017**	0.018**
	(0.005)	(0.007)	(0.008)
x Aug2007	0.029***	0.016**	0.022***
	(0.005)	(0.007)	(0.008)
x Sep2007	0.013**	0.011*	0.015**
	(0.005)	(0.006)	(0.007)
x Oct2007	0.008	0.009	0.009
	(0.005)	(0.006)	(0.006)
Control Group (200-300% poverty)		X	X
Triple Difference (dummies for July - October)			X
Observations	52	104	104
R-Squared	0.960	0.957	0.958

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

NOTE: This table performs the difference-in-difference regressions analogous to those in Table 2.4 (see the note to that table for additional information), but with a treatment group of enrollees 150-200% poverty, whose affordable amount dropped from $40 to $35 in June 2008. The control group continues to be enrollees 200-300% poverty, whose affordable amounts were essentially unchanged at that time.

Appendix C Appendices for Chapter 3

C.1 Data and Sample Construction

I study Social Security claiming using data from the Health and Retirement Study (HRS), a nationally representative panel survey of older Americans. I use data from the RAND version of the HRS for first nine biennial waves from 1992-2008. Starting from the full HRS cohort of 13,596 individuals, I drop 3,212 Social Security Disability Insurance recipients and others (mostly widows) who start Social Security before 62. Next, to ensure a sufficient period of observation, I drop 4,908 people not born between 1931 and 1938. Finally, I drop 869 individuals who enter the sample after turning 62 or who exit before age 62, and 428 people whose claiming age is unavailable. The resulting sample contains 4,179 individuals. Summary statistics for this sample are shown in Appendix Table C.1.1. I use this sample for all of the main analysis of Propositions 1 and 2. All analysis is weighted using the respondent survey weight from the first interview an individual is in the data.

For two analyses, I use data from older cohorts contained in the same RAND HRS dataset. For median assets at ages 76-85 in Figure 3.3, I use data for individuals born from 1917-1923 in the parallel Asset and Health Dynamics of the Oldest Old (AHEAD) survey. As in the HRS sample, I drop individuals who ever received Disability Insurance or claimed Social Security before age 62. However, I do not drop individuals for the other reasons, which are not applicable for estimating assets. For the estimates of probabilities of first entering a nursing home for a long-term stay (>60 days) shown in Figure 3.6, I use data from the HRS, AHEAD and intermediate Children of the Depression (CODA) cohort (born 1924-1930). I use the full AHEAD sample, not just those born from 1917-1923. I again drop anyone who received Disability Insurance or claimed Social Security before 62 but do not make any other exclusions.

Social Security claiming time is a self-reported variable in the public-use data, which I use for the current draft (though I recently gained access to linked administrative Social Security records, which I will use in a future draft). Social Security claiming age is a monthly variable in the data, but the month of

claiming is missing for many individuals and has been recoded to the month after an individual's birthday. Because of the bias this may create, I ignore the monthly component and treat Social Security claiming age as an annual variable. My analysis also requires calculating the Social Security benefit received at claiming. Annual Social Security retirement income is a self-reported variable, usually for the calendar year before the interview. Because the beneficiary only receives a partial year's benefit during the year of claiming, I use the annual Social Security income in the first wave in which the retiree was at least a year older than his claiming age in the income reporting year. I scale this benefit down using the actual Cost-of-Living Adjustments that occurred between the claiming age and this income observation. This method introduces more noise into a measure that is already subject to measurement error, but given the high level of assets, the error is unlikely to significantly affect the test of Proposition 1. Finally, I use two measures of assets in the analysis:

- "Total assets" is a relatively comprehensive measure, which includes all housing, financial, real estate, business, vehicle, and other wealth (except for second homes in wave 3 when a survey error resulted in this variable being lost), net of any mortgages and debts (RAND HRS variable HwATOTB, except in wave 3 when it is H3ATOTA). From this, I also subtract the value of "other savings" (RAND variable HwAOTHR), since this could in theory include annuity wealth (though in practice, most people have zero wealth in this category). The only major asset this measure misses is 401(k) balances that are maintained in the employer account after retirement (rather than withdrawn or rolled over into an IRA).

- "Non-housing assets" is a more conservative measure, which starts from total assets and excludes net equity in primary and secondary residences, as well as vehicle wealth (RAND HRS variable HwATOTN, less HwATRANS and HwAOTHR).

All of these measures of assets exclude defined benefit and defined contribution pension wealth that has not yet been received as a lump sum or that has been promised as a future annuity. This makes the asset measures even more conservative estimates of retirees' access to liquidity.

Appendix Table C.1.1

HRS Sample Summary Statistics

Variable	Mean	Median	Std. Dev.
Male	0.492	---	0.500
Married (Wave 1)	0.811	---	0.391
Year of Birth	1934.4	1934.0	2.3
Education: High School Dropout	0.187	---	0.390
High School Grad / GED	0.406	---	0.491
Some College	0.202	---	0.402
College Graduate	0.205	---	0.404
Social Security Claiming:			
Claiming Age (in years)	63.14	62.00	1.50
Claim at Age 62	0.538	---	0.499
Claim at Ages 63-65	0.403	---	0.490
Claim at Ages 66+	0.059	---	0.236
Annual Benefit at Claiming*	$9,341	$8,999	$4,398
Self-Reported Probabilities:			
Live to 75**	0.678	0.717	0.238
Live to 80**	0.547	0.500	0.278
Live to 85**	0.442	0.450	0.284
Nursing Home w/in 5 Yrs.***	0.095	0.000	0.170

* In year 2000 dollars, ** Avg. of self-reports within 2 years of claiming, *** First report

C.2 Proofs of Propositions

Proposition 1: Starting from the path of consumption, assets, and Social Security that the individual actually chose, $\{c_t, a_{t+1}\}_{t=s}^{\infty}$, execute the following arbitrage transaction. Delay claiming benefits until period $s+k$. The change in benefit income is $\Delta ben_t = -b_s$ for $t = s,...,s+k-1$ and $\Delta ben_t = (b_{s+k} - b_s)$ for $t \geq s+k$. Compensate for this by reducing conventional savings in each period by:

$$\Delta a_{t+1} = \begin{cases} -\sum_{i=s}^{t} R^{i-s} b_s & \text{for } t = s,\ldots,s+k-1 \\ \min\left\{0, -\sum_{i=s}^{t} R^{i-s} b_s + \sum_{i=s+k}^{t} R^{i-(s+k)} b_{s+k}\right\} & \text{for } t \geq s+k \end{cases}$$

The conditions in the proposition ensure that asset reductions of at least this size are feasible. Because $R_{SS}^{s,s+k} > R^k$, the second term in braces in the expression for Δa_{t+1} for $t \geq s+k$ will eventually turn positive, meaning no further asset reductions after some date. To see this, factor R^{t-s} from the expression and reorder the indexing, so that it equals:

$$= R^{t-s}\left(-\sum_{i=0}^{t-s} R^{-i} b_s + \sum_{i=k}^{t-s} R^{-i} b_{s+k}\right)$$

As $t \to \infty$, the expression in parentheses converges to:

$$\begin{aligned} -\sum_{i=0}^{\infty} R^{-i} b_s + \sum_{i=k}^{\infty} R^{-i} b_{s+k} &= -\left(\tfrac{R}{R-1}\right) b_s + \left(\tfrac{R^{1-k}}{R-1}\right) b_{s+k} \\ &= \left(\tfrac{R^{1-k}}{R-1}\right) b_s \cdot \left(R_{SS}^{s,s+k} - R^k\right) \\ &> 0 \end{aligned}$$

where the inequality uses condition (a).

In each period, the change in consumption is determined by the budget constraint: $\Delta c_t = R\Delta a_t - \Delta a_{t+1} + \Delta ben_t$. It is simple but tedious to verify that up to the date when $\Delta a_{t+1} = 0$, consumption will be unchanged and afterward, consumption will increase. Therefore, because the asset reductions are costless by Assumption 1, Social Security delay increases utility. ■

Proposition 2: Consider the arbitrage transaction that delays Social Security from *s* to *s*+1 but holds consumption fixed in every period. Initially, assets fall to fund consumption during period *s*, and from *s*+1 on, assets are rebuilt using the incremental Social Security benefits $b_{s+1} - b_s$. The change in savings in every period for this transaction is exactly equal to the change in the PDV of benefits accumulated through time *t*:

$$\Delta a_{t+1} = \sum_{k=s+1}^{t} R^{t-k} b_{s+1} - \sum_{k=s}^{t} R^{t-k} b_s = \Delta Ben_t$$

Because of the minimum asset assumptions in (c), this transaction never violates any liquidity constraint. The change in utility from this transaction is, to first-order approximation:

$$\begin{aligned} dU_s \approx \widetilde{dU_s} &= \sum_{t=s}^{\infty} \beta^{t-s} q_t v'(a_{t+1}) \Delta a_{t+1} \\ &= \sum_{t=s}^{\infty} \beta^{t-s} q_t v'(a_{t+1}) \Delta Ben_t \\ &= \sum_{t=s}^{T_{s,s+1}^{BE}-1} \beta^{t-s} q_t \Delta Ben_t \cdot v'(\overline{a}_{early}) + \sum_{t=T_{s,s+1}^{BE}}^{\infty} \beta^{t-s} q_t \Delta Ben_t \cdot v'(\overline{a}_{late}) \end{aligned}$$

where the last line is by definition of the certainty equivalence values $\overline{a}_{early}$ and $\overline{a}_{late}$. By definition of the break-even age, the change in assets terms are positive for every $t \geq T_{s,s+1}^{BE}$. Therefore, because of condition (b) that $\overline{a}_{early} \geq \overline{a}_{late}$ and the weak-concavity of $v(.)$, we have $v'(\overline{a}_{early}) \leq v'(\overline{a}_{late})$ and therefore:

$$\begin{aligned} \widetilde{dU} &\geq \sum_{t=s}^{T_{s,s+1}^{BE}-1} \beta^{t-s} q_t \Delta Ben_t \cdot v'(\overline{a}_{early}) + \sum_{t=T_{s,s+1}^{BE}}^{\infty} \beta^{t-s} q_t \Delta Ben_t \cdot v'(\overline{a}_{early}) \\ &= v'(\overline{a}_{early}) \sum_{t=s}^{\infty} \beta^{t-s} q_t \cdot \Delta Ben_t \\ &\geq 0 \end{aligned}$$

where the last line uses condition (a). Therefore, $\widetilde{dU} \geq 0$, and the transaction is optimal, proving the proposition. ■

Claim: If the discount rate equals the interest rate $(\beta = R^{-1})$ and the risk in question is mortality – so $q_t = p_{s,t}(1 - p_{t,t+1})$ where $p_{x,y}$ is the probability of surviving from *x* to *y* – condition (a) of Proposition 2 is satisfied iff Social Security delay is actuarially fair or better, that is iff $R_{SS}^{s,s+1} \geq R_{fair}^{s,s+1}$ $= 1 + \left(\sum_{t=s+1}^{\infty} R^{s-t} p_{s,t} \right)^{-1}$.

Proof: Suppose $R_{SS}^{s,s+1} \geq R_{fair}^{s,s+1}$. Then:

$$\sum_{t=s}^{\infty}\beta^{t-s}q_t\cdot\Delta B_t=\sum_{t=s}^{\infty}R^{s-t}q_t\cdot\left[\left(\frac{R^{t-s}-1}{R-1}\right)R_{SS}^{s,s+1}-\left(\frac{R^{t-s+1}-1}{R-1}\right)\right]b_s$$
$$\geq\sum_{t=s}^{\infty}R^{s-t}q_t\cdot\left[\left(\frac{R^{t-s}-1}{R-1}\right)R_{fair}^{s,s+1}-\left(\frac{R^{t-s+1}-1}{R-1}\right)\right]b_s$$

I claim that this last expression equals zero. Rearranging it, this is true iff:

$$R_{fair}^{s,s+1}=\frac{\sum_{t=s}^{\infty}q_t\left(R-R^{s-t}\right)}{\sum_{t=s}^{\infty}q_t\left(1-R^{s-t}\right)}=\frac{\sum_{t=s}^{\infty}p_{s,t}\left(1-p_{t,t+1}\right)\left(R-R^{s-t}\right)}{\sum_{t=s}^{\infty}p_{s,t}\left(1-p_{t,t+1}\right)\left(1-R^{s-t}\right)}$$

Now because $p_{s,t}\left(1-p_{t,t+1}\right)$ is the probability that the time of death occurs right after period t and because the individual has to die at some point after s (at which point he is alive), we have $\sum_{t=s}^{\infty}p_{s,t}\left(1-p_{t,t+1}\right)=1$. Therefore, the RHS of this condition simplifies to:

$$R_{fair}^{s,s+1}=\frac{R-\sum_{t=s}^{\infty}p_{s,t}\left(1-p_{t,t+1}\right)R^{s-t}}{\sum_{t=s}^{\infty}p_{s,t}\left(1-p_{t,t+1}\right)\left(1-R^{s-t}\right)}$$

$$=1+\frac{R-1}{\sum_{t=s}^{\infty}p_{s,t}\left(1-p_{t,t+1}\right)\left(1-R^{s-t}\right)}$$

Comparing this to the expression for $R_{fair}^{s,s+1}$, this is true iff:

$$\sum_{t=s+1}^{\infty}R^{s-t}p_{s,t}=\sum_{t=s}^{\infty}p_{s,t}\left(1-p_{t,t+1}\right)\frac{\left(1-R^{s-t}\right)}{R-1}\tag{4.4}$$

Using the fact that $p_{s,t}\cdot p_{t,t+1}=p_{s,t+1}$, the RHS of this expression simplifies as follows:

$$\begin{aligned}
\sum_{t=s}^{\infty} p_{s,t}\left(1-p_{t,t+1}\right)\frac{\left(1-R^{s-t}\right)}{R-1} &= \sum_{t=s}^{\infty} p_{s,t}\frac{\left(1-R^{s-t}\right)}{R-1} - \sum_{t=s}^{\infty} p_{s,t+1}\frac{\left(1-R^{s-t}\right)}{R-1} \\
&= \sum_{t=s}^{\infty} p_{s,t}\frac{\left(1-R^{s-t}\right)}{R-1} - \sum_{t=s+1}^{\infty} p_{s,t}\frac{\left(1-R^{s-t+1}\right)}{R-1} \\
&= p_{s,s}\frac{\left(1-R^{0}\right)}{R-1} + \sum_{t=s+1}^{\infty} p_{s,t}\left(\frac{1-R^{s-t}-\left(1-R^{s-t+1}\right)}{R-1}\right) \\
&= \sum_{t=s+1}^{\infty} R^{s-t} p_{s,t}\left(\frac{R-1}{R-1}\right) \\
&= \sum_{t=s+1}^{\infty} R^{s-t} p_{s,t}
\end{aligned}$$

which is precisely the LHS of (4.4). ■