Moral hazard in health insurance: How important is forward looking behavior?*

Aviva Aron-Dine, Liran Einav, Amy Finkelstein, and Mark Cullen†

July 2012

Abstract. We investigate whether individuals exhibit forward looking behavior in their response to the non-linear pricing common in health insurance contracts. Our empirical strategy exploits the fact that employees who join an employer-provided health insurance plan later in the calendar year face the same initial ("spot") price of medical care but a higher expected end-of-year ("future") price than employees who join the same plan earlier in the year. Our results reject the null of completely myopic behavior; medical utilization appears to respond to the future price, with a statistically significant elasticity of medical utilization with respect to the future price of -0.4 to -0.6. To try to quantify the extent of forward looking behavior, we develop a stylized dynamic model of individual behavior and calibrate it using our estimated behavioral response and additional data from the RAND Health Insurance Experiment. Our calibration suggests that the elasticity estimate may be substantially smaller than the one implied by fully forward-looking behavior, yet it is sufficiently high to have an economically significant effect on the response of annual medical utilization to a non-linear health insurance contract. Overall, our results point to the empirical importance of accounting for dynamic incentives in analyses of the impact of health insurance on medical utilization.

JEL classification numbers: D12, G22

Keywords: Health insurance, moral hazard, forward looking behavior, dynamic incentives

---

*We are grateful to David Molitor and James Wang for outstanding research assistance, and to Amitabh Chandra, Kate Ho, Jeff Liebman, and numerous seminar participants for helpful comments and suggestions. The Alcoa portion of the data were provided as part of an ongoing service and research agreement between Alcoa, Inc. and Stanford, under which Stanford faculty, in collaboration with faculty and staff at Yale University, perform jointly agreed-upon ongoing and ad hoc research projects on workers' health, injury, disability, and health care, and Mark Cullen serves as Senior Medical Advisor for Alcoa, Inc. We gratefully acknowledge support from the NIA (R01 AG032449), the National Science Foundation Grant SES-0643037 (Einav), the John D. and Catherine T. MacArthur Foundation Network on Socioeconomic Status and Health, and Alcoa, Inc. (Cullen), and the U.S. Social Security Administration through grant #5 RRC08098400-04-00 to the National Bureau of Economic Research as part of the SSA Retirement Research Consortium. The findings and conclusions expressed are solely those of the authors and do not represent the views of SSA, any agency of the Federal Government, or the NBER.

†Aron-Dine: Department of Economics, MIT, arondine@mit.edu; Einav: Department of Economics, Stanford University, and NBER, leinav@stanford.edu; Finkelstein: Department of Economics, MIT, and NBER, afink@mit.edu; Cullen: Department of Internal Medicine, School of Medicine, Stanford University, and NBER, mrcullen@stanford.edu.

1 Introduction

The size and rapid growth of the healthcare sector – and the pressure this places on public sector budgets – has created great interest among both academics and policymakers in possible approaches to reducing healthcare spending. On the demand side, the standard, long-standing approach to constraining healthcare spending is through consumer cost sharing in health insurance, such as deductibles and coinsurance. Not surprisingly therefore, there is a substantial academic literature devoted to trying to quantify how the design of health insurance contracts affects medical spending. These estimates have important implications for the costs of alternative health insurance contracts, and hence for the optimal design of private insurance contracts or social insurance programs.

One aspect of this literature that we find remarkable is the near consensus on the nature of the endeavor: the attempt to quantify the response of medical spending with respect to its (out-of-pocket) price to the consumer. Yet, health insurance contracts in the United States are highly non-linear, so trying to estimate the behavioral response to a single out-of-pocket price is, in most cases, not a well-posed exercise, as it begs the question "which price?". A typical private health insurance plan has a deductible, a coinsurance rate, and an out-of-pocket maximum (or "stop loss"). The consumer faces a price of 100% of medical expenditures until he has spent the deductible, at which point the marginal price falls sharply to the coinsurance rate (typically around 10-20%), and then falls to zero once out-of-pocket expenditures have reached the stop-loss amount. Public health insurance programs, such as Medicare, also involve non-linear schedules, including occasionally schedules in which the marginal price rises over some expenditure range and then falls again (as in the famous "doughnut hole" in Medicare Part D prescription drug coverage).

In the context of such non-linear budget sets, trying to characterize an insurance policy by a single price could produce very misleading inferences. For example, one cannot extrapolate from estimates of the effect of coinsurance on health spending to the effects of introducing a high-deductible health insurance plan without knowing how forward looking individuals are in their response to health insurance coverage. A completely myopic individual would respond to the introduction of a deductible as if his "price" has sharply increased to 100%, whereas a fully forward looking individual with annual health expenditures that are likely to exceed the new deductible would experience little change in the effective marginal price of care and therefore might not change his behavior much.[1] Understanding how medical spending responds to the design of health insurance contracts therefore requires that we understand how consumers account for the non-linear budget schedule they face

[1] Indeed, once one accounts for the non-linear contract design, even characterizing which insurance contract would provide greater incentives to economize on medical spending becomes a complicated matter. Consider, for example, two plans with a coinsurance arm that is followed by an out-of-pocket maximum of $5,000. Imagine that Plan A has a 10% coinsurance rate and plan B has a 50% coinsurance rate. Which plan would induce less spending? The naive answer would be that Plan B is less generous and would therefore lead to lower medical utilization. Yet, the answer depends on the distribution of medical spending without insurance, as well as on how forward looking individuals are. For example, an individual who suffers a compound fracture early in the coverage period and spends $10,000 on a surgery would effectively obtain full insurance coverage for the rest of the year under Plan B, but would face a 10% coinsurance rate (with a remaining $4,000 stop loss) under Plan A. We would therefore expect this individual to have greater medical utilization under Plan B.

in making their medical consumption decisions. A fully rational, forward-looking individual who is not liquidity constrained should recognize that the "spot" price applied to a particular claim is not relevant; this nominal price should not affect his consumption decisions. Rather, the decision regarding whether to undertake some medical care should be a function only of the end-of-year price.

In this paper, we therefore investigate whether the common practice of summarizing health insurance contracts with a single price is a reasonable approximation by examining whether and to what extent individuals respond to the expected end-of-year price, or "future price," of medical care. We do so in the context of employer-provided health insurance in the United States, which is the source of over 85% of private health insurance coverage. Assessing whether individuals respond to the future price is empirically challenging, which may explain why there has been relatively little work on this topic. The key empirical difficulty arises because the spot price and the future price often vary jointly. A low spending individual faces both a high spot price (because all his spending falls below the deductible) and a high expected end-of-year price (because he does not expect to hit the deductible), while the opposite is true for a high spending individual. Similarly, the types of variation that have most often been used to estimate the impact of health insurance on medical spending – such as variation in deductibles or coinsurance rates – will change the spot price and the future price jointly. This makes it challenging to identify whether individuals respond to the future price without a tightly specified model of expectation formation, which in turn raises concerns about the extent to which any elasticity estimates are driven by these modeling assumptions.

The primary empirical exercise in this paper addresses this challenge by identifying situations in which individuals face the same spot price for their consumption decision, but have substantially different expected end-of-year prices. The key insight behind our empirical strategy is that, as a result of certain institutional features of employer-provided health insurance in the United States, individuals who join the same deductible plan in different months of the year initially face the same spot price, but different expected end-of-year prices. Employer-provided health insurance resets every year, typically on January 1. When new employees join a firm in the middle of the year, they obtain coverage for the remainder of the year. While their premiums are pro-rated, deductible amounts are fixed at their annual level. As a result, all else equal, the expected end-of-year price is increasing with the join month over the calendar year; individuals who join a plan later in the year have fewer months to spend past the deductible.

We use this feature in order to test for forward looking behavior in the response to health insurance contracts. In other words, we test the null of completely myopic behavior, which we define as consumption decisions that depend only on the spot price. We do so by comparing initial medical utilization across individuals who join the same deductible PPO plan in different months of the year. If individuals are forward looking in their healthcare consumption decisions, an individual who joins the plan earlier in the calendar year should (initially) spend more than an otherwise identical individual who joins the same plan later in the calendar year. By contrast, if individuals are myopic, the initial spending of an individual who joins the plan earlier should be the same as the initial spending of the individual who joins the same plan later. To account

for potential confounders, such as seasonality in healthcare spending, we use patterns of initial utilization by join month for individuals who join the same PPO plan with no deductible, in which the future price hardly varies over the course of the year. To operationalize this strategy empirically, we draw on data from several large employers with information on their plan details as well as their employees' plan choices, demographics, and medical claims.

All of our analyses are thus within a set of employees who join a plan mid-year. They rely on comparisons of initial utilization patterns by join month for those who join a deductible PPO plan mid-year relative to the initial utilization patterns by join month for those who join a no deductible PPO plan mid-year. Moreover, although we draw on data from several firms, our analysis relies on only within-firm variation in the pattern of initial utilization across plans by join month. Figure 1 (whose construction we describe in much more detail later in the paper) provides one way of summarizing this empirical exercise. It shows – separately for each firm – that as employees join a plan later in the year (and the expected end-of-year price rises for those in the deductible plan) initial medical utilization in the deductible plan tends to fall (i.e., initial spending falls or time to first claim increases), both in absolute terms as well as relative to the corresponding pattern in the no-deductible plan.

We note that individuals may fail to exhibit forward-looking behavior not only because they are myopic but also if they are liquidity constrained or lack an understanding of their future budget constraint. If we had failed to reject the null of completely myopic behavior, we would have been unable to distinguish which of these factors was behind our result. In practice, however, we reject the null and estimate that conditional on the spot price of medical care, individuals who face a higher future price consume statistically significantly less (initial) medical care. It therefore appears that individuals understand something about the nature of their dynamic budget constraint and make their healthcare consumption decisions with at least some attention to forward-looking considerations.

In the last section of the paper we attempt to move beyond testing the null of complete myopia and toward quantifying the extent of forward looking behavior. We estimate that a ten cent increase in the future price (for a dollar of medical spending) is associated with a 6 to 8 percent decline in initial medical utilization. This implies an elasticity of initial medical utilization with respect to the future price of -0.4 to -0.6. To provide an economic interpretation of this estimate, we develop a stylized dynamic model in which utilization behavior in response to medical shocks depends on both the underlying willingness to substitute between health and residual income and the degree of forward looking behavior. We draw on additional data from the RAND Health Insurance Experiment to calibrate the model, and use the calibrated model to assess the extent of forward looking behavior implied by our estimates of the response of initial medical utilization to the future price. On the spectrum between full myopia (individuals respond only to the spot price) and textbook forward looking behavior (individuals respond only to the future price), our calibration results generally suggest that individuals' behavior is much closer to the former. Nonetheless, we illustrate that the degree of forward looking behavior we find still has a substantial effect on the response of annual medical spending to health insurance contracts relative to the spending response

that would be predicted under either completely myopic or completely forward looking behavior. Thus, failing to account for dynamic considerations can greatly alter the predicted impact of non-linear health insurance contracts on annual medical expenditures.

Our paper links to the large empirical literature that tries to estimate moral hazard in health insurance, or the price sensitivity of demand for medical care. As already mentioned, much of this literature tries to estimate a demand elasticity with respect to a single price,[2] although different studies consider a different "relevant" price to which individuals are assumed to respond. For example, the famous RAND elasticity of -0.2 is calculated assuming individuals respond only to the spot price (Manning et al., 1987; Keeler and Rolph, 1988; Zweifel and Manning, 2000), while more recent estimates have assumed that individuals respond only to the expected end-of-year price (Eichner, 1997) or to the actual (realized) end-of-year price (Eichner, 1998; Kowalski, 2010). Our findings highlight the importance of thinking about the entire budget set rather than about a single price; this point was emphasized in some of the early theoretical work on the impact of health insurance on health spending (Keeler, Newhouse, and Phelps, 1977; Ellis, 1986) but until recently has rarely been incorporated into empirical work. Several papers on the impact of health insurance on medical spending – Ellis (1986), Cardon and Hendel (2001), and more recently Kowalski (2011), Marsh (2011), and our own work (Einav et al., 2011) – explicitly account for the non-linear budget set, but a (fully forward-looking) behavioral model is assumed, rather than tested.[3]

Outside of the context of health insurance, a handful of papers address the question of whether individuals respond at all to the non-linearities in their budget set, and which single price may best approximate the non-linear schedule to which individuals respond. This is the focus of Liebman and Zeckhauser (2004), Feldman and Katuscak (2006), and Saez (2010) in the context of the response of labor supply to the progressive income tax schedule, and of Borenstein (2009) and Ito (2010) in the context of residential electricity utilization. In most of these other contexts, as well as in our own previous work on moral hazard in health insurance (Einav et al., 2011), the analysis of demand in the presence of a non-linear pricing schedule is static. This is partly because in most non-health contexts information about intermediate utilization levels (within the billing or tax cycle) is not easy to obtain (for both consumers and researchers) and partly because dynamic modeling often introduces unnecessary complications in the analysis. In this sense, our current study – utilizing the precise timing of medical utilization within the contract year – is virtually unique within this literature in its explicit focus on the dynamic aspect of medical utilization, and its explicit account of expectation formation.[4]

Forward looking decision making plays a key role in many economic problems, and interest in

[2] See Chandra, Gruber, and McKnight (2007) for a recent review of this literature and its estimates.

[3] Non-linear pricing schedules are not unique to health insurance. Indeed, a large literature, going back at least to Hausman (1985), develops methods that address the difficulties that arise in modeling selection and utilization under non-linear budget sets, and applies these methods to other setting in which similar non-linearities are common, such as labor supply (Burtless and Hausman, 1978; Blundell and MaCurdy, 1999; Chetty et al., 2011), electricity utilization (Reiss and White, 2005), or cellular phones (Grubb and Osborne, 2009; Yao et al., 2011).

[4] An exception in this regard is Keeler and Rolph (1988), who, like us, test for forward looking behavior in health insurance contracts (but use a different empirical strategy and reach a different conclusion).

the extent of forward looking behavior is therefore quite general. From this perspective, a closely related work to ours is Chevalier and Goolsbee's (2009) investigation of whether durable goods consumers are forward looking in their demand for college textbooks (they find that they are). Despite the obvious difference in context, their empirical strategy is similar to ours. They use the fact that static, spot incentives remain roughly constant (as pricing of textbook editions doesn't change much until the arrival of new editions), while dynamic incentives (the expected time until a new edition is released) change. A slightly cleaner aspect of our setting is that the constant spot prices and varying dynamic incentives are explicitly stipulated in the coverage contract rather than empirical facts that need to be estimated from data.

The rest of the paper proceeds as follows. Section 2 sketches a simple, stylized model of medical care utilization that is designed to provide intuition for the key concepts and our empirical strategy; the model serves as both a guide to some of our subsequent empirical choices, and as a framework that we use to benchmark the extent of forward looking behavior we estimate. In Section 3 we test for forward looking behavior. We start by describing the basic idea and the data we obtained to implement it, and then present the results. In Section 4 we calibrate the model from Section 2 to try to quantify the extent to which individuals are forward looking. Section 5 concludes.

2 A simple model

Consider a model of a risk-neutral forward-looking individual who faces uncertain medical expenditure, and is covered by a contract of (discrete) length T and deductible D.[5] That is, the individual pays all his expenditures out of pocket up to the deductible level D, but any additional expenditure is fully covered by the insurance provider.

The individual's utility is linear and additive in health and residual income, and we assume that medical events that are not treated are cumulative and additively separable in their effect on health. Medical events are given by a pair (θ, ω), where $\theta > 0$ denotes the total expenditure (paid by either the individual or his insurance provider) required to treat the event, and $\omega > 0$ denotes the (monetized) health consequences of the event if left untreated. We assume that individuals need to make a discrete choice whether to fully treat an event or not; events cannot be partially treated. We also assume that treated events are "fully" cured, and do not carry any other health consequences. Thus, conditional on an event (θ, ω), the individual's flow utility is given by

$$u(\theta, \omega; d) = \begin{cases} -min\{\theta, d\} & if\ treated \\ -\omega & if\ not\ treated \end{cases} \tag{1}$$

where $min\{\theta, d\}$ is the out-of-pocket cost associated with expenditure level θ, which is a function of d, the amount left to satisfy the deductible.

[5] Assuming risk neutrality in the context of an insurance market may appear an odd modeling choice. Yet, it makes the model simpler and more tractable and makes no difference for any of the qualitative insights we derive from the model.

Medical shocks arrive with a per-period probability λ, and when they arrive they are drawn independently from a distribution $G(\theta, \omega)$. Given this setting, the only choice individuals make is whether to treat or not treat each realized medical event. Optimal behavior can be characterized by a simple finite horizon dynamic problem. The two state variables are the time left until the end of the coverage period which we denote by t, and the amount left until the entire deductible is spent which we denote by d. The value function $v(d, t)$ represents the present discounted value of expected utility along the optimal treatment path. Specifically, the value function is given by the solution to the following Bellman equation:

$$v(d,t) = (1-\lambda)\delta v(d,t-1) + \lambda \int max \left\{ \begin{array}{l} -min\{\theta, d\} + \delta v(max\{d-\theta, 0\}, t-1), \\ -\omega + \delta v(d, t-1) \end{array} \right\} dG(\theta, \omega), \quad (2)$$

with terminal conditions of $v(d, 0) = 0$ for all d. If a medical event arrives, the individual treats the event if the value from treating, $-min\{\theta, d\} + \delta v(max\{d-\theta, 0\}, t-1)$, exceeds the value obtained from not treating, $-\omega + \delta v(d, t-1)$.

The model implies simple and intuitive comparative statics: the treatment of a medical event is more likely when the time left on the contract, t, is higher and the amount left until the deductible is spent, d, is lower. This setting nests a range of possible behaviors. For example, "fully" myopic individuals ($\delta = 0$) would not treat any shock as long as the immediate negative health consequences of the untreated shock, ω, are less than the immediate out-of-pocket expenditure costs associated with treating that shock, $\min\{\theta, d\}$. Thus, if $\theta < d$, fully myopic individuals ($\delta = 0$) will not treat if $\omega < \theta$. By contrast, "fully" forward looking individuals ($\delta \approx 1$) will not treat shocks if the adverse health consequences, ω, are less than the expected end-of-year cost of treating this illness, which is given by $fp \cdot \theta$, where fp (for "future price") denotes the expected end-of-year price of medical care, which is the relevant price for a "fully" forward looking individual in deciding whether to consume care today. Thus, if $\theta < d$, fully forward looking individuals will not treat if $\omega < fp \cdot \theta$. That is, while fully myopic individuals consider the current, "spot", or nominal price of care (which in our example is equal to one), fully forward looking individuals only care about the future price.

To illustrate the implications of the model that will serve as the basis of our empirical strategy, we solve the model for a simple case, where we assume that $\lambda = 0.2$ and that medical events are drawn uniformly from a two-point support of $(\theta = 50, \omega = 50)$ and $(\theta = 50, \omega = 45)$. We use two different deductible levels (of 600 and 800) and up to 52 periods (weeks) of coverage. Figure 2 presents some of the model's implications for the case of $\delta = 1$. It uses metrics that are analogous to the empirical objects we later use in the empirical exercise. The top panel presents the expected end-of-year price of the individual as we change the deductible level and the coverage horizon. The expected end-of-year price in this example is $1 - \Pr(hit)$, where $\Pr(hit)$ is the fraction of individuals who hit the deductible by the end of the year. Individuals are, of course, more likely to hit the deductible as they have more time to do so or as the deductible level is lower. This ex-ante probability of hitting the deductible determines the individual's expectations about his end-of-year price. This future price in turn affects a forward looking individual's willingness to treat medical events. The bottom panel of Figure 2 presents the (cumulative) expected spending

over the initial three months (12 weeks). Given the specific choice of parameter values, expected spending over the initial 12 periods is at least 60 (due to the per-period 0.1 probability of a medical event ($\theta = 50, \omega = 50$) that would always be treated) and at most 120 (if all medical events are treated).

The key comparative static that is illustrated by Figure 2 – and that will form the basis of our empirical work – is how the expected end-of-year price (and hence initial spending by a forward looking individual) varies with the coverage horizon. For a given deductible, the expected end-of-year price is increasing as the coverage horizon declines (top panel of Figure 2) and therefore, for a forward looking individual, expected *initial* spending also declines as the coverage horizon declines (bottom panel of Figure 2). Specifically, when the coverage horizon is long enough and the deductible level low enough, forward looking individuals expect to eventually hit the deductible and therefore treat all events, so expected spending is 120. However, as the horizon gets shorter there is a greater possibility that the deductible would not get exhausted by the end of the year, so the end-of-year price could be 1 (rather than zero), thus making forward looking individuals not treat the less severe medical events of ($\theta = 50, \omega = 45$).

The graphs also illustrate how the spot price of current medical care misses a great deal of the incentives faced by a forward looking individual. In the bottom panel of Figure 2 we see a fully forward looking individual's initial medical utilization (i.e., spending in the first 12 weeks) varying greatly with the coverage horizon despite a spot price that is always one. By contrast, for the cases we consider, a fully myopic individual ($\delta = 0$) who only responds to the spot price has expected 12-week spending of 60, regardless of the coverage horizon t (see bottom panel).[6] Likewise, the expected three-month spending of individuals in a no-deductible plan does not vary with the coverage horizon, regardless of their δ, since the expected end-of-year price does not vary with the coverage horizon.

Finally, we note that while we have referred to δ as a measure of how "forward looking" the individual is, in practice a variety of different factors can push δ below 1 and induce a behavioral response to the current, "spot" price. These factors include not only myopia but also liquidity constraints (e.g., Adams, Einav, and Levin, 2009) and salience (e.g., Chetty and Saez, 2009; Liebman and Luttmer, 2011). Our research strategy does not distinguish between these, nor is it necessary to do so for predicting how spending will respond to changes in a non-linear budget set. However, these different sources that may affect behavior can be important for forecasting the effects of alternative public policy interventions or for extrapolating our results to alternative populations. We return to these issues briefly in the conclusions.

[6] A fully myopic individual ($\delta = 0$) would (like the fully forward looking individual) always treat ($\theta = 50, \omega = 50$) shocks but as long as he is still in the deductible range would never treat ($\theta = 50, \omega = 45$) shocks. Given this behavior, with a 600 or 800 deductible, there is a zero probability that the deductible would be reached within the first 12 weeks.

3 Testing for forward looking behavior

3.1 Basic idea

To test whether individuals exhibit forward looking behavior in their behavioral response to their health insurance contract, we design a test for whether individuals respond to the future price of medical consumption in a setting in which similar individuals face the same spot price (i.e., the nominal price at the time they make their medical consumption decision) but different future prices. In such a situation, we can test whether medical utilization changes with the future price, holding the spot price fixed, and interpret a non-zero coefficient as evidence of forward looking behavior and as a rejection of the null of complete myopia.

The central empirical challenge therefore is to identify individuals who face the same spot price but different future prices for medical consumption. Our novel observation is that the institutional features of employer-provided health insurance in the United States provide such variation. Specifically, we use the fact that unlike other lines of private insurance (e.g., auto insurance or home insurance), the coverage period of employer-provided health insurance is not customized to individual employees. This presumably reflects the need for synchronization within the company, such as benefits sessions, open enrollment periods, and tax treatment. Therefore, (annual) coverage begins (and ends, unless it is terminated due to job separation) at the same date – typically on January 1 – for almost all employees. Although all employees can choose to join a new plan for the subsequent year during the open enrollment period (typically in October or November), there are only two reasons employees can join a plan in the middle of the year: either they are new hires or they have a qualifying event that allows them to change plans in the middle of the year.[7] In order to transition new employees (and occasionally existing employees who have a qualifying event) into the regular cycle, the common practice is to let employees choose from the regular menu of coverage options, to pro-rate linearly the annual premium associated with their choices, but to maintain constant (at its *annual* level) the deductible amount. As a result, individuals who are hired at different points in the year, but are covered by the same (deductible) plan, face the same spot price (of one) but different future prices. Thus, as long as employees join the company at different times for reasons that are exogenous to their medical utilization behavior, variation in hire date (or in the timing of qualifying events) generates quasi-experimental variation in the future price that allows us to test for forward looking behavior.

To illustrate, consider two identical employees who select a plan with an \$800 (annual) deductible. The first individual is hired by the company in January and the second in July. The difference in their incentives is analogous to the simple model presented in Figure 2. Individuals who join in a later month during the year have a shorter coverage horizon t until coverage resets (on January 1). Individuals who join early in the year have a longer coverage horizon. The early joiners are therefore more likely to hit their deductible by the time their coverage resets. Therefore, as in

[7] Qualifying events include marriage, divorce, birth or adoption of a child, a spouse's loss of employment, or death of a dependent.

the top panel of Figure 2, early joiners have a lower expected end-of-year price. As in the bottom panel of Figure 2, if individuals are forward looking, then early joiners have a greater incentive to utilize medical care upon joining the plan. Crucially, just after they get hired, both January and July joiners have yet to hit their deductible, so their spot price is (at least initially) the same. Thus, differences in (initial) spending cannot be attributed to differences in spot prices, and therefore must reflect dynamic considerations. By contrast, as Figure 2 also illustrates, if individuals are completely myopic (or join a plan with no deductible so that the expected end-of-year price does not vary with the month they join the plan), initial utilization will not vary for the early and later joiners.

3.2 Data

Data construction With this strategy in mind, we obtained claim-level data on employer-provided health insurance in the United States. We limited our sample to firms that offered at least one PPO plan with a deductible (which would generate variation in expected end-of-year price based on the employee's join month, as in the top panel of Figure 2) and at least one PPO plan with no deductible. The relationship between initial utilization and join month in the no-deductible plan is used to try to control for other potential confounding patterns in initial medical utilization by join month (such as seasonal flu); in a typical no-deductible plan, the expected end-of-year price is roughly constant by join month, so – absent confounding effects that vary by join month – initial medical utilization of employees covered by a no-deductible plan should not systematically vary with join month (bottom panel of Figure 2).

The data come from two sources. The first is Alcoa, Inc., a large multinational producer of aluminum and related products. We have four years of data (2004-2007) on the health insurance options, choices, and medical insurance claims of its employees (and any insured dependents) in the United States. We study the two most common health insurance plans at Alcoa, one with a deductible for in-network expenditure of $250 for single coverage ($500 for family coverage), and one with no deductible associated with in-network spending. While Alcoa employed (and the data cover) about 45,000 U.S.-based individuals every year, the key variation we use in this paper is driven by mid-year plan enrollment by individuals not previously insured by the firm, thus restricting our analysis to only about 7,000 unique employees (over the four years) that meet our sample criteria.[8] Of the employees at Alcoa who join a plan mid-year and did not previously have insurance at Alcoa that year, about 80% are new hires, while the other 20% are employees who were at Alcoa but uninsured at the firm, had a qualifying event that allowed them to change plans in the middle of the year, and chose to switch to Alcoa-provided insurance.

The Alcoa data are almost ideal for our purposes, with the important exception of sample size. Ex ante, sample size was a key concern given the large variation in medical spending across individuals. To increase statistical power we examined the set of firms (and plans) available through

[8] We restrict our analysis to employees who are not insured at the firm prior to joining a plan in the middle of the year because if individuals change plans within the firm (due to a qualifying event), the deductible would not reset.

the National Bureau of Economic Research's (NBER) files of Medstat's MarketScan database. The data on plan choices and medical spending are virtually identical in nature and structure across the three firms (indeed, Alcoa administers its health insurance claims via Medstat); they include coverage and claim-level information from an employer-provided health insurance context, provided by a set of (anonymous) large employers.

We selected two firms that satisfied our basic criteria of being relatively large and offering both deductible and no-deductible options to their employees. Each firm has about 60,000 employees who join one of these plans in the middle of the year over the approximately six years of our data. This substantially larger sample size is a critical advantage over the Alcoa data. The disadvantages of these data are that we cannot tell apart new hires from existing employees who are new to the firm's health coverage (presumably due to qualifying events that allow them to join a health insurance plan in the middle of the year), we cannot distinguish between in-network and out-of-network spending, there is less demographic information on the employees, and the coinsurance rate for one of the plans in one of the firms is not known.

Because employers in MarketScan are anonymous (and we essentially know nothing about them), we will refer to these two additional employers as firm B and firm C. We focus on two plans offered by firm B. We have five years of data (2001-2005) for these plans, during which firm B offered one plan with no in-network deductible and one plan that had a $150 ($300) in-network single (family) deductible. The data for firm C are similar, except that the features of the deductible plan have changed slightly over time. We have seven years of data for firm C (1999-2005), during which the firm continuously offered a no-deductible plan (in-network) alongside a plan with a deductible. The deductible amount increased over time, with a single (family) in-network deductible of $200 ($500) during 1999 and 2000, of $250 ($625) during 2001 and 2002, and $300 ($750) during 2004 and 2005.

Table 1 summarizes the key features of the plans (and their enrollment) that are covered by our final data set. In all three firms, we limit our sample to employees who join a plan between February and October, and who did not have insurance at the firm immediately prior to this join date. We omit employees who join in January for reasons related to the way the data are organized that make it difficult to tell apart new hires who join the firm in January from existing employees. We omit employees who join in November or December because, as we discuss in more detail below, we use data from the first three months after enrollment to construct our measures of "initial" medical utilization.[9] Table 1 also summarizes, by plan, the limited demographic information we observe on each covered employee, namely the type of coverage they chose (family or single), and the employee's gender, age, and enrollment month.[10]

[9] In practice we only observe the join month rather than the join date. Thus, throughout the paper, when we speak of the "first three months" after enrollment, more precisely we are using the first 2-3 months after enrollment. As long as the join day within the month is similar across months, the average time horizon should also be similar by join month.

[10] In each firm we lose roughly 15 to 30 percent of new plan joiners because of some combination of missing information about the employee's plan, missing plan details, or missing claims data (because the plan is an HMO or a partially or fully capitated POS plan).

Measuring the expected end-of-year price Table 2 describes the key variation we use in our empirical analysis. For each plan, we report the expected end-of-year price as a function of the time within the year an employee joined the plan.[11] Specifically, we define the expected end-of-year price, or future price, fp, as

$$fp_{jm} = 1 - \Pr(hit_{jm}), \tag{3}$$

where $\Pr(hit_{jm})$ is the probability an employee who joins plan j in month m will hit (i.e., spend more than) the in-network deductible by the end of the year; we calculate $\Pr(hit)$ as the fraction of employees in a given plan and join month who have spent more than the in-network deductible by the end of the year.[12] For example, consider a plan with a \$500 deductible and full coverage for any medical expenditures beyond the deductible. If 80% of the employees who joined the plan in February have hit the deductible by the end of the year, the expected end-of-year price would be $0.8 \cdot 0 + 0.2 \cdot 1 = 0.2$. If only 40% of the employees who joined the plan in August have hit the deductible by the end of the year, their expected end-of-year price would be $0.4 \cdot 0 + 0.6 \cdot 1 = 0.6$. Thus, the future price is the average (out-of-pocket) end-of-year price of an extra dollar of in-network spending. It is a function of one's plan j, join month m, and the annual spending of all the employees in one's plan and join month.[13]

Table 2 summarizes the average future price for each plan based on the quarter of the year in which one joins the plan. For plans with no deductible ($A0$, $B0$, and $C0$), the future price is mechanically zero (since everyone "hits" the zero deductible), regardless of the join month. For deductible plans, however, the future price varies with the join month. Only a small fraction of the individuals who join plans late in the year (August through October) hit their deductible, so their future price is greater than 0.8 on average. In contrast, many more employees who join a deductible plan early in the year (February to April) hit their deductible, so for such employees the future price is just over 0.5. Thus, early joiners who select plans with a deductible face an average end-of-year price that is about 30 percentage points lower than the end-of-year price faced by late joiners. Yet, initially (just after they join) both types of employees have yet to hit their deductible, so they all face a spot price of one. Differences in initial spending between the groups therefore plausibly reflects their dynamic response to the future price. This baseline definition of the future price – the fraction of employees who join a given plan in a given month whose spending does not exceed the in-network deductible by the end of the calendar year – will be used as the key right hand variable in much of our subsequent empirical work.

[11] In this and all subsequent analyses we pool the three different deductible plans in firm C which were offered at different times over our sample period.

[12] We calculate $\Pr(hit)$ separately for employees with individual and family coverage (since both the deductible amount and spending patterns vary with the coverage tier), and therefore in all of our analyses fp varies with coverage tier. However, for conciseness, in the tables we pool coverage tiers and report the (weighted) average across coverage tiers within each plan.

[13] To the extent that individuals have private information about their future health spending, and thus about their expected end-of-year price, our "average" measure would introduce measurement error and would bias our results toward zero. For testing, this bias would make us less likely to find evidence for forward looking behavior. For quantification, as will be described later, we apply an estimation strategy that tries to accommodate this concern.

Our baseline measure of the future price abstracts from several additional characteristics of the plans, which are summarized in Appendix Table A1. First, it ignores any coinsurance features of the plans. Plans A0, A1, and C1-C3 all have a 10% coinsurance rate, while plans B0 and C0 have a zero coinsurance rate. The coinsurance rate for plan B1 is unknown (to us). Second, we use only the in-network plan features and assume that all spending occurs in network. In practice, each plan (including the no-deductible plan) has deductibles and higher consumer coinsurance rates for medical spending that occurs out of network.

There are two consequences of these abstractions, both of which bias any estimated impact of the future price on behavior toward zero. First, abstracting from these features introduces measurement error into the future price. Second, our analysis assumes that for the no-deductible plans there is no variation in the future price for employees who join in different months (i.e., the spot price and the future price are always the same). In practice, both a positive in-network coinsurance rate (prior to the stop-loss) and the existence of out-of-network deductibles in all of the no-deductible (in-network) plans mean that the future price also increases with the join month for employees in the no-deductible plans. In the robustness section below we show that accounting for these additional features – to the extent we are able to – makes little quantitative difference to either our measurement of the future price or its estimated effect.

A final point worth noting about our definition of the future price is that it is constructed based on the observed spending patterns of people who join a specific plan (and coverage tier) in a specific month. For forward looking individuals, this spending may of course be influenced by the future price. As we discuss in more detail below, this is not a problem for testing the null of complete myopia (because under this null spending is not affected by the future price). Yet, for quantifying the extent of forward looking behavior in Section 4 we will implement an instrumental variable strategy designed to purge the calculated future price of any endogenous spending response.

3.3 Estimating equations and results

Patterns of initial utilization by plan and join month We proxy for "initial" utilization with two alternative measures. The first is a measure of the time (in days) to the first claim, while the second is a measure of total spending (in dollars) over some initial duration (we will use three months). In both cases, the measures of utilization encompass the utilization of the employee and any covered dependents.

Average three month spending in our sample is about $600. It is zero for about 42% of the sample. Since time to first claim is censored at as low a value as 92 days (for individuals who join in October), we censor time to first claim at 92 for all the individuals (regardless of join month) who have their first claim more than 92 days after joining the firm's coverage. The average time to first claim for the remaining 58% of the individuals is 35 days, so with 42% of the sample censored at 92 days, the sample average for the censored variable is 58 days.

Figure 1 reports, for each firm separately, the pattern of initial medical utilization by join month for the deductible plan and the no deductible plan. The left hand panel reports the results for initial

three month spending; the right hand panel for time to first claim. These statistics already indicate what appears to be a response to dynamic incentives. For the deductible plan, initial medical spending generally tends to fall (and time to first claim to rise) with join month (and expected end of year price), while there is generally no systematic relationship between join month and initial medical utilization (by either measure) in the corresponding no deductible plan. As illustrated in the bottom panel of Figure 2, this is exactly the qualitative pattern one would expect from forward looking individuals.

We operationalize this analysis a little more formally by regressing the measures of initial utilization on join month. A unit of observation is an employee e who joins health insurance plan j during calendar month m. As mentioned, we limit attention to employees who join new plans between February and October, so $m \in \{2, ...10\}$. As a result, all of our comparisons of patterns of initial utilization by join month, both within and across plans, are among a set of employees who are new to their plan.

The simplest way by which we can implement our strategy is to look within a given health plan that has a positive deductible associated with it and regress a measure of initial medical utilization y_e on the join month m_e and possibly a set of controls x_e, so that:

$$y_e = \beta_j m_e + x_e' \gamma + u_e. \tag{4}$$

Absent any confounding influences of join month on y_e, we would expect an estimate of $\beta_j = 0$ for deductible plans if individuals are fully myopic ($\delta = 0$) and $\beta_j < 0$ for spending ($\beta_j > 0$ for time to first claim) if individuals are not ($\delta > 0$). We include an additional covariate for whether the employee has family (as opposed to single) coverage to account for the fact that the deductible varies within a plan by coverage tier (see Table 1) and that there naturally exist large differences in average medical utilization in family vs. single coverage plans.

For our analysis of initial spending, our baseline dependent variable is $\log(s + 1)$, where s is total medical spending (in dollars) by the employee and any covered dependents during their first three months in the plan. Given that medical utilization is highly skewed, the log transformation helps in improving precision and reducing the effect of outliers.[14] An added attraction of the log specification is that it facilitates comparison of the results to those from our analysis of time to first claim. For the latter analysis, we use a Tobit specification on $\log(time)$, where $time$ measures the time to first claim (in days) by the employee and any covered dependents; the Tobit is used to account for the censoring at 92 days described above. We explore alternative functional forms for both dependent variables below.

Columns (1) and (3) of Table 3 report results from estimating equation (4) on these two dependent variables, separately for each plan. The key right-hand-side variable is the join month, enumerated from 2 (February) to 10 (October). In plans that have a deductible ($A1$, $B1$, and $C1$-

[14] While conceptually a concave transformation is therefore useful, we have no theoretical guidance as to the "right" functional form; any transformation therefore (including the one we choose) is ad hoc, and we simply choose one that is convenient and easy to implement. We note however that Box-Cox analysis of the $s + 1$ variable suggests that a log transformation is appropriate.

$C3$), dynamic considerations would imply a negative relationship between join month and initial spending and positive relationship between join month and time to first claim. The results show exactly this qualitative pattern.

Patterns of initial utilization by join month for deductible vs. no-deductible plan If seasonality in medical utilization is an important factor, it could confound the interpretation of the estimated relationship that we have just discussed as a test for the null of full myopia. For example, if spending in the spring is greater than spending in the summer due to, say, seasonal flu, then we may incorrectly attribute the decline in "spot" utilization for late joiners as a response to dynamic incentives. To address this concern (and other possible confounding differences across employees who join plans at different months of the year), we use as a control group employees within the same firm who join the health insurance plan with no deductible in different months. As discussed earlier, such employees are in a plan in which the spot price and future price are (roughly) the same so that changes in their initial utilization over the year (or lack thereof) provides a way to measure and control for factors that influence initial utilization by join month that are unrelated to dynamic incentives.

Columns (1) and (3) of Table 3, discussed earlier, also show the plan-level analysis of the relationship between initial medical utilization and join month for the no-deductible plans ($A0$, $B0$, and $C0$). The coefficient on join month for the no-deductible plans tends to be much smaller than the coefficient for the deductible plan in the same firm (and is often statistically indistinguishable from zero). This suggests that the difference-in-difference estimates of the pattern of spending by join month in deductible plans relative to the analogous pattern in no-deductible plans will look very similar to the patterns in the deductible plans. Indeed, this is what we find, as reported in columns (2) and (4) of Table 3, which report this difference-in-difference analysis in which the no-deductible plan (within the same firm) is used to control for the seasonal pattern of initial utilization by join month in the "absence" of dynamic incentives. Specifically, the difference-in-differences specification is

$$y_e = \beta' m_e D_j + \mu_j + \tau_m + x_e' \gamma' + \upsilon_e, \tag{5}$$

where μ_j are plan fixed effects, τ_m are join-month fixed effects, and D_j is a dummy variable that is equal to one when j is a deductible plan. The "plan fixed effects" (the μ_j's) include separate fixed effects for each plan by coverage tier (family or single) since the coverage tier affects the deductible amount (see Table 1). Again, our coefficient of interest is β', where $\beta' = 0$ would be consistent with the lack of response to dynamic incentives (i.e., full myopia) while $\beta' < 0$ (for spending; or $\beta' > 0$ for time to first claim) implies that the evidence is consistent with forward looking behavior. Since we are now pooling results across plans (deductible and no-deductible plans), the parameter of interest β' no longer has a j subscript.

The results in Table 3 indicate that, except at Alcoa where we have much smaller sample sizes, the difference-in-difference estimates for each firm are all statistically significant and with the sign that is consistent with dynamic considerations. For example, in Firm B we find that enrollment a month later in a plan with a ($150 or $300) deductible relative to enrollment a month later in a

plan with no deductible is associated with an 8% decline in medical expenditure during the first three months, and a 3% increase in the time to first claim. In Firm C these numbers are a 2% decline and a 2% increase, respectively.

Of course, employees who self select into a no-deductible plan are likely to be sicker and to utilize medical care more frequently than those employees who select plans with a deductible (due to both selection and moral hazard effects). Indeed, Table 1 shows that there are, not surprisingly, some observable differences between employees within a firm who choose the no-deductible option instead of the deductible option. Our key identifying assumption is that while initial medical utilization may differ on average between employees who join deductible plans and those who join no-deductible plans, the within-year pattern of initial utilization by join month does not vary based on whether the employee joined the deductible or no-deductible plan except for dynamic incentives. In other words, we assume that any differences in initial utilization between those who join the no-deductible plan and the deductible plan within a firm can be controlled for by a single (join month invariant) dummy variable. We return to this below, when we discuss possible threats to this identifying assumption and attempt to examine its validity.

Testing the relationship between expected end-of-year price and initial utilization In order to provide an economic interpretation to the parameter of interest, it is useful to convert the key right-hand-side variable, join month (m_e), into a variable that is closer to the underlying object of interest: the expected end-of-year price. We therefore start by analyzing variants of the single-plan analysis (equation (4)) and the difference-in difference analysis (equation (5)) in which we replace the join month variable (m_e) with the future price variable fp defined earlier (recall equation (3) for a definition, and Table 2 for summary statistics). The estimating equations are thus modified to

$$y_e = \widetilde{\beta}_j fp_m + x_e'\widetilde{\gamma} + \widetilde{u}_e, \tag{6}$$

and

$$y_e = \widetilde{\beta}' fp_{jm} + \widetilde{\mu}_j + \widetilde{\tau}_m + x_e'\widetilde{\gamma}' + \widetilde{v}_e, \tag{7}$$

where (as before) $\widetilde{\mu}_j$are plan (by coverage tier) fixed effects, and $\widetilde{\tau}_m$ are join-month fixed effects. This transformation also aids in addressing the likely non-linear effect of join month on both expected end-of-year price and on expected spending. Figure 2 illustrates how this relationship may be non-linear, and Table 2 indicates that, indeed, our measure of the end-of-year price varies non-linearly over time.

Table 4 reports the results. The first three rows report the results for each firm. We report the results for the deductible plan in each firm in columns (1) and (3) and the difference-in-difference results that use the deductible and no-deductible plan within each firm in columns (2) and (4).[15] The difference-in-difference results in Firm B and Firm C (where the sample sizes are much bigger) suggest that a 10 cent increase in the expected end-of-year price is associated with an 8 to 17

[15] Note that, by design, fp is constant for no-deductible plans, so that we cannot estimate the single-plan analysis of the relationship between initial medical utilization and future price for the no-deductible plans.

percent reduction in initial medical spending and with a 2.5 to 7 percent increase in the time to first claim. These results are almost always statistically significant.

Thus far, all of the analysis has been of single plans or pairs of plans within a firm. The use of future price (rather than join month) also allows us to more sensibly pool results across firms and summarize them with a single number, since the relationship between join month and future price will vary both with the level of the deductible (see Figure 2) and with the employee population. In pooling the data, however, we continue to rely on only within firm variation, since we know little about the different firms or about how comparable (or not) their employee populations are (although we show in the appendix that in practice this does not make a substantive difference to the results). Thus, our final specification allows the join month dummy variables $\widetilde{\tau}_m$'s to vary by firm, so that all of the identification is coming from the differential effect of the join month on employees in deductible plans relative to no-deductible plans within the same firm. That is, we estimate

$$y_e = \widetilde{\widetilde{\beta}}' fp_{jm} + \widetilde{\widetilde{\mu}}_j + \widetilde{\widetilde{\tau}}_{mf} + x_e' \widetilde{\widetilde{\gamma}}' + \widetilde{\widetilde{\upsilon}}_e, \tag{8}$$

where $\widetilde{\widetilde{\tau}}_{mf}$ denotes a full set of join month by firm fixed effects. The bottom rows of Table 4 reports the results from this regression. The OLS results (penultimate row of Table 4) will represent our baseline specification in the rest of this section. We defer discussion of the IV results (last row of Table 4) to the next section.

The effect of future price is statistically significant for both dependent variables. The OLS results in the penultimate row indicate that an increase of 10 cents in the future price is associated with an 11% decline in initial medical spending and a 3.6% increase in time to first claim. Overall, the results suggest that we can reject the null of complete myopia ($\delta = 0$). Individuals appear to respond to the future price of medical care in making current medical care utilization decisions. In other words, among individuals who face the same spot price of medical care, individuals who face a higher expected end-of-year price – because they join the plan later in the year – initially consume less medical care.

We also investigated the margin on which the individual's response to the future price occurs. About three quarters of medical expenditures in our data represent outpatient spending; per episode, inpatient care is more expensive and perhaps less discretionary than outpatient care. Perhaps not surprisingly therefore, we find clear evidence of a response of outpatient spending to the future price, but we are unable to reject the null of no response of inpatient spending to the future price (although the point estimates are of the expected sign); Appendix Table A2 contains the results. Unfortunately, we lack the statistical power to be able to examine in more detail the responsiveness of different particular types of spending or of people with particular conditions. Rather, like much of the previous empirical work on moral hazard in health insurance, we focus on the average responsiveness, although we consider this type of heterogeneity an important and interesting topic for further study.

Robustness We explored the robustness of our results to a number of our modeling choices. The first six rows of Table 5 shows that our finding is quite robust across alternative functional forms

for the dependent variable. The first row shows the baseline results, where for initial spending the dependent variable is $\log(s+1)$, where s is total medical spending in the three months after joining the plan, and for time-to-first-claim we estimate a Tobit model for $\log(time)$, where $time$ is the number of days until the first claim, censored at 92.

Row (2) of Table 5 uses levels (rather than logs) of s and $time$ (maintaining the Tobit specification for the $time$ analysis). The statistically significant estimates are comparable in magnitude to those in the baseline specification. Relative to the mean of the dependent variable, the results in row (2) suggest that a 10 cent increase in the future price is associated with a 7% decline in initial spending (compared to an 11% decline estimated in the baseline specification), and a 2.5% increase in the time to first claim (compared to a 3.6% increase in the baseline). In row (3) we report results from quasi-maximum likelihood Poisson estimation and calculate the fully-robust variance covariance matrix (Wooldridge, 2002, pp. 674-676); this is an alternative proportional model to the log specification, and one that allows us to incorporate the frequent occurrence of zero initial spending without adding 1 when the dependent variable is based on three-month spending. The estimate is still statistically significant, although somewhat smaller than our baseline estimate for initial spending (suggesting that a 10 cent increase in the future price is associated with a 7% rather than 11% decline in initial spending).

The next three rows investigate alternative ways of handling the time to first claim analysis. Row (4) shows that estimating the baseline specification by OLS instead of Tobit produces estimates that are still statistically significant but are somewhat smaller than the baseline (a 1.1% rather than 3.6% increase in time to first claim). Row (5) reports result from estimating a censored-normal regression on our baseline dependent variable $\log(time)$, which allows for the censoring value to vary across observations. This allows us to make use of the fact that while we only observe 92 days of medical claims for individuals who join in October, we can expand the observation period for individuals who join in earlier months. The advantage of such a specification is that it makes use of more information; the disadvantage is that it may not be as comparable to the spending estimates since it implicitly gives more weight to individuals who join earlier in the year. The results are virtually identical to the baseline specification. In row (6) we estimate a Cox semi-parametric proportional hazard model of the time to first claim (censored at 92 days for all observations).[16] Consistent with the previous specifications, the results indicate that an increase in the future price is associated with a statistically significant decline in the probability of a claim arrival (i.e., a longer time to first claim).

In Appendix Table A3 (Panels A and B) we further show the robustness of our results to alternative choices of covariates regarding the firm and coverage tier fixed effects. We also explore an alternative measure of the future price which, unlike our baseline measure, accounts for the in-network coinsurance rates in both the deductible and no-deductible plans for the two firms in which this information is available (Alcoa and Firm C; see Appendix Table A1). Accounting for the in-

[16] Van der Berg (2001) discusses the trade-offs involved in analyzing a duration model using a linear model with the logarithm of the duration as the dependent variable, relative to a proportional hazard model. As he explains, neither model strictly dominates the other.

network coinsurance rates for Alcoa and Firm C makes little difference to either our measurement of the future price (Appendix Table A4) or its estimated effect (Appendix Table A3, Panel C), although the results in Appendix Table A3 suggest that, as expected (see discussion in Section 3.2), not accounting for the coinsurance rate slightly biases downward the estimated impact of the future price in our baseline specification.[17]

3.4 Assessing the identifying assumption

The results suggest that we can reject the null of complete myopia in favor of some form of forward looking behavior. The key identifying assumption behind this interpretation of the results is that there are no confounding differences in initial medical utilization among employees by their plan and join month. In other words, any differential patterns of initial medical utilization that we observe across plans by join month is caused by differences in expected end-of-year price. This identifying assumption might not be correct if for some reason individuals who join a particular plan in different months vary in their underlying propensity to use medical care. In particular, one might be concerned that the same forward looking behavior that may lead to differential medical care utilization might also lead to differential selection into a deductible compared to a no-deductible plan on the basis of join month.

A priori, it is not clear if forward looking individuals would engage in differential selection into a deductible vs. no-deductible plan based on the month they are joining the plan. A selection story that would be most detrimental to the interpretation of our results is that individuals who have high expected initial medical expenditure would be more likely to select the no-deductible plan later in the year. For example, if an individual knows that all he needs is a single (urgent) doctor's appointment of $100 (which is below the deductible amount), it may be worth joining the no-deductible plan (and paying the higher monthly premium) if he joins the plan later in the year but not earlier in the year, as late in the year the incremental premium of a no-deductible plan is lower and would be less than the $100 benefits it would provide. This would introduce a positive relationship between individuals who join the no-deductible plan in later months and initial medical utilization and could cause us to erroneously interpret the lack of such a pattern in the deductible plans as evidence that individuals respond to the future price.

On the other hand, there are many reasons to expect no selection, even in the context of forward looking individuals, if there are additional choice or preference parameters governing insurance plan selection that do not directly determine medical utilization. For example, if individuals anticipate the apparently large switching costs associated with subsequent re-optimization of plan choice (as in Handel, 2011) they might make their initial, mid-year plan choice based on their subsequent optimal

[17] We do not observe the breakdown of spending by in-network vs. out-of-network in Firm B or Firm C, so we cannot account for out-of-network spending in our construction of the future price at either of these firms. We do know that in Alcoa, where the data allow us to tell apart in-network spending from out-of-network spending, about 95% of the spending is done in network. We therefore suspect that the accounting for out-of-network spending and out-of-network features of the plan would have little quantitative impact on our estimates of either the future price or the response to it.

open enrollment selection for a complete year. In such a case, we would not expect differential selection into plans by join month. Ultimately, whether there is quantitatively important differential selection and its nature is an empirical question.

The summary statistics in Table 2 present some suggestive evidence that individuals may be (slightly) more likely to select the deductible plan relative to the no-deductible plan later in the year.[18] Quantitatively, however, the probability of selecting the deductible vs. no-deductible plan is very similar over the course of the year. When we regress an indicator variable for whether the employee chose a deductible plan on the employee's join month (enumerated, as before, from 2 to 10), together with a dummy variable for coverage tier and firm fixed effects to parallel our main specification, the coefficient on join month is 0.0034 (standard error 0.0018). Qualitatively, the pattern of greater probability of choosing a deductible later in the year is the opposite of what could produce a confounding explanation for our main empirical results. More importantly, quantitatively the results suggest trivial differential plan selection by join month; joining a month later is associated with only a 0.3 percentage point increase in the probability of choosing the deductible plan, or 0.9% relative to the 32% probability of choosing the deductible plan in the sample. This very similar share of choices of deductible vs. no-deductible plans over the course of the year implies that differential plan selection is unlikely to drive our findings. Consistent with a lack of strategic plan selection based on join month, we also find no evidence in the employees' second year at the firm of differential selection out of the deductible plan relative to the no deductible plan by join month.[19]

We also examined whether the observable characteristics – i.e. age and gender – of individuals joining a deductible vs. no-deductible plan within each of the three firms varied by join month. In general, the results (shown in Appendix Table A5) show little evidence of systematic differences by join month.[20] To examine whether our findings are sensitive to these observable differences, in Row 7 of Table 5 we re-estimate our baseline specification (equation (8)) adding controls for the observable demographic characteristics of the employees: employee age, gender, and join year (see Table 1). In keeping with the "within-firm" spirit of the entire analysis, we interact each of these variables with the firm fixed effects. This specification controls for potential observable differences across employees within a firm by plan type and join month. The results indicate that the impact of these controls is neither substantial nor systematic. The effect of a 10 cent increase in the expected end-of-year price on initial spending declines from 11% in the baseline specification to 10% with the demographic controls, while the effect on time to first claim increases from 3.6% in the baseline

[18] Over the three join quarters shown in Table 2, the share joining the deductible plan varies in Alcoa from 0.49 to 0.53 to 0.53, in firm B from 0.20 to 0.22 to 0.19, and in firm C from 0.38 to 0.40 to 0.44.

[19] On average only about four percent of employees change their plan in the second year, which is consistent with low rates of switching found in other work (Handel, 2011).

[20] While there are two exceptions that show statistically significant differential selection by join month, they are both quantitatively trivial. Employees at Alcoa who join a deductible vs. no-deductible plan one month later in the year are 0.9 percentage points (about 2%) more likely to be female. Employees at Firm B who join a deductible vs. no-deductible plan one month later in the year are 0.6 percentage points (about 2%) less likely to be over 45 (or 0.2 months younger (not shown in the table)).

specification to 5.2% with the demographic controls. All the results remain statistically significant.

As another potential way to investigate the validity of the identifying assumption, we implement an imperfect "placebo test" by re-estimating our baseline specification (equation (8)) with the dependent variable as the "initial" medical utilization in the second year the employee is in the plan. In other words, we continue to define "initial medical utilization" relative to the join month (so that the calendar month in which we measure initial medical utilization varies in the same way as in our baseline specification across employees by join month) but we now measure it in the second year the employee is in the plan. For example, for employees who joined the plan in July 2004, we look at their medical spending during July through September 2005. In principle, when employees are in the plan for a full year there should be no effect of join month (of the previous year) on their expected end-of-year price, and therefore no difference in "initial" utilization by join month across the plans. In practice, the test suffers from the problem that the amount of medical care consumed in the join year could influence (either positively or negatively) the amount consumed in the second year, either because of inter-temporal substitution (which could generate negative serial correlation) or because medical care creates demand for further care (e.g., follow up visits or further tests), which could generate positive serial correlation.

Row (8) of Table 5 shows the baseline results limited to the approximately 60% of the employees who remain at the firm for the entire second year. We continue to find the same basic pattern in this smaller sample although the point estimate declines (in absolute value) and the time to first claim results are no longer statistically significantly different from zero. For this subsample of employees, row (9) shows the results when we now measure "initial medical spending" in the same three months but in the second year.[21] Here we find that an increase in the future price is associated with a much smaller and statistically insignificant decline in medical spending measured over the same three month period but in the second year. We interpret this as generally supportive of the identifying assumption, and suggestive of positive serial correlation in medical spending.

Finally, in row (10) we investigate the extent to which the decrease in utilization in response to a higher future price represents inter-temporal substitution of medical spending to the next year. Such inter-temporal substitution would not be a threat to our empirical design – indeed, it might be viewed as evidence of another form of forward-looking behavior in medical spending – but it would affect the interpretation of our estimates and is of interest in its own right. We therefore re-run our baseline specification but now with the dependent variables measured in January to March of the second year. The results indicate that individuals who face a higher future price (and therefore consume less medical care) also consume less medical care in the beginning of the subsequent year. This suggests that inter-temporal substitution, in the form of postponement of care to the subsequent calendar year, is unlikely to drive the decrease in care associated with a higher future price.

[21] We perform this "second year" analysis only for the dependent variable "initial medical spending" as it seemed awkward to us to examine "time to first claim" from an arbitrary starting point in the second year (when in fact the individual has had prior months to make his first claim).

4 Quantifying forward looking behavior

Our results thus far have rejected the null of no response to the future price and presented evidence consistent with some form of forward looking behavior. A natural subsequent question is to ask how forward looking the behavior is. In other words, having rejected one extreme of complete myopia ($\delta = 0$), we would like to get a sense of where individuals lie on the spectrum between complete myopia ($\delta = 0$) and "full" forward looking behavior ($\delta \approx 1$). Relatedly, we are also interested in the implications (relative to either of these extremes) of the amount of forward looking behavior we detect for the the impact of alternative health insurance contracts on annual medical spending.

4.1 Quantifying the effect of the future price on initial medical utilization

We start by quantifying the elasticity of initial medical utilization with respect to the future price. The results reported in the previous section tested whether there was a relationship between the future price and initial medical utilization. However, a concern with interpreting this relationship as the causal effect of the future price on initial medical utilization is that there is a mechanical relationship between initial medical utilization (the dependent variable) and our measure of the future price (the right-hand-side variable). The future price is a function of the plan (by coverage tier) chosen, the month joined, and the monthly medical spending of people who join that plan (by coverage tier) in that month; thus, the future price is a function of medical spending which is also used in constructing the dependent variable. This is not a concern for testing the null of complete myopia (i.e., testing whether the coefficient on the future price is zero) which was the focus of the last section, because under the null of complete myopia medical spending is not a function of the future price. However, under the alternative hypothesis that individuals are forward looking, this can bias away from zero the estimated response to the future price.

To address this concern, we present results from estimating an instrumental variable version of equation (8) in which we instrument for the future price with a simulated future price. Like the future price, the simulated future price is computed based on the characteristics of the plan (by coverage tier) chosen and the month joined. However, unlike the future price which is calculated based on the spending of people who joined that plan (by coverage tier) that month, the simulated future price is calculated based on the spending of all employees in that firm and coverage tier in our sample who joined either the deductible or no-deductible plan, regardless of join month.[22] By using a *common sample* of employee spending that does not vary with plan or join month, the

[22] Specifically, for every employee in our sample in a given firm and coverage tier (regardless of plan and join month) we compute their monthly spending for all months that we observe them during the year that they join the plan, creating a common monthly spending pool. We then simulate the future price faced by an employee in a particular plan and join month by drawing (with replacement) 110,000 draws of monthly spending from this common pool, for every month we need a monthly spending measure. For the first month we draw from the pool of first month spending (since people may join the plan in the middle of the month, the first month's spending has a different distribution from other months) whereas for all other months in the plan that year we draw from the pool (across families and months) of non first month spending. For each simulation we then compute the expected end-of-year price based on the draws.

instrument is "purged" of any potential variation in initial medical utilization that is correlated with plan and join month, in very much the same spirit as Currie and Gruber's (1996) simulated Medicaid eligibility instrument. An additional attraction of this IV strategy is that it helps correct for any measurement error in our calculated future price (which would bias the coefficient toward zero). On net, therefore, the OLS may be biased upward or downward relative to the IV.

The bottom row of Table 4 shows the results from this IV strategy. As would be expected, the first stage is very strong and the IV estimates are statistically significant.[23] For the dependent variable log initial spending, the point estimate from the IV results suggests that a 10 cent increase in the expected end-of-year price is associated with a 7.8% decline in initial medical spending. Given an average expected end-of-year price for people in our sample who choose the deductible plan of about 70 cents, this suggests an elasticity of initial medical utilization with respect to the future price of -0.56. For the dependent variable log time to first claim, the IV results suggest that a 10 cent increase in the expected end-of-year price is associated with a 5.6% increase in the time to first claim, or an elasticity of initial medical utilization with respect to the future price of about -0.39.[24]

4.2 Mapping the estimated elasticity to economic primitives of interest

There are (at least) two related reasons why this estimate of the elasticity of initial medical utilization with respect to the future price is an unsatisfactory answer to the question: how important is forward looking behavior? The first reason is that this elasticity measures the effect of future price on *initial* spending, while we suspect that *total* (annual) spending is the outcome variable of interest for most research or policy questions associated with health insurance utilization. The importance of dynamic incentives for annual spending may well be much lower than for initial spending since the wedge between the spot price and the future price becomes smaller as health shocks accumulate within the year and/or the end of the coverage period nears.

The second reason is that it is difficult to assess whether the elasticity is large or small without

[23] The first stage coefficient from the regression of the future price on the instrument (as well as plan-by-coverage tier fixed effects and firm-by-start month fixed effects) yields a coefficient (on the instrument) of 0.56 (standard error 0.024).

[24] In principle, the IV estimate of the impact of the future price on the first three months' spending could be biased upward since, over the first three months, 17% of the individuals in deductible plans spend past the deductible. If individuals are at least partially forward looking, the probability of hitting the deductible in the first three months could be correlated with join month, which would introduce variation during the first three month in the spot price among individuals who join the same plan in different months. Once again, this is not a problem for testing the null of complete myopia; nor is it a problem when the dependent variable is the time to first claim (since the spot price is the same for all individuals within a plan at the time of first claim). In practice, moreover, any upward bias is likely unimportant quantitatively. We estimate a virtually identical response to the future price when the dependent variable is based on two-month (instead of three-month) spending, even though the fraction hitting the deductible within the initial utilization period (and therefore the likely magnitude of the bias) drops by almost a half. Moreover, there is no noticeable trend in the likelihood of hitting the deductible within the first three months by the join month; hitting the deductible within a short time after enrollment therefore appears to be primarily determined by large and possibly non-discretionary health shocks, rather than an endogenous spending response to the future price.

appropriate benchmarks. We would like to compare our estimated elasticity with respect to the future price to the "primitive" price elasticity, i.e. the underlying elasticity that is driven by substitution between health and income and is purged of dynamic incentives. However, the same motivation that prompted us to write this paper also implies that the prior empirical literature does not provide such benchmarks. As noted in the introduction, most papers in this literature estimate the elasticity of demand for medical care with respect to its price under a specific assumption about how forward looking individuals are. For example, the commonly cited price elasticity of demand for medical care of -0.2 from the RAND Health Insurance Experiment was estimated under the assumption that individual behavior is completely myopic (Keeler and Rolph, 1988), which is precisely the question we are investigating.[25]

Some assumption about the nature and extent of forward looking behavior is required in the existing literature because it has not examined the impact of linear contracts on spending. If we had an estimate of the utilization response to the coinsurance rate in a linear contract, for which the price of medical care is constant for an individual throughout the year, this would be a useful benchmark against which to compare our estimated response to the future price. In a linear contract, dynamic considerations should not affect utilization decisions, so that the behavioral response to different prices (coinsurance rates) would be invariant to the extent of forward looking behavior, and could therefore shed light on the "primitive" substitution between health and income. However, we know of no estimates of the response to a linear contract, nor a source of clean variation in the (constant) coinsurance rate that could be used to identify this response.[26] In the remainder of this section, we therefore calibrate a stylized dynamic model in order to translate our baseline estimate of the response to the future price into economic primitives of interest.

Calibration exercise To try to gauge what our estimated elasticity with respect to the future price implies for how forward looking individuals are, as well as to assess the implications of this finding for the impact of alternative health insurance contracts on *annual* medical spending, we turn to the stylized model of medical utilization decisions in response to health shocks that we developed in Section 2. We investigate what degree of forward looking behavior (δ) is needed in that model to generate the magnitude of the response of initial medical utilization to the future price that we estimated in our data. Specifically, we calibrate the other parameters of the model and then simulate the response of initial medical utilization to the future price under alternative

[25] More precisely, Keeler and Rolph (1988) assume that individuals are completely myopic about the possibility of future health shocks in making current medical spending, but that they have perfect foresight regarding all of the year's medical spending associated with a given health shock.

[26] In the Appendix we show how we can use the experimental variation from the RAND Health Insurance Experiment in both the coinsurance rate and the out-of-pocket maximum to extrapolate (out of sample) to the effect of the coinsurance rate in a plan with an infinite out-of-pocket maximum, which thus approximates the response to a linear contract. Our point estimates, while quite imprecise in most specifications, tend to suggest a semi-elasticity of medical utilization with respect to the price of a linear contract that ranges from our estimate of the semi-elasticity with respect to the future price to up to twice as large as this estimate. We interpret the results of this exercise as suggestive of potentially substantial, but perhaps not full, forward looking behavior.

assumptions about δ; we search for the value of δ that, in this calibrated model, produces the response to the future price that we estimated in the foregoing empirical work.

To do this exercise requires that we calibrate the other primitives of the model in Section 2. These are the arrival rate λ of medical shocks, and the distribution of medical shocks $G(\theta, \omega)$ when they arrive. The latter can be rewritten as $G(\theta, \omega) \equiv G_2(\omega|\theta)G_1(\theta)$. That is, $G_1(\theta)$ represents the unconditional distribution of the total spending that would be required to treat medical shocks and $G_2(\omega|\theta)$ represents the distribution of the (monetized) utility loss from not treating a medical shock of size θ; in that sense, the distribution of ω relative to θ (or simply the distribution of the ratio ω/θ) can be thought of as the "primitive" price elasticity that captures substitution between health and income. As ω/θ is higher (lower), the utility loss is greater (smaller) relative to the cost of treating the shock, so (conditional on the price) the medical shock is more (less) likely to be treated.

We draw on data from the RAND Health Insurance Experiment to calibrate these additional parameters.[27] Conducted over three to five years in the 1970s on a representative population of individuals under 65, the key feature of the RAND experiment was to experimentally vary the health insurance plans to which individuals were assigned. In particular, the coinsurance in the plans varied from "free care" (zero coinsurance rate) to 100% coinsurance rate, with individuals also assigned to 25%, 50%, and 95% coinsurance rates. The details of the experimental design as well as the main results in terms of the impact of consumer cost sharing on healthcare spending and health have been summarized elsewhere (Manning et al., 1987; Newhouse et al., 1993).[28] The estimates from this famous study still remain the standard reference for calibration exercises that require a moral hazard estimate for health insurance (e.g., Finkelstein, Luttmer, and Notowidigdo, 2008; Mahoney, 2010; Gross and Notowidigdo, 2011) and the standard benchmark with which to compare newer estimates of the impact of health insurance on health spending (e.g., Finkelstein, 2007; Chandra, Gruber, and McKnight, 2010; Finkelstein et al., 2011).

Two features of the RAND experiment are very useful for our particular calibration exercise. First, the existence of detailed data on medical claims under a zero cost sharing (free care) plan is not something that, to our knowledge, exists elsewhere. Such data allow us to calibrate the distribution of medical shocks (λ and $G_1(\theta)$) from data that is "uncensored" by any response to cost-sharing; by contrast, any other plan with positive consumer cost sharing only provides information on the medical shocks that are endogenously treated. Second, the experimental variation in plan assignment helps us calibrate the primitive price elasticity $G_2(\omega|\theta)$.

We defer many of the calibration details to the Appendix, and only summarize them here briefly. In the first step of our calibration exercise, we perform a simple statistical exercise to calibrate the weekly arrival and distribution of medical shocks (λ and $G_1(\theta)$) based on the detailed utilization data for the approximately 2,400 family-years we observe in the RAND's free care plan.

In the second step, we use the experimental plan variation in the RAND data to calibrate

[27] The data from the RAND experiment have, very helpfully, been made publicly available by the RAND investigators through ICPSR.

[28] In the Appendix we provide some more details on the experimental design and the data.

$G_2(\omega|\theta)$. As mentioned above and discussed in more detail in the Appendix, the RAND experiment does not involve variation in linear contracts that would allow us to directly estimate the "primitive" price elasticity $G_2(\omega|\theta)$. Rather, families in the experiment were randomized into plans with different coinsurance rates and then, within each positive coinsurance rate, they were further randomized into plans with different out-of-pocket maximums. The observed changes in behavior, as both the coinsurance rate and the out-of-pocket maximum are experimentally varied, are therefore influenced by both $G_2(\omega|\theta)$ and δ. Our second step of the calibration exercise uses the random assignment of families to plans and our calibrated model of the arrival and distribution of medical shocks, to map the spending response to different plans to values of $G_2(\omega|\theta)$ and δ that would rationalize this spending response. Fortunately, the resultant values of $G_2(\omega|\theta)$ are quite stable, and are not at all sensitive to the value of δ, so that we can use the RAND experiment to calibrate $G_2(\omega|\theta)$ without knowledge of δ.[29] We can thus use the experimental variation to calibrate $G_2(\omega|\theta)$, and are left with δ as the only remaining unknown primitive.

In the final step of the calibration exercise, we use the calibrated parameters of the model that we have just described to simulate initial medical utilization under deductible contracts with coverage horizons of 3 to 11 months, artificially replicating the setting in which we obtained our estimated elasticity of initial medical utilization with respect to the future price. We repeat this simulation under alternative assumptions about the value of δ. Higher values of δ correspond to greater changes in initial medical utilization as the coverage horizon varies. To quantify this, we regress, for each δ, initial medical utilization in the simulated data on the future price.[30] We then ask what value of δ gives rise to the magnitude of the change in initial medical utilization with respect to the future price that we estimated based on variation in the coverage horizon in our employer-provided data (see last two rows of Table 4).

Calibrated value of δ Figure 3 illustrates our exercise by plotting the semi-elasticity of initial (three month) medical spending with respect to the future price implied by each value of δ. Our point estimate of the relationship between initial medical spending and future price was -1.08 in the OLS estimation in the penultimate row of Table 4, with the 95% confidence interval ranging

[29] Less fortunately, the converse is not true: the RAND experiment by itself does not allow us to pin down δ with any confidence. In principle, the experimental variation in both coinsurance rates and out of pocket maximums makes the RAND data seem perfectly suited to test and quantify forward looking behavior (since there is experimental variation in the future price conditional on the experimentally determined spot price). In practice, however, using the RAND data to estimate the behavioral response to the future price encounters two important obstacles. The first is conceptual: the combination of non-trivial risks of fairly large expenditure shocks and a preponderance of relatively low out-of-pocket maximums means that is difficult to isolate variation in the future price, as it mechanically generates variation in spot prices that is driven by large medical shocks that are greater than the (lower) out of pocket maximums. The second obstacle is practical: given its much smaller sample size, our attempt to use the RAND variation (despite the first issue) to estimate the behavioral response to the future price produced extremely noisy estimates. The Appendix provides additional details and results of this analysis.

[30] These regressions, as in our earlier empirical analysis, abstract from private information that individulas may have regarding their future spending, and thus their expected future price. The "symmetric" treatment of this private information in both the RAND analysis and the primary analysis makes us worry less that this concern regarding private infromation would drive our quantification results in one direction or another.

from -1.66 to -0.50. Figure 3 indicates that this point estimate in the simulated data is achieved with $\delta = 0.2$, with the 95% confidence interval ranging from 0.06 to 0.45. Table 6 summarizes the implied δ's from the simulation exercises using the alternative dependent variable (time to first claim) and based on IV estimation rather than OLS estimation in both the actual and simulated data.[31] Across the four specifications, the point estimate of δ are centered around 0.2, with a low of around 0.1 and a high of around 0.7.

These calibration results therefore suggest that while we find evidence of forward looking behavior, the extent of forward looking behavior is far from what would be implied by a perfectly rational, fully forward looking individual ($\delta \approx 1$) and closer to what would be implied by a fully myopic individual ($\delta = 0$). Of course, as we noted at the outset, δ – or "forward looking" behavior in our context – should not be interpreted as a pure rate of time preference; liquidity constraints and/or imperfect understanding of the coverage details can push the estimated δ below the rate of time preference, and presumably do so in our context.

Implications for impact of health insurance on spending behavior We can also use our calibrated model to try to assess whether the positive but low δ we have calibrated is quantitatively important for understanding the response of medical utilization to non-linear health insurance contracts. In other words, we try to get a feel for whether, despite the fact that our testing exercise in the main part of the paper rejects fully myopic behavior, myopia could be a reasonable way to approximate behavior. The answer will depend of course not only on our estimate of δ but also on the other parameters of the model and the contracts examined. For example, if the deductible level is low and the vast majority of individuals will exhaust it quickly, most individuals will spend most of the time past the deductible, where they are effectively covered by a linear contract, so that the extent of forward looking behavior would not matter much for the impact of the health insurance contract on medical utilization.

Figure 4 uses the calibrated model to report total annual spending for contracts with different deductible levels in the range of what is common in employer-provided health insurance contracts, and full coverage (zero coinsurance rate) beyond the deductible. It shows results under alternative assumptions about δ. The annual spending levels are based on simulated results from the calibrated model. We are interested in whether low values of δ (of, say, 0.1 or 0.2) can be reasonably approximated by an assumption of either complete myopia ($\delta = 0$) – as underlies for example the famous RAND estimate of the price elasticity of demand for medical care – or an assumption of perfectly forward looking behavior ($\delta \approx 1$) – as has been assumed by other papers estimating the responsiveness of medical care to health insurance contracts. The results in the figure suggest that both these extremes produce substantively different results for the impact of these health insurance contracts on total spending relative to our calibrated estimates of δ. For example, across all the deductible levels we consider, as we move from the no-deducible plan to a positive deductible plan

[31] Since the endogeneity of the measured future price to initial medical utilization exists in both the actual and simulated data, comparing the OLS estimates from the actual data to the OLS estimates of the simulated data – or comparing the IV estimates from the actual data to the IV estimates from the simulated data – are both meaningful.

the decrease in spending implied by $\delta = 0.2$ is 25 to 50 percent smaller than what would be implied by myopic behavior ($\delta = 0$), and 50 to 270 percent greater than what would be implied by $\delta = 1$. These results point to the empirical importance of accounting for dynamic incentives in analyses of the impact of health insurance on medical utilization, and relatedly to the dangers in trying to summarize health insurance contracts with a single price.

5 Conclusions

Our paper rejects the null of completely myopic behavior in individuals' response to the non-linear price of medical care. This result jointly indicates that individuals understand something about the non-linear pricing schedule they face, and that they take account of the future price of care in making current medical decisions. Calibration results from our stylized, dynamic model of medical utilization suggests that, at least in the populations we study, individuals may be far from fully forward looking, but that, nonetheless, the extent of forward looking behavior we detect has a non-trivial impact for forecasting how medical spending will respond to changes in non-linear health insurance contracts.

These findings have important implications for estimating or forecasting the impact of alternative health insurance contracts on medical spending, which is a topic that receives great interest and attention both by academics and in the public policy arena. As we noted at the outset, the work to date has almost exclusively focused on estimating (and then using) the elasticity of demand for medical care with respect to its "price". However, faced with a non-linear budget set, unless individuals are completely myopic or completely forward looking in their decision making, characterizing moral hazard in health insurance using a single elasticity estimate is neither informative as to how it should be used (relative to which price?) nor is it conceptually well-defined (there are at least two price elasticities that are relevant). Thus, our results highlight the need for more complete modeling of medical utilization induced by the health insurance contract in estimating and forecasting the likely effects of these non-linear pricing schedules among forward looking individuals. More generally, our results speak to the question of whether individuals understand and respond to the incentives embodied in non-linear pricing schedules, of which health insurance contracts are just one of many common examples.

Of course, our findings are specific to our population, which consists of individuals with employer-provided health insurance. Such individuals may be more forward looking than the general population, or may be less liquidity constrained and therefore less responsive to the spot price. It is therefore very possible that in other populations, particularly populations with lower education or income, the extent or even the existence of forward looking behavior might be very different. In settings where individuals appear to behave mostly or entirely myopically it becomes both interesting and important to understand the sources of this apparent myopia, such as the relative roles of time horizon and liquidity constraints. We think that extending our analysis to other settings and attempting to decompose the sources of any myopic component of behavior are promising directions for future work.

References

Adams, William, Liran Einav, and Jonathan Levin (2009). "Liquidity Constraints and Imperfect Information in Subprime Lending." *American Economic Review* 99(1), 49-84.

Blundell, Richard, and Thomas MaCurdy (1999). "Labor Supply: A Review of Alternative Approaches" in Ashenfelter, Orley, and David Card (eds.), *Handbook of Labor Economics*. Oxford: Elsevier North Holland.

Borenstein, Severin (2009). "To What Electricity Price Do Consumers Respond? Residential Demand Elasticity Under Increasing-Block Pricing." Mimeo, UC Berkeley.

Burtless, Gary, and Jerry Hausman (1978). "The Effect of Taxation on Labor Supply: Evaluating The Gary Negative Income Tax Experiment." *Journal of Political Economy* 86(6), 1103–1130.

Cardon, James H., and Igal Hendel (2001). "Asymmetric Information in Health Insurance: Evidence from The National Medical Expenditure Survey." *Rand Journal of Economics* 32, 408-427.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight (2007). "Patient Cost-Sharing, Hospitalization Offsets, and the Design of Optimal Health Insurance for the Elderly." NBER Working Paper No. 12972.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight (2010). "Patient Cost-Sharing, Hospitalization Offsets, and the Design of Optimal Health Insurance for the Elderly." *American Economic Review* 100(1): 193-213.

Chetty, Raj, John Friedman, Tore Olsen, and Luigi Pistaferri (2011). "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *Quarterly Journal of Economics* 126(2), 749-804.

Chetty, Raj, and Emmanuel Saez (2009). "Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients." Mimeo, Harvard University.

Chevalier, Judith, and Austan Goolsbee (2009). "Are Durable Goods Consumers Forward-Looking? Evidence from College Textbooks." *Quarterly Journal of Economics* 124(4), 1853-1884.

Currie, Janet, and Jonathan Gruber (1996). "Health Insurance Eligibility, Utilization of Medical Care, and Child Health." *Quarterly Journal of Economics* 111(2), 431-466.

Eichner, Matthew J. (1997). "Medical Expenditures and Major Risk Health Insurance." MIT Ph.D. Dissertation, Chapter 1.

Eichner, Matthew J. (1998). "The Demand for Medical Care: What People Pay Does Matter." *American Economic Review Papers and Proceedings* 88(2), 117-121.

Einav, Liran, Amy Finkelstein, Stephen Ryan, Paul Schrimpf, and Mark R. Cullen (2011). "Selection on Moral Hazard in Health Insurance." NBER Working Paper No. 16969.

Ellis, Randall (1986). "Rational Behavior in the Presence of Coverage Ceilings and Deductibles." *RAND Journal of Economics* 17(2), 158-175.

Feldman, Naomi E., and Peter Katuscak (2006). "Should the Average Tax Rate Be Marginalized?" Working Paper No. 304, CERGE-EI.

Finkelstein, Amy (2007). "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare." *Quarterly Journal of Economics* 122(1), 1-37.

Finkelstein, Amy, Erzo Luttmer and Matthew Notowidigdo (2008). "What Good Is Wealth Without Health? The Effect of Health on the Marginal Utility of Consumption." NBER Working Paper No. 14089.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group (2011). "The Oregon Health Insurance Experiment: Evidence from the First Year." NBER Working Paper No. 17190.

French, Eric, and John B. Jones (2004). "On the Distribution and Dynamics of Health Costs." *Journal of Applied Econometrics* 19(6), 705–721.

Gross, Tal, and Matthew Notowidigdo (2011). "Health Insurance and the Consumer Bankruptcy Decision: Evidence from Expansions of Medicaid." *Journal of Public Economics* 95(7-8), 767-778.

Grubb, Michael D., and Matthew Osborne (2011). "Cellular Service Demand: Tariff Choice, Usage Uncertainty, Biased Beliefs, and Learning." Mimeo, MIT.

Handel, Benjamin (2011). "Adverse Selection and Switching Costs in Health Insurance Markets: When Nudging Hurts." Mimeo, UC Berkeley.

Hausman, Jerry (1985). "The Econometrics of Nonlinear Budget Sets." *Econometrica* 53, 1255–1282.

Ito, Koichiro (2010). "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing." Mimeo, UC Berkeley.

Keeler, Emmett, Joseph P. Newhouse, and Charles Phelps (1977). "Deductibles and The Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule under Uncertainty." *Econometrica* 45(3), 641-655.

Keeler, Emmett B., and John E. Rolph (1988). "The Demand for Episodes of Treatment in the Health Insurance Experiment." *Journal of Health Economics* 7, 337-367.

Kowalski, Amanda (2010). "Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care." NBER Working Paper No. 15085.

Kowalski, Amanda (2011). "Estimating the Tradeoff Between Risk Protection and Moral Hazard with a Nonlinear Budget Set Model of Health Insurance." Mimeo, Yale University.

Liebman, Jeffrey B., and Erzo F. P. Luttmer (2011). "Would People Behave Differently If They Better Understood Social Security? Evidence From a Field Experiment." NBER Working Paper No. 17287.

Liebman, Jeffrey B., and Richard J. Zeckhauser (2004). "Schmeduling." Mimeo, Harvard University.

Manning, Willard, Joseph Newhouse, Naihua Duan, Emmett Keeler, Arleen Leibowitz, and Susan Marquis (1987). "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." *American Economic Review* 77(3), 251-277.

Mahoney, Neale (2010). "Bankruptcy as Implicit Health Insurance." Unpublished mimeo. Available at http://www.stanford.edu/~nmahoney/Research/Mahoney_Bankruptcy.pdf.

Marsh, Christina (2011). "Estimating Health Expenditure Elasticities using Nonlinear Reimbursement." Mimeo, University of Georgia.

Medstat (2006). "MarketScan Commercial Claims and Encounters Database, Description of Deliverables."

Newey, Whitney K. (1987). "Specification Tests for Distributional Assumptions in the Tobit Model." *Journal of Econometrics* 34, 124-145.

Newhouse, Joseph P., and the Insurance Experiment Group (1993). *Free for All? Lessons from the RAND Health Insurance Experiment*. Harvard University Press, Cambridge, MA.

Reiss, Peter C., and Matthew W. White (2005). "Household Electricity Demand, Revisited." *Review of Economic Studies* 72, 853–883.

Saez, Emmanuel (2010). "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* 2, 180–212.

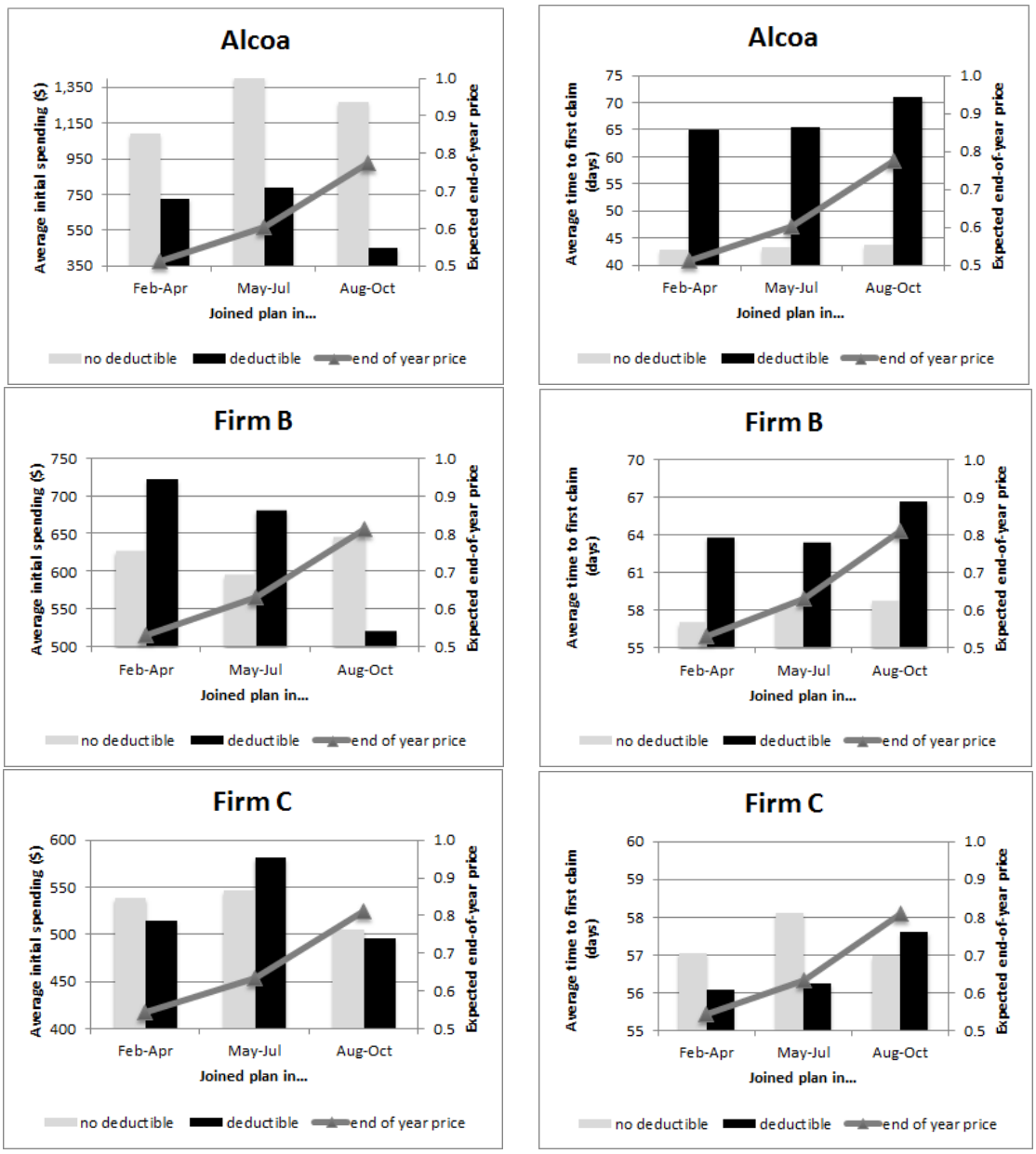
Van den Berg, Gerard J. (2001). "Duration Models: Specification, Identification and Multiple Durations." In J. J. Heckman and E. E. Leamer (eds.), *Handbook of Econometrics* (First Ed.), Vol. 5, Amsterdam: Elsevier, Chapter 55, 3381–3460.

Yao, Song, Yuxin Chen, Carl F. Mela, and Jeongwen Chiang (2011). "Determining Consumers' Discount Rates with Field Studies." Mimeo, Duke University.

Zweifel, Peter and Willard Manning. (2000). "Moral hazard and consumer incentives in health care." In A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics* Vol. 1, Amsterdam: Elsevier, Chapter 8, 410-459.

Figure 1: Initial medical utilization by join quarter

Initial (3 month) spending | Time to first claim



All utilization measures refer to utilization by the employee and any covered dependents. Note that lower utilization implies *less* initial spending (all panels on the left) but *longer* time to first claim (all panels on the right). Days to first claim is censored for all employees at 92 days (42% of observations are censored). Expected end-of-year price is computed for the deductible plan only and corresponds to the end-of-year prices reported in Table 2. Sample sizes by plan and join quarter are reported in Table 2.

Figure 2: Model illustration

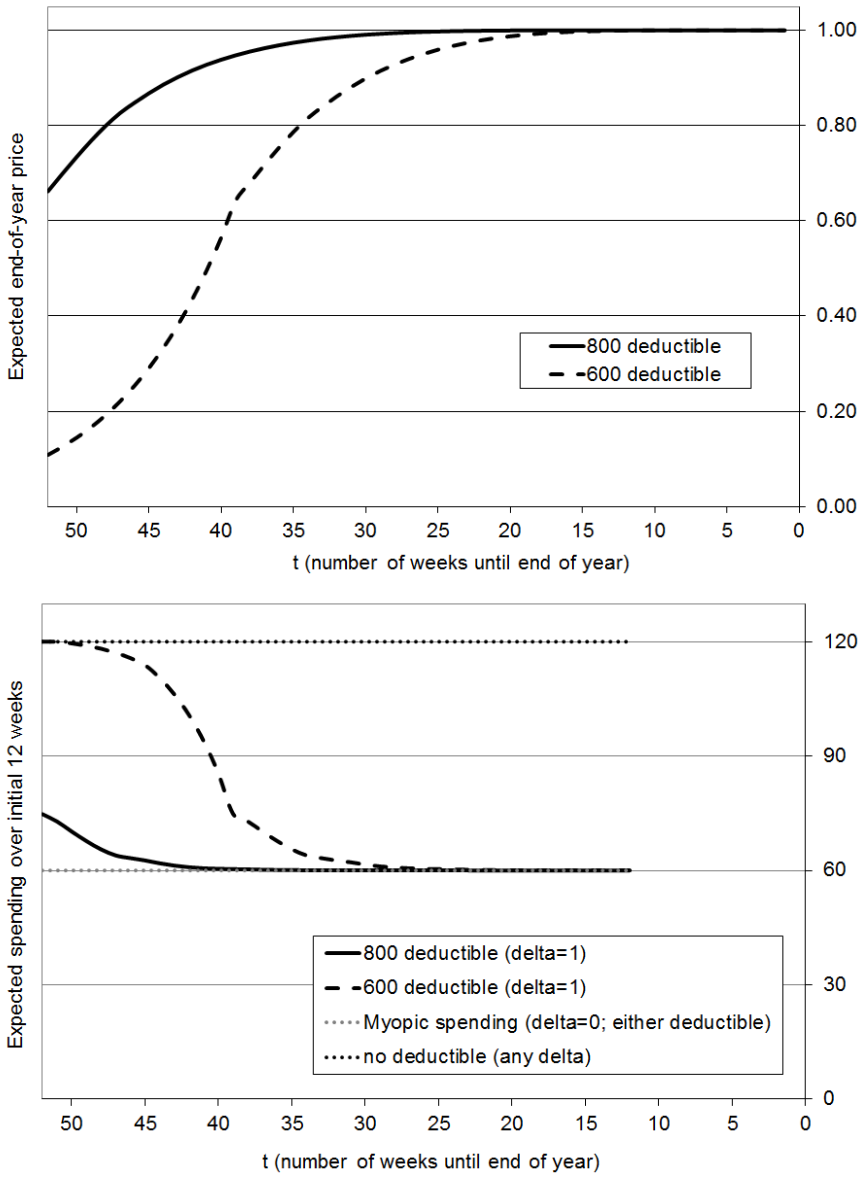


Figure illustrates the implications from a numerical solution to a simple version of the model described in Section 2. We assume $\lambda = 0.2$ and medical events are drawn uniformly from a two-point support of $(\theta = 50, \omega = 50)$ and $(\theta = 50, \omega = 45)$. Expected end-of-year price is equal to one minus the probability of hitting the deductible by the end of the year.

Figure 3: Calibration of δ

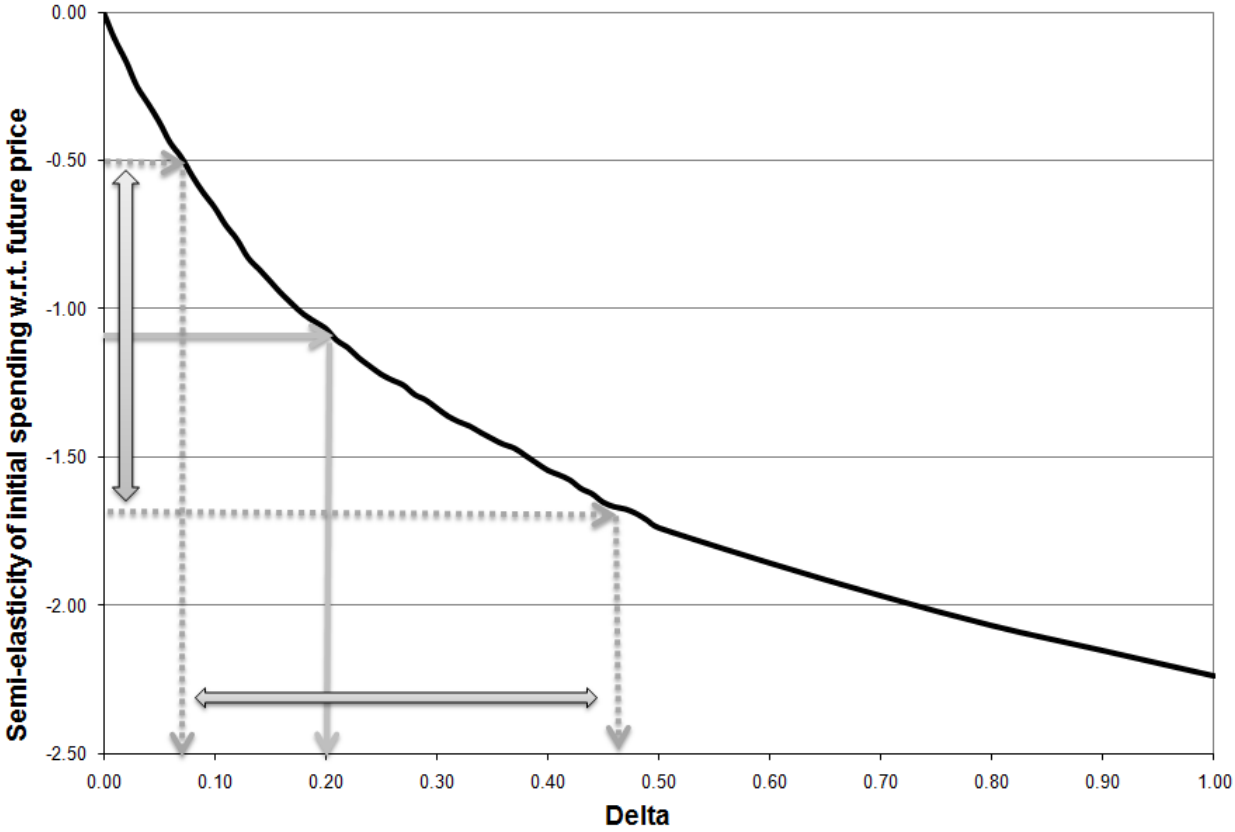


Figure illustrates our calibration exercise. The plot presents the relationship implied by our calibration exercise (see the Appendix for details) between δ and the semi-elasticity of initial medical spending with respect to the future price. The arrows then illustrate how the point estimate and the confidence interval of our semi-elasticity OLS estimate of the impact of the future price on initial spending (penultimate row of Table 4) translate to a point estimate and a confidence interval for δ.

Figure 4: The effect of δ on *annual* spending

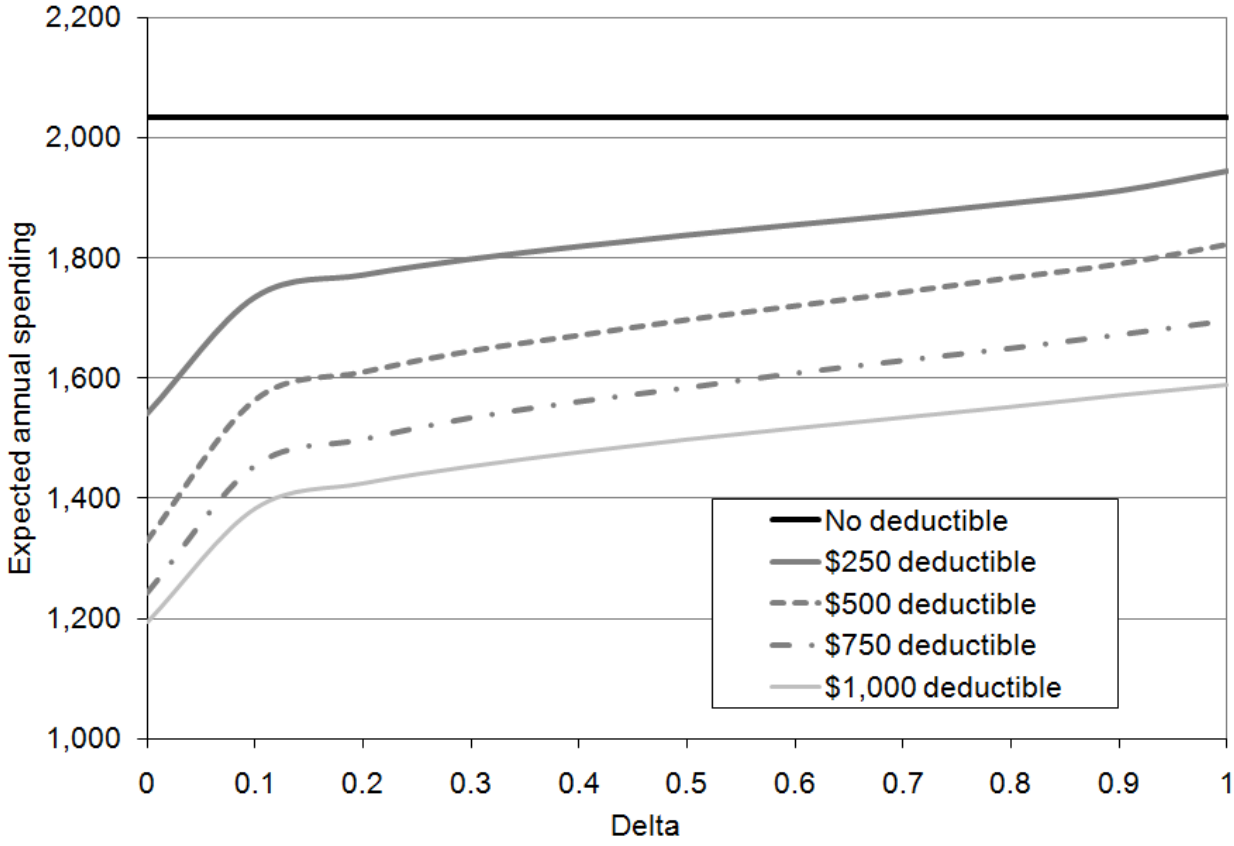


Figure illustrates the implications δ on overall (annual) spending, given the calibration exercise (see the Appendix for details), for a range of possible contracts. The black line represents a case of full insurance, in which overall spending is highest and does not depend on δ. The other lines represent overall spending for deductible contracts which provide full insurance (zero coinsurance rate) once the deductible level has been reached.

Table 1: Summary statistics

Employer	Plan	Years offered	Mid-year new enrollees[a]	In-network deductible ($)		Fraction family	Fraction Female	Average Age	"Average" enrollment month[b]
				Single	Family				
Alcoa	A0	2004-07	3,269	0	0	0.622	0.379	38.56	6.28
	A1	2004-07	3,542	250	500	0.408	0.254	35.68	6.42
Firm B	B0	2001-05	37,759	0	0	0.530	0.362	36.77	6.35
	B1	2001-05	9,553	150	300	0.382	0.341	36.87	6.29
Firm C[c]	C0	1999-2002, 2004-05	27,968	0	0	0.348	0.623	36.40	7.35
	C1	1999-2000	6,243	200	500	0.348	0.622	37.53	7.50
	C2	2001-02	8,055	250	625	0.323	0.606	38.66	7.56
	C3	2004-05	5,633	300	750	0.299	0.660	38.51	7.67

[a] The sample includes employees who enroll in February through October.

[b] In computing the "average" enrollee month we number the join months from 2 (February) through 10 (October).

[c] We omit 2003 from the analysis since the plan documentation regarding the deductible plan was incomplete in that year.

Table 2: Variation in expected end-of-year price

Employer	Plan	Deductible (Single/Family) [N = enrollees]	Expected end-of-year price[a] Joined plan in Feb-Apr	May-Jul	Aug-Oct
Alcoa	A0	0 [N = 3,269]	0.000 (N = 1,007)	0.000 (N = 981)	0.000 (N = 1,281)
	A1	250/500 [N = 3,542]	0.512 (N = 975)	0.603 (N = 1,114)	0.775 (N = 1,453)
Firm B	B0	0 [N = 37,759]	0.000 (N = 8,863)	0.000 (N = 15,102)	0.000 (N = 13,794)
	B1	150/300 [N = 9,553]	0.529 (N = 2,165)	0.630 (N = 4,175)	0.806 (N = 3,213)
Firm C	C0	0 [N = 27,968]	0.000 (N = 6,504)	0.000 (N = 6,158)	0.000 (N = 15,306)
	C1-C3[b]	200-300/500-750 [N = 19,931]	0.543 (N = 4,001)	0.633 (N = 4,143)	0.811 (N = 11,787)

[a] Expected end-of-year price is equal to the fraction of individuals who do not hit the deductible by the end of the calendar year (and therefore face a marginal price of 1). It is computed based on the plan's deductible level(s), the join month, and the annual spending of all the employees who joined that plan in that month; we compute it separately for family and single coverage within a plan and report the enrollment-weighted average.

[b] In firm C, we pool the three different deductible plans (C1, C2, and C3) that are offered in different years.

Table 3: The relationship between join month and initial medical utilization

Employer	Plan	Deductible (Single/Family) [N = enrollees]	Log Initial Spending[a] Difference (1)	Log Initial Spending[a] DD (2)	Log Time to First Claim[b] Difference (3)	Log Time to First Claim[b] DD (4)
Alcoa	A0	0 [N = 3,269]	-0.003 (0.023)		0.007 (0.010)	
	A1	250/500 [N = 3,542]	-0.015 (0.021)	-0.012 (0.027)	0.003 (0.014)	-0.005 (0.015)
Firm B	B0	0 [N = 37,759]	-0.015 (0.007)		0.024 (0.008)	
	B1	150/300 [N = 9,553]	-0.091 (0.026)	-0.075 (0.025)	0.059 (0.014)	0.033 (0.010)
Firm C	C0	0 [N = 27,968]	-0.004 (0.013)		0.003 (0.006)	
	C1-C3	200-300/500-750 [N = 19,931]	-0.027 (0.012)	-0.022 (0.010)	0.019 (0.007)	0.016 (0.006)

Table reports coefficients (and standard errors in parentheses) from regressing a measure of initial medical care utilization (defined in the top row) on join month (which ranges from 2 (February) to 10 (October)). Columns (1) and (3) report the coefficient on join month separately for each plan, based on estimating equation (4); the regressions also include an indicator variable for coverage tier (single vs. family). Columns (2) and (4) report the difference-in-differences coefficient on the interaction of join month and having a deductible plan, separately for each firm, based on estimating equation (5); the regressions also include plan by coverage tier fixed effects and join month fixed effects. Standard errors are clustered on join month by coverage tier.

[a] Dependent variable is $log(s+1)$ where s is the total medical spending of the employee and any covered family members in their first three months in the plan.

[b] Dependent variable is $log(time)$ where "time" is the number of days to first claim by any covered family member, censored at 92. We estimate the regressions in columns (3) and (4) by Tobit.

Table 4: The relationship between expected end-of-year price and initial medical utilization

Sample	N	Log Initial Spending[a]		Log Time to First Claim[b]	
		Difference	DD	Difference	DD
		(1)	(2)	(3)	(4)
Alcoa	6,811	-0.92	-0.76	0.294	0.046
		(0.30)	(0.51)	(0.176)	(0.191)
Firm B	47,312	-2.02	-1.73	1.171	0.677
		(0.57)	(0.54)	(0.363)	(0.227)
Firm C	47,899	-0.89	-0.81	0.357	0.254
		(0.39)	(0.37)	(0.234)	(0.143)
Pooled (OLS)	102,022		-1.08		0.357
			(0.29)		(0.113)
Pooled (IV)	102,022		-0.78		0.564
			(0.27)		(0.135)

Table reports coefficients (and standard errors in parentheses) from regressing a measure of initial medical care utilization (defined in the top row) on the expected end-of-year price, computed for each plan (by coverage tier) and join month. Columns (1) and (3) report the coefficient on expected end-of-year price (fp) separately for each deductible plan in each firm, based on estimating equation (6); the regressions also include an indicator variable for coverage tier (single vs. family). Columns (2) and (4) report the coefficient on expected end-of-year price (fp) from estimating equation (7), which now includes the no-deductible plans as well; these regressions also include plan by coverage tier fixed effects and join month fixed effects. In the bottom two rows, we report the coefficient on expected end-of-year price (fp) from estimating equation (8) using OLS and IV (respectively) by pooling the data from all firms and plans; in addition to plan by coverage tier and join month fixed effects, these regressions now also include firm by join month fixed effects. Standard errors are clustered on join month by coverage tier by firm. The IV specification makes use of a simulated end-of-year price as an instrument for the expected end-of-year price (see text for details). The coefficient on the instrument in the first stage is 0.56 (standard error 0.024); the F-statistic on the instrument is 524.

[a] Dependent variable is $log(s+1)$ where s is the total medical spending of the employee and any covered family members in their first three months in the plan

[b] Dependent variable is $log(time)$ where "time" is the number of days to first claim by any covered family member, censored at 92 days. We estimate the regressions in columns (3) and (4) by Tobit.

Table 5: Robustness and specification checks

Specification	N	Initial Spending		Time to First Claim	
		Coeff on fp (1)	Std. Err. (2)	Coeff on fp (3)	Std. Err. (4)
(1) Baseline (logs)	102,022	-1.08	(0.29)	0.357	(0.113)
(2) Level	102,022	-394.43	(162.12)	14.842	(4.429)
(3) QMLE Poisson	102,022	-0.70	(0.25)	--	--
(4) OLS (No Tobit)	102,022	--	--	0.114	(0.057)
(5) Varying censor points	102,022	--	--	0.330	(0.114)
(6) Cox proportional hazard model	102,022	--	--	-0.347	(0.109)
(7) Control for Demographics	102,014	-0.98	(0.26)	0.524	(0.121)
(8) Only those who remain for 2nd year	64,398	-0.73	(0.34)	0.161	(0.133)
(9) Dep. Var. measured in 2nd year	64,398	-0.17	(0.31)	--	--
(10) Dep. Var. measured Jan-Mar of 2nd year	64,398	-0.44	(0.26)	0.172	(0.106)

Table reports results from alternative analyses of the relationship between initial medical utilization and expected end-of-year price. Row (1) shows the baseline results from estimating equation (8) by OLS in columns 1 and 2 and by Tobit in columns 3 and 4, as in the penultimate row of Table 4. Alternative rows report single deviations from this baseline as explained below. In row (2) the dependent variables are defined in levels rather than logs. Mean dependent variables are 596.2 dollars (initial spending) and 58.3 days ((censored) time to first claim). In row (3) the dependent variable is defined in levels (not logs) and the regression is estimated by quasi-maximum likelihood Poisson instead of OLS. In row (4) the regression is estimated by OLS rather than Tobit. In row (5) we estimate the same regression as in the baseline, but we now allow the censoring point to vary with join month, from 92 days if the employee joined in October to 334 days if the employee joined in February. In row (6) we estimate a Cox semi-parametric proportional hazard model on the time to first claim (censored at 92 days); note that here a longer time to first claim is indicated by a negative coefficient (a lower "failure" rate). In row (7) we add controls for age, gender, and start year (as well as interactions of each of those with the firm fixed effects) to the baseline specification. In row (8) we estimate the baseline specification on a smaller sample of employees who remain in the firm for the entire subsequent year; in row (9) we estimate the baseline specification on this same sample, but defining the dependent variable based on utilization in the same three months of the subsequent year (i.e., their first full year in the firm); in row (10) we estimate the baseline specification on this same sample but now define the dependent variable based on utilization in January to March of the first full year in the firm.

Table 6: Calibrating δ

	Log (Three month Spending)		Log(Time to First Claim)	
	OLS	IV	Tobit	Tobit IV[a]
A. Estimated semi-elasticity				
Point estimate	-1.08	-0.78	0.36	0.56
CI Lower Bound	-0.50	-0.24	0.14	0.30
CI Upper Bound	-1.67	-1.33	0.58	0.83
Implied delta				
Point estimate	0.20	0.12	0.26	0.67
CI Lower Bound	0.06	0.02	0.06	0.19
CI Upper Bound	0.45	0.28	0.76	1.00

Panel A reports the estimated semi-elasticities of initial medical utilization with respect to the future price; these are taken directly from the last two rows of Table 4. Panel B shows the implied values of δ associated with each estimate based on the calibration exercise described in the text.

[a] We impose 1 for the upper bound of the confidence interval for the implied δ in the Tobit IV case based on our a priori knowledge that δ cannot be higher than 1; no δ less than 1 produces a semi-elasticity as large as 0.83 in our model.

Appendix
Not for publication

Appendix

This appendix describes in more detail the uses we make of data from the RAND Health Insurance Experiment.[32] Appendix A describes our attempt to use the RAND experiment random assignment of out-of-pocket maximums as the basis for an additional test for forward looking behavior. Appendix B discusses our attempt to use the RAND data to approximately the "primitive" price elasticity of demand , to serve as a benchmark for our estimated response to the future price. Appendix C provides a detailed explanation of how we use the RAND data for the calibration exercises described in Section 4.

As explained in the main text, the RAND experiment, conducted in 1974-1981, randomly assigned participating families to health insurance plans with different levels of cost sharing. Each plan was characterized by two parameters: the coinsurance rate (the share of initial expenditures paid by the enrollee) and the out-of-pocket maximum, referred to as the "Maximum Dollar Expenditure" (MDE). Families were assigned to plans with coinsurance rates ranging from 0% ("free care") to 100%. Within each coinsurance rate, families were randomly assigned to plans with MDEs set equal to 5%, 10%, or 15% of family income, up to a maximum of $750 or $1,000.[33]

A. Testing forward-looking behavior using the RAND data

The latter feature of the RAND plan assignment process – random assignment of MDEs – would seem to provide an ideal experimental setting for a test of forward looking behavior since it potentially provides random variation in the future price among individuals assigned to the same coinsurance (and hence the same spot price). While differences in MDEs across individual families were due in part to differences in family income, differences in average MDE and average end-of-year price across plans can be treated as randomly assigned. Appendix Table A6 provides sample counts and various summary statistics for the RAND plans.[34] As the table shows, average MDEs were considerably higher in plans where the MDE was set equal to 10% or 15% of family income than in plans where the MDE was set to 5% of income. These differences generated corresponding differences in the share of families hitting the MDE and in expected end-of-year price (columns (5) and (6)).

Columns (8) and (9) of Appendix Table A6 present results from a regression of time to first claim on expected end-of-year price in the RAND. Specifically, we run the regression

$$\text{Log(Time-to-First-Claim)}_f = \beta \cdot fp_j + \gamma \cdot coins_j + X_f'\chi + \epsilon_f, \tag{9}$$

where fp is the future price (or expected end-of-year price), $coins$ is the coinsurance rate, f indexes families, j indexes plans, and X_f is a vector of dummy variables for site and start month in the

[32] The original RAND investigators have very helpfully made their data publicly available. We accessed the data through the Inter-University Consortium for Political and Social Research.

[33] For a detailed description of the plans and other aspects of the experiment, see Newhouse et al. (1993).

[34] Appendix Table A6 omits the RAND's "individual deductible plans," which had coinsurance rates of 0% for inpatient services and 100% or 95% for outpatient services, because there was no MDE variation within this coinsurance rate structure.

experiment by year.[35] As shown, we run the regression separately for each coinsurance rate group and then pool all groups (or all groups except the free care plan) to maximize power (we also run a specification with a full set of coinsurance rate dummies in place of the coinsurance rate term). We run both OLS and Tobit regressions, where the latter account for censoring of time to first claim at 367 days.

There are two important limitations to this analysis, so that despite its apparent advantages, the RAND variation is in fact inferior to the variation generated by employee hire dates (the primary variation used in the paper). First, as a practical matter, the RAND setting gives us much less power to detect differences in spending by expected end-of-year price. The samples are smaller (with a total sample size of 5,653 family-year observations across all plans, relative to more than 100,000 in the combined employer-provided sample), and there is much less variation in end-of-year price. As a result, most of our estimates based on the RAND data are quite imprecise, although in our most inclusive specification (bottom row of Table A6), we are able to reject the null of no response to the future price.

Second, conceptually, the variation in the MDE in the RAND data is not as clean for testing for forward looking behavior as the variation in the coverage horizon that we use in the paper. To see this, note that differences in expected end-of-year price are correlated with differences in spot price even under the null hypothesis of no forward-looking behavior. Even if people are fully myopic, families in low MDE plans will meet their MDEs sooner and will spend more of the year facing a 0% spot price. As a result, they will have higher spending even if they are not at all forward-looking.[36] Because 12% of families in the lowest MDE, highest coinsurance rate plan hit the MDE within the very first month of the experiment, this is a concern even when looking just at initial (e.g., one month) spending.

We attempted to solve this problem by using time to first claim as the outcome variable. Unfortunately, however, some of RAND's MDE levels are quite low, so they can affect even the spot prices families face when making decisions about their very first health expenditure. To see this, consider two families in plans with a 100% coinsurance rate. The first family has an MDE of \$150, the second an MDE of \$300. Suppose that, before either family has any other health expenditures, each experiences a health shock that would cost \$300 to treat. The out-of-pocket cost of treating this shock would be \$150 for the low MDE family but \$300 for the high MDE family, meaning that the low MDE family faces a spot price of only 50% for the episode, compared to 100% for the high MDE family. Hence, the low MDE family will be more likely to treat the episode, even if both families are fully myopic.

Because about half of outpatient episodes (defined as in Appendix B below) are larger than

[35] Plan assignment was random only conditional on which of the experiment's six sites a family lived at and when the family enrolled in the experiment. For details, see Newhouse et al. (1993, Appendix B).

[36] In contrast, this is not a problem when using the variation in end-of-year price generated by month of hire. If people are fully myopic, then early hires will have the same levels of three-month spending as late hires, and so the two groups will be equally likely to hit their deductibles within three months and will face the same average spot price.

the smallest MDEs in the RAND sample, this problem is potentially quite significant. Indeed, in simulations mimicking the RAND setting, we obtain a large and statistically significant coefficient on end-of-year price in a regression for time to first claim, even when we assume complete myopia. Thus, we conclude that, even apart from the precision problems, we cannot use the RAND setting to generate variation in the future price conditional on the spot price to test for forward looking behavior.[37]

B. Approximating the "primitive" price elasticity using RAND data

Before turning to the model-based calibration exercise in the next sub-section, we first present a loose way of trying to gauge the extent of forward looking behavior by using the experimental variation in contracts in the RAND data to generate an estimate of the "primitive" price elasticity of medical care utilization which we then compare to our previously estimated response to the future price from the main empirical work in the paper.

The variation used in the main empirical work in the paper is not useful in this regard, as we observe neither linear contracts nor identifying variation for plan assignment. The RAND experiment does not provide this ideal variation either, since all of the RAND contracts (except for the free care contract) involve a non-linear pricing schedule; families were randomized into coinsurance rates and then, within each positive coinsurance rate, they were further randomized into plans with an out-of-pocket maximum (known as the "maximum dollar amount" or MDE in the RAND context) of either 5%, 10%, or 15% of income (up to a maximum of $1,000 or $750); above the MDE the price of care is zero.[38] However, RAND's experimental variation (within each coinsurance rate) in the out-of-pocket maximum allows us to estimate its effect, and then to extrapolate out of sample to obtain the behavioral response to a contract where the out-of-pocket maximum is sufficiently high, thus approximating a linear contract.

Specifically, we estimate the regression

$$y_{fj} = \eta_1 \cdot coins_j + \eta_2 \cdot Share\_Hit_j + \eta_3 \cdot coins_j \cdot Share\_Hit_j + v_{fj}, \tag{10}$$

where y_{fj} is a measure of medical utilization by family f in plan j, $coins_j$ is the coinsurance rate of the plan the family was randomized into (which is either 0%, 25%, 50%, or 95%), and $Share\_Hit_j$ is the fraction of families within the same coinsurance rate and MDE assignment that hit (i.e., spent past) the MDE during the year. For the positive coinsurance plans this number ranges from

[37] Two of the original RAND investigators, Keeler and Rolph (1988), also attempt to use the RAND data to test for forward looking behavior, but they use a different empirical strategy. They do not exploit the MDE variation, and instead rely on within-year variation in how close families are to their MDEs. They test whether spending is higher among families who are closer to hitting their MDEs, as would be expected - all else equal - if people are forward looking. They make several modeling assumptions to try to address the (selection) problem that families with higher underlying propensities to spend are more likely to come close to hitting their MDEs. They also assume that individuals have perfect foresight regarding all the subsequent medical expenses within a year associated with a given health shock. They conclude that they cannot reject the null of complete myopia with respect to future health shocks.

[38] All dollar amounts are reported in current (1970s) dollars.

8 percent to 40 percent depending on the plan assignment (see Appendix Table A6, column (5)). The coefficient of interest is η_1, which we interpret as the responsiveness of medical utilization to a change in the coinsurance rate of a linear contract; this involves extrapolating out of sample to where $Share\_Hit_j = 0$, which would be the case for a sufficiently high MDE.

Because the share of families in a given plan assignment that hit the MDE depends on family spending behavior, which itself may be endogenous to plan assignment, we also present IV specifications in which we instrument for the share of families in a given plan that hit the MDE with the simulated share hitting the MDE. The "simulated share" is calculated as the share of the entire (common) sample of individuals across all plans that would have hit the MDE if assigned to the given plan, in a similar spirit to the IV exercise we reported in the previous section.

Appendix Table A7 presents the results. Our sample size is approximately 1,500 families (about 5,600 family-years).[39] As in the previous analysis, we analyze both the responsiveness of the first three months of spending and the time to first claim. Here, we also add total (annual) spending as an additional outcome (as the proportional response to a linear contract should not, in principle, be different for initial and total spending).

The response to the linear coinsurance – while fairly imprecise in most specifications – can now be compared to our estimates of the response to the future price from the previous sub-section. Using the IV specification, we find a spending semi-elasticity with respect to the price of a linear contract that ranges from -1.2 to -1.7, which is roughly twice as large as the semi-elasticity of -0.78 with respect to the future price that we found in last section (see last row of Table 4). Similarly, we estimate that the semi-elasticity of the time to first claim with respect to the price of a linear contract is 0.51, which is virtually identical to our analogous semi-elasticity estimate of 0.56 with respect to the future price. Thus, overall the results are indicative of substantial, but perhaps not full, forward looking behavior.

C. Model calibration

In Section 4 of the main text, we explain how we use the RAND data to calibrate a model that allows us to map the estimated elasticity of initial spending with respect to the end-of-year price to the parameter δ. Here, we provide more details about this calibration exercise.

Calibrating the medical shock process (λ and $G_1(\theta)$) We calibrate the medical shock process using data from the 620 families (approximately 2,400 family-years) participating in the RAND's "free care" plan. We calibrate the distribution of inpatient and outpatient shocks separately and also allow for heterogeneity across families in the distribution of shocks. Specifically, letting f index families and ζ index types of spending (inpatient or outpatient), we assume that in each period t,

[39] Appendix Table A6 shows the exact plans we study and the distribution of families across those plans. The entire RAND experiment involved about 2,400 families. We exclude from this analysis the approximately 400 families randomized into the 95% coinsurance plan with a fixed ($150 per person) MDE plan (also know as the "individual deductible" plan) because for this MDE only the coinsurance rate differed (it was 95% for outpatient care but free for inpatient care), and the approximately 400 families randomized into an HMO.

family f draws a shock of type ζ with probability $\lambda_{f\zeta}$. In periods where a family does experience shocks of type ζ, the shocks are drawn i.i.d. from a lognormal distribution with mean $\mu_{f\zeta}$ and variance σ.

Our procedure for obtaining the various spending distribution parameters is as follows. We define a period t as a week. We group together all claims of a given type separated by less than one week and define each grouped set of claims as one episode, assigning it to the first week of the episode; this generated about 6,000 inpatient episodes and about 77,000 outpatient episodes over the course of the entire experiment. For each family and each type of spending, we then compute: $\lambda_{f\zeta}$ as the share of weeks (over the course of the entire experiment[40]) in which family f experienced an episode of type ζ, we set $\mu_{f\zeta}$ as the average size of family f's episodes of type ζ, and $\sigma_{f\zeta}$ as the variance of family f's episodes of type ζ. Because $\sigma_{f\zeta}$ is extremely noisy (even more so than $\mu_{f\zeta}$) and because it is unavailable for families with only one shock of a given type, we set σ to be the average of $\sigma_{f\zeta}$ for all families.

Partly to reduce noise and partly to make simulating the model computationally feasible, we next divide families into five-percentile groups based on their values of λ. We replace each value of $\lambda_{f\zeta}$ and $\mu_{f\zeta}$ with the mean of the respective variable for family f's percentile group. This approach eliminates cases where the probability of outpatient shock is zero, but leaves 55% of the sample with a zero probability of inpatient medical shocks. This is consistent with our intuition that every family faces some meaningful risk of experiencing an outpatient shock, but some families (specifically, those who experience no inpatient episodes at any point during the experiment) may face so little risk of an inpatient episode that they perceive it as approximately zero.

Appendix Figure A1 and Appendix Table A8 compare the actual distributions of expenditures in the free care plan with the simulated distributions. Appendix Figure A1 presents a histogram of total (the sum of inpatient and outpatient) spending, while Appendix Table A8 reports the means and standard deviations of log inpatient, outpatient, and total spending, as well as the share of families with no inpatient, outpatient, or total spending over the course of a year ($T = 52$).[41]

The fit is notably better for outpatient than inpatient spending, basically because, as others have also found (see, e.g., French and Jones, 2004), the lognormal distribution is a better fit for outpatient than inpatient spending. Nonetheless, the fit is fairly good for both categories of spending, and seems (to us) to capture the main properties of health spending for the purpose of our calibration exercise.

Calibrating the distribution of valuations ($G_2(\omega/\theta)$) Recall that ω represents the (monetized) health cost of a given shock, and so ω/θ represents the health cost of a given shock relative to the cost of treatment. For example, if $\omega/\theta = 0.5$ then the health cost of not treating a given shock is equal to half the cost of the treatment.

We calibrate the distribution of ω/θ for outpatient shocks, but assume $\omega/\theta = 1$ for all inpatient shocks. That is, we assume that individuals treat all inpatient shocks, regardless of what share of

[40] Families participated in the experiment for periods of either three or five years.

[41] Throughout, we define log spending as $log(spending + 1)$ in order to avoid missing values.

the cost of treatment they pay out of pocket. This analytic choice is done primarily to make the calibration exercise much more feasible (inpatient shocks are sufficiently rare relative to outpatient shocks that it is much harder to use the data to calibrate ω/θ for them). It also reflects our intuition that most health shocks for which treatment requires hospitalization are much less discretionary than outpatient care; this is consistent with the basic findings from the RAND experiment itself (Newhouse et al., 1993) as well as subsequent quasi-experimental evidence (e.g., Einav et al., 2011) and our findings in this paper that only outpatient care appears to respond to the future price (Appendix Table A2).

For outpatient shocks, we assume that ω/θ follows a Beta distribution with parameters a and b, so that $\omega/\theta \sim \beta(a, b)$. Thus, $a > 0$ and $b > 0$ are the key primitive price elasticity parameters of the model. The ratio $a/(a + b)$ gives the mean value of ω/θ. We use data on the 95%, 50%, and 25% coinsurance RAND plans to calibrate a and b.[42] We simulate the model described in Section 2 to generate utilization data for each coinsurance rate. We then try to match three moments of the actual RAND data for each coinsurance rate: the mean of log spending, the standard deviation of log spending, and the share of the sample with zero spending.[43] Specifically, we minimize the sum of squared differences, weighting by the different coinsurance rates' RAND sample sizes.

So far, we have glossed over a tension with our calibration strategy. Namely, that the distribution of spending (using the model of Section 2) depends not only on the distributions of λ, θ, and ω/θ, but also on δ. And yet our goal is to obtain the parameters of the ω/θ distribution without knowing δ so that we can then determine what value of δ yields the elasticities we obtained from the employer-provided data.

Our strategy succeeds simply because it happens that the objective function is quite flat in δ but quite steep (and generally invariant to δ) in a and b. Panel A of Appendix Table A9 shows, for 11 values of δ ranging from zero to one, the optimal values of a and b and the resulting values of the objective function. As the table shows, the model selects very similar values of a and b regardless of the assumed value of δ, and yields similar values of the objective function (at the optimal values of a and b). Basically, whatever the choice of δ, the best fit involves $E[\omega/\theta] \approx 0.55$ and a highly bimodal distribution for ω/θ, with modes near 0 and 1.[44]

Based on eyeballing the simulation results, we select $a = 0.3$ and $b = 0.25$ for our calibration exercise; these are the values that minimize the objective function averaged over the possible values of δ we examine. Panel B of Appendix Table A9 shows that these values yield a fairly tight fit to the RAND data for any assumed value of δ.

[42] We do not make use of the data from the "mixed coinsurance rate" plans included in Appendix Table A6. Incorporating these plans would have required further complicating the model in order to introduce multiple types of outpatient spending.

[43] As before, we define log spending as $log(spending + 1)$ to avoid missing values.

[44] Intuitively, the bimodal distribution reflects the fact that the sample means from the RAND data are almost the same for the 25% and 50% coinsurance rate plans, implying that, for most outpatient shocks, people either will not treat the shock at a coinsurance rate of 25% or will treat it unless the coinsurance rate is quite high.

Mapping the elasticity of initial spending with respect to end-of-year price to δ Having calibrated the key elements of the model, the final step in our calibration exercise is to simulate the data generating process from our employer-provided data and obtain estimates of the responsiveness of initial spending to the expected end-of-year price in the simulated data.

We consider plans with deductible of \$0, \$250, \$750, and \$1,000, with no cost-sharing above the deductible. For each of these plans, we use the calibrated parameters described above, and a range of values for δ (the only remaining free parameter), and simulate spending given time horizons of 47, 42, 37, 32, 27, 22, 17, or 12 weeks (analogous to hire dates ranging from February to October). For each of 10,000 simulated families in each deductible-horizon combination, we obtain simulated spending in the first 12 periods (analogous to first three month spending) and time to first claim (here measured in weeks and censored at 12); in addition, for each deductible/horizon combination, we obtain average "end-of-year"price (here, just average price at the end of the horizon).

Letting d denote levels of the deductible, h index horizon lengths, and f index families, we use the simulated data to estimate the regression

$$Outcome_i = \beta \cdot fp_{dh} + \gamma_d + \epsilon_f. \tag{11}$$

Here, β is the coefficient of interest, while the γ_d's are dummy variables for deductible level. We estimate the regression for log("three month") spending (spending in the first 12 periods) and for time to first claim. For reasons explained in the main text, we run both OLS and IV regressions, in the latter case instrumenting for fp_{dh} with the average end-of-period price after h periods among families with the maximum time horizon (47 weeks).

We repeat the above exercise for 101 values of δ ranging from 0 to 1. We can then obtain point estimates and confidence intervals for δ by comparing the estimates of the responsiveness of initial spending to end-of-year price obtained in the simulations with the estimates and the bounds of the confidence intervals obtained from the employer-provided data.[45] The results are presented in Figure 3 in the main text.

[45] Technically, the confidence intervals on δ should also take into account the standard errors on β from the regressions in the simulated data. However, because we can make the simulations so large – we simulate 10,000 families for each deductible horizon paid – the standard errors on β are effectively zero. β simply describes the relationships imposed by the model and the calibrated parameters.

Appendix Figure A1: Fit of the calibration exercise of medical events

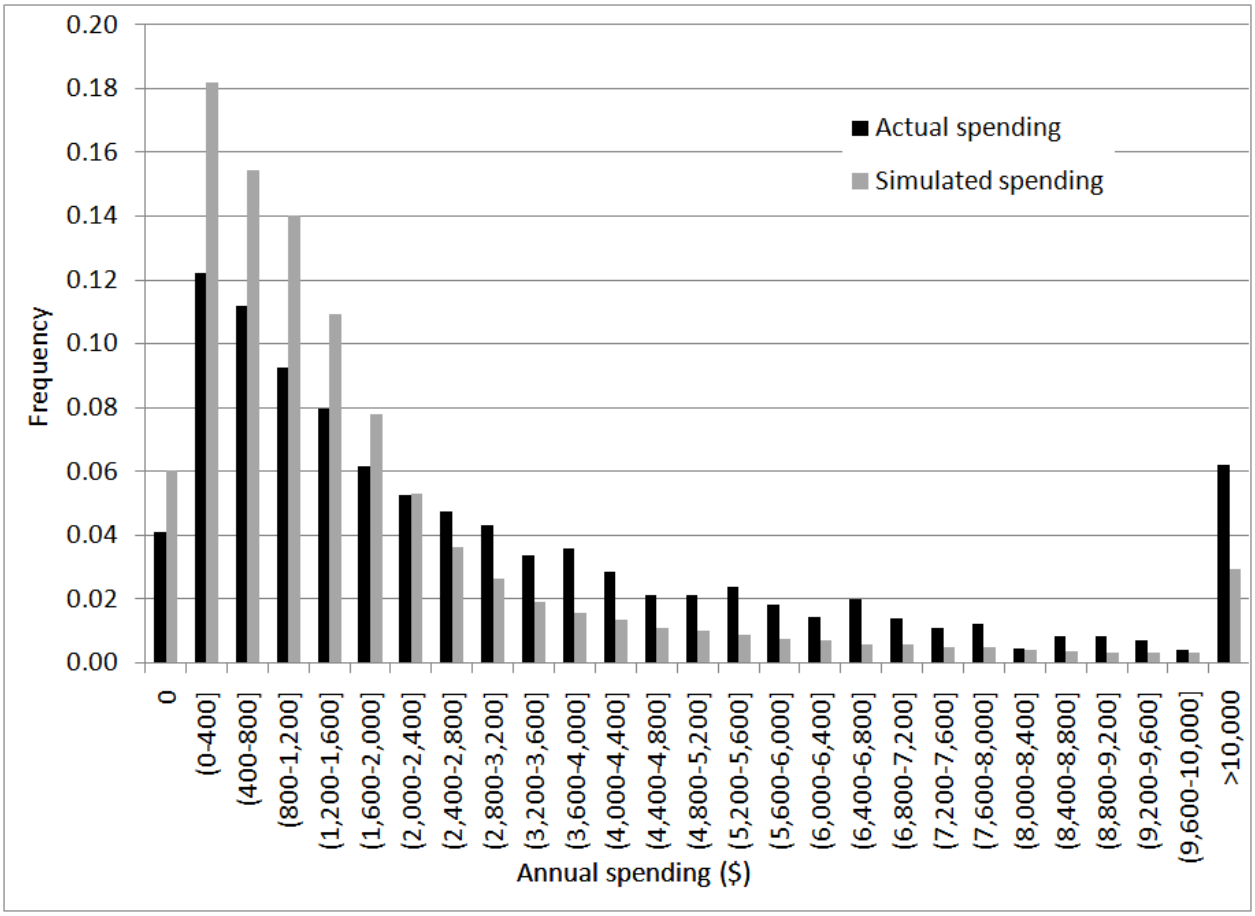


Figure shows the distribution of annual medical spending in the "free care" RAND data based on the actual (black bars) and simulated (gray bars) data. The simulations use the calibrated parameters, as explained in the Appendix. The actual data is based on the 2,376 family-years of data in the free care plan.

Appendix Table A1: Additional plan details

Employer	Plan	Years offered	Mid-year new enrolees[a]	In-network features						Out-of-network features					
				Deductible ($)		Coins[b]	Copay ($)	Stop loss ($)		Deductible ($)		Coins[b]	Copay ($)	Stop loss ($)	
				Single	Family			Single	Family	Single	Family			Single	Family
Alcoa	A0	2004-07	3,269	0	0	0.10	0	2,500	5,000	250	500	0.3	0	5,000	10,000
	A1	2004-07	3,542	250	500	0.10	0	2,750	5,500	500	1,000	0.3	0	5,500	11,000
Firm B	B0	2001-05	37,759	0	0	0.00	15	0	0	250	500	0.2	0	1,250	2,500
	B1	2001-05	9,553	150	300	??	??	??	1,100	??	??	??	0	??	??
Firm C	C0	1999-05	27,968	0	0	0.00	15	0	0	300	750	0.3	0	3,000	6,000
	C1	1999-00	6,243	200	500	0.10	0	1,000	2,000	??	??	0.3	0	3,750	7,500
	C2	2001-02	8,055	250	625	0.10	0	1,250	2,500	250	625	0.3	0	3,900	7,800
	C3	2004-05	5,633	300	750	0.10	0	1,300	2,600	300	750	0.3	0	3,900	7,800

"??" denotes an unknown feature of a plan.

[a] The sample includes employees who enroll in February through October.

[b] Coinsurance denotes the fraction of medical expenditures the insured must pay out of pocket after hitting the deductible and prior to reaching the "stop loss."

Appendix Table A2: Responsiveness of different types of care to the future price

	Dependent variable	Mean of the dep. var.	Coeff. on future price	Std. Error
(1)	Log initial spending	3.32	-1.08	(0.29)
(2)	Log initial outpatient spending	3.29	-1.06	(0.29)
(3)	Initial spending	596.2	-394.4	(162.1)
(4)	Initial outpatient Spending	445.0	-375.8	(107.7)
(5)	Initial inpatient Spending	147.5	-19.8	(99.1)
(6)	Any initial inpatient spending	0.014	-0.008	(0.006)

Table reports results for different types of medical spending of the analysis of the relationship between initial medical spending and expected end-of-year price ("future price"). All rows show the results from estimating equation (8) by OLS using different dependent variables; in addition to "future price" the covariates in this regression include plan by coverage tier fixed effects, join month fixed effects and firm by join month fixed effects. Standard errors are clustered on join month by coverage tier by firm. The first row shows the baseline results (see penultimate row in Table 4) for the dependent variable log initial spending (plus 1). In row 2 the dependent variable is the log of initial outpatient spending (plus 1). Rows 3 through 5 show results for the level of initial medical spending, the level of initial outpatient spending and the level of initial inpatient spending respectively. The last row shows the results for an indicator variable for any initial inpatient spending. "Initial" spending is defined as spending in the first three months of the plan for all covered members of the plan. $N = 102,022$.

Appendix Table A3: Additional robustness exercises

	Specification	N	Log Initial Spending		Log Time to First Claim	
			Coeff on fp	(S.E.)	Coeff on fp	(S.E.)
(1)	Baseline	102,022	-1.08	(0.29)	0.357	(0.114)
	Panel A: Altnerative sets of fixed effects					
(2)	Don't limit to within firm	102,022	-1.07	(0.30)	0.320	(0.121)
(3)	Don't control for Tier	102,022	-3.98	(0.76)	1.943	(0.373)
(4)	Tier x firm interactions	102,022	-1.04	(0.29)	0.355	(0.114)
	Panel B: Family vs Single Tier					
(5)	Family Tier	43,358	-0.90	(0.42)	0.132	(0.124)
(6)	Single Tier	58,664	-1.15	(0.40)	0.579	(0.193)
	Panel C: Using Additional Plan Characteristics to construct mp					
(7)	Baseline (Firms A and C)	54,710	-0.81	(0.32)	0.263	(0.127)
(8)	Firms A and C, refined fp measure	54,710	-0.90	(0.36)	0.293	(0.141)

Table reports results from alternative analyses of the relationship between initial medical utilization and expected end of year marginal price. The first row shows the baseline results (see penultimate row in Table 4) from estimating equation (8) which pools the data across firms. In addition to the expected end of year marginal price, the regressions also include plan by coverage tier fixed effects, join month fixed effects and firm by month fixed effects. Standard errors are clustered on join month by coverage tier by firm. Alternative rows report single deviations from this baseline. In Row 2 we remove the firm by join month fixed effects from the baseline. In Row 3 we remove the controls for coverage tier (so that there are plan fixed effects but not plan by coverage tier fixed effects) from the baseline. In row 4 we add firm by coverage tier fixed effects and firm by coverage tier by join month fixed effects to the baseline. In rows 5 and 6 we stratify the sample by coverage tier. In Panel C we limit the analysis to the two firms (Alcoa and Firm C) in which we observe the in-network coinsurance rate for all plans (see Appendix Table A1 for details). Row 7 reports the baseline results limited to those two firms; Row 8 shows the sensitivity to using a refined measure of future price which accounts for the coinsurance rate (see Appendix Table A4 for details). As expected, not accounting for the coinsurance rate in our baseline future price measure (row 7) biases downward our estimated impact of the future price (compare rows 7 and 8).

Appendix Table A4: Alternative construction of future price

Employer	Plan	Expected end-of-year price[a]			Refined expected end-of-year price[b]		
		Joined plan in			Joined plan in		
		Feb-Apr	May-Jul	Aug-Oct	Feb-Apr	May-Jul	Aug-Oct
Alcoa	A0	0.000	0.000	0.000	0.0994	0.0995	0.0997
	A1	0.512	0.603	0.775	0.560	0.643	0.798
Firm C	C0	0.000	0.000	0.000	0.000	0.000	0.000
	C1-C3	0.543	0.633	0.811	0.589	0.670	0.830

[a] Expected end-of-year price is equal to the fraction of individuals who do not hit the deductible by the end of the calendar year (and therefore face a marginal price of 1). It is computed based on the plan's deductible level(s), join month, and the annual spending of all the employees in one's plan and join month; we compute it separately for family and single coverage within a plan and report the enrollment-weighted average.

[b] "Refined" expected end-of-year price is equal to the coinsurance rate times the fraction of individuals who hit the deductible but not the out-of-pocket maximum by the end of the year (and therefore face a marginal price equal to the coinsurance rate) + the fraction of individuals who do not hit the deductible by the end of the calendar year (and therefore face a marginal price of 1.) The refined expected end-of-year price is computed in the same manner as described above for the expected end-of-year price.

Appendix Table A5: Differences in observables by plan and join month

Employer	Plan	Deductible (Single/Family) [N = enrollees]	Indicator for Old (>=45) Difference (1)	Indicator for Old (>=45) DD (2)	Indicator for Female Difference (3)	Indicator for Female DD (4)
Alcoa	A0	0	-0.009		-0.011	
		[N = 3,269]	(0.004)		(0.003)	
	A1	250/500	-0.008	0.0020	-0.002	0.009
		[N = 3,542]	(0.002)	(0.0041)	(0.003)	(0.004)
Firm B	B0	0	-0.004		-0.003	
		[N = 37,759]	(0.003)		(0.002)	
	B1	150/300	-0.010	-0.0059	-0.004	-0.001
		[N = 9,553]	(0.004)	(0.0026)	(0.004)	(0.003)
Firm C	C0	0	-0.014		0.009	
		[N = 27,968]	(0.002)		(0.002)	
	C1-C3	200-300/500-750	-0.019	-0.0045	0.009	0.000
		[N = 19,931]	(0.003)	(0.0032)	(0.003)	(0.003)

Table reports coefficients (and standard errors in parentheses) from regressing the dependent variable on join month (which ranges from 2 (February) to 10 (October)). The dependent variables are demographic characteristics (defined in the top row) with overall means for "old" (i.e. age 45+) of 0.27 and for female of 0.48. Columns (1) and (3) report the coefficient on join month separately for each plan, based on estimating equation (4); the regressions also include an indicator variable for coverage tier (single vs. family). Columns (2) and (4) report the difference-in-differences coefficient on the interaction of join month and having a deductible plan, separately for each firm, based on estimating equation (5); the regressions also include plan by coverage tier fixed effects and join month fixed effects. Standard errors are clustered on join month by coverage tier.

Appendix Table A6: Summary statistics of the RAND data

Coinsurance Rate	Maximum Dollar Expenditure (MDE)	Number of family years (Number of families in year 1)	Avgerage MDE (Adjusted[a])	Share of family-years who hit the MDE	Expected End-of-Year Price[b]	Avg. Time to First Claim (Days)[c]	Effect of End-of-Year Price on Log(Time to First Claim)[d] OLS	Tobit
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Plan-by-plan analysis								
100%	5% of income up to $1,000	33 (33)	$533	0.33	0.67	70	-2.94	-2.66
	10% of income up to $1,000	29 (29)	$801	0.21	0.79	82	(2.65)	(2.84)
	15% of income up to $1,000	33 (33)	$794	0.21	0.79	64		
95%	5% of income up to $1,000	418 (84)	$559	0.40	0.57	88	-0.10	-0.70
	10% of income up to $1,000	342 (80)	$746	0.34	0.63	86	(2.19)	(2.41)
	15% of income up to $1,000	470 (101)	$817	0.33	0.63	99		
50%	5% of income up to $1,000	111 (26)	$535	0.28	0.36	58	4.16	5.30
	10% of income up to $1,000	76 (17)	$779	0.16	0.42	84	(4.98)	(5.25)
	15% of income up to $1,000	308 (84)	$847	0.19	0.40	80		
50% for dental & mental health; 25% for all other	5% of income up to $750	189 (41)	$499	0.28	0.22	78		
	10% of income up to $750	226 (44)	$584	0.31	0.22	59		
	15% of income up to $750	159 (30)	$689	0.16	0.26	62	4.77	4.89
	5% of income up to $1,000	18 (18)	$523	0.28	0.23	27	(3.82)	(3.80)
	10% of income up to $1,000	19 (19)	$600	0.16	0.26	40		
	15% of income up to $1,000	13 (13)	$837	0.08	0.29	65		
25%	5% of income up to $750	192 (22)	$518	0.17	0.21	73		
	10% of income up to $750	208 (31)	$617	0.17	0.21	61		
	15% of income up to $750	207 (26)	$683	0.18	0.21	61	0.09	-2.26
	5% of income up to $1,000	86 (52)	$535	0.14	0.22	71	(21.71)	(22.15)
	10% of income up to $1,000	70 (43)	$818	0.11	0.22	38		
	15% of income up to $1,000	70 (44)	$816	0.16	0.21	37		
0%	--	2,376(620)	--	1.00	0.00	46		
Panel B: Pooling across plans								
All positive coinsurance plans, with coinsurance dummies		3,277 (870)					-0.36 (1.51)	-0.08 (1.64)
All positive coinsurance plans, pooled		3,277 (870)					0.52 (1.12)	0.73 (1.23)
All plans, pooled		5,653 (1,490)					1.90 (0.83)	1.96 (0.90)

[a] Regression adjusted for differences in site, start month, and year across plans (see Newhouse et al. (1993, Appendix B) for more details).

[b] Expected end-of-year price equals the share of families not hitting the MDE (in the given plan) times the coinsurance rate. For the mixed coinsurance rates plans, we weight the two coinsurance rates based on their shares of initial claims in the full sample; 25% of initial claims are for mental/dental.

[c] For families with no claims in a given year, time to first claim is coded as 367.

[d] Columns (8) and (9) show the coefficient on the expected end-of-year price (fp) from estimating variants of equation (9). In Panel A we regress log time-to- first-claim on the expected end-of-year price (see column (6)) and site and enrollment month by year dummies; plan assignment in the RAND experiment was random conditional on the location (site) and when the family enrolled in the experiment (see Newhouse et al. (1993, Appendix B) for more details). In Panel B we pool across plans and therefore add additional controls in the form of either coinsurance dummies (first row) or the coinsurance level directly (bottom two rows); the final row adds in the free care (0% coinsurance) plan. Standard errors are clustered on family.

Appendix Table A7: Approximating the response to a linear contract in the RAND data

Regressor	Dependent Variable					
	Log Initial Spending[a]		Log Annual Spending[b]		Log Time to First Claim[c]	
	OLS	IV	OLS	IV	Tobit	Tobit IV
Coins. Rate	-1.21	-1.19	-1.78	-1.65	0.88	0.51
	(0.73)	(1.03)	(0.73)	(1.03)	(0.55)	(0.49)
Share hit MDE	0.45	0.43	0.20	0.21	-0.23	-0.25
	(0.21)	(0.25)	(0.20)	(0.24)	(0.15)	(0.12)
Coins. Rate * Share Hit MDE	0.58	0.52	1.76	1.40	-0.54	0.14
	(1.79)	(2.53)	(1.78)	(2.53)	(1.32)	(1.21)

Sample consists of 5,653 family-years (1,490 unique families) in the RAND data in one of the positive coinsurance plans or the free care plan. "Share hit MDE" is the share of families in a given coinsurance and maximum dollar expenditure (MDE) plan who spend past the MDE during the year. Because plan assignment in the RAND experiment was random only conditional on site and month of enrollment in the experiment, all regressions control for site and start month fixed effects (see Newhouse et al. (1993, Appendix B) for more details). All regressions cluster standard errors on the family, except for the Tobit IV specifications, which is estimated using a minimum distance estimator (Newey, 1987). In the IV specifications, we instrument for the share of families in a given coinsurance and MDE plan who hit the MDE with the "simulated" share hitting the MDE; the "simulated" share is calculated as the share of the full ($N = 5,653$) sample which, given their observed spending, would have hit the MDE if (counterfactually) assigned to the given plan; the coefficient on the instrument in the first stage is 1.05 (standard error 0.003); the F-statistic on the instrument is 120,000. Appendix Table A6 provides more details on the plans in the RAND experiment, the distribution of the sample across the different plans, and the share of families who hit the MDE in each plan.

[a] Dependent variable is $log(s+1)$ where s is the total medical spending of the employee and any covered family members in their first three months in the plan.

[b] Dependent variable is $log(s+1)$ where s is the total medical spending of the employee and any covered family members in their full year in the plan.

[c] Dependent variable is $log(time)$ where "*time*" is the number of days to first claim by any covered family member, censored at 367 days

Appendix Table A8: Fit of the calibration exercise of medical events

	Total Spending		Inpatient		Outpatient	
	Actual	Simulated	Actual	Simulated	Actual	Simulated
Mean of log spending	6.57	6.53	2.08	1.61	6.18	6.06
Standard deviation of log spending	2.17	2.10	3.58	3.28	1.96	2.04
Share with any spending	93.7%	93.8%	25.8%	19.7%	93.5%	92.5%

Table reports summary statistics of the actual and simulated moments of the spending distribution for the RAND "free care" plan. Log spending is computed as log(spending+1) to avoid missing values. Simulated data are generated as described in Appendix B.

Appendix Table A9: Calibration and fit of the "primitive" price elasticity parameters

Imposed value of δ	Value of obj. fn at the optimum	Value of *a* at the optimum	Value of *b* at the optimum	Implied E(ω)
0	19.9	0.30	0.20	0.60
0.1	10.1	0.25	0.20	0.56
0.2	15.5	0.25	0.20	0.56
0.3	11.6	0.30	0.25	0.55
0.4	11.9	0.30	0.25	0.55
0.5	14.0	0.35	0.30	0.54
0.6	16.6	0.35	0.30	0.54
0.7	24.5	0.35	0.30	0.54
0.8	34.6	0.35	0.35	0.50
0.9	28.7	0.35	0.35	0.50
1	29.7	0.35	0.35	0.50

Imposed value of δ	Value of obj. fn	Mean log spending			Std. Dev. of log spending			Share with zero spending		
		Coins. 25%	Coins. 50%	Coins. 95%	Coins. 25%	Coins. 50%	Coins. 95%	Coins. 25%	Coins. 50%	Coins. 95%
Actual (observed moments)		6.08	6.04	5.53	2.36	2.35	2.71	90.7%	90.7%	85.2%
0	113.9	6.09	5.92	5.28	2.28	2.39	2.81	91.8%	90.6%	84.0%
0.1	45.3	6.09	5.93	5.39	2.28	2.39	2.77	91.8%	90.6%	84.8%
0.2	20.1	6.09	5.94	5.46	2.28	2.39	2.75	91.8%	90.6%	85.3%
0.3	11.6	6.10	5.95	5.52	2.28	2.39	2.73	91.8%	90.7%	85.7%
0.4	11.9	6.10	5.95	5.57	2.28	2.39	2.72	91.8%	90.7%	86.0%
0.5	17.7	6.10	5.96	5.61	2.28	2.39	2.71	91.8%	90.7%	86.3%
0.6	27.6	6.10	5.97	5.65	2.28	2.39	2.70	91.8%	90.7%	86.5%
0.7	40.3	6.11	5.98	5.68	2.28	2.38	2.69	91.8%	90.7%	86.7%
0.8	55.8	6.11	5.99	5.72	2.28	2.38	2.68	91.9%	90.8%	86.9%
0.9	74.1	6.11	6.00	5.75	2.28	2.38	2.67	91.9%	90.8%	87.1%
1	97.5	6.11	6.01	5.79	2.28	2.38	2.67	91.9%	90.8%	87.2%

The top panel reports the values of a and b that minimize the objective function for different values of δ. The bottom panel reports goodness of fit measures for our choice of $a = 0.3$ and $b = 0.25$ for different values of δ. Log spending is computed as log(spending+1) to avoid missing values. Simulated data are generated as described in Appendix B.