NBER WORKING PAPER SERIES

INTERTEMPORAL SUBSTITUTION IN HEALTH CARE DEMAND: EVIDENCE FROM THE RAND HEALTH INSURANCE EXPERIMENT

Haizhen Lin
Daniel W. Sacks

Working Paper 22802
http://www.nber.org/papers/w22802

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2016

We thank our discussant Chris Cronin, as well as Guy Arie, Anthony Defusco, Johannes Haushofer, Anita Mukherjee, Jeff Prince, Eric Rasmusen, Marc Rysman, Brad Shapiro, Jonathan Skinner, Matthijs Wildenbeest, and audiences at Indiana University, the University of Rochester, the Junior Health Economics Summit, and the Midwest Health Economics Conferences for helpful comments and feedback. We are grateful to the RAND investigators for providing publicly available, well documented replication data. Jordan Keener provided excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Intertemporal Substitution in Health Care Demand: Evidence from the RAND Health Insurance Experiment
Haizhen Lin and Daniel W. Sacks
NBER Working Paper No. 22802
November 2016
JEL No. D12,G22

ABSTRACT

Nonlinear cost-sharing in health insurance encourages intertemporal substitution be- cause patients can reduce their out-of-pocket costs by concentrating spending in years when they hit the deductible. We test for such intertemporal substitution using data from the RAND Health Insurance Experiment, where people were randomly assigned either to a free care plan or to a cost-sharing plan which had coinsurance up to a maximum dollar expenditure (MDE). Hitting the MDE—leading to an effective price of zero—has a bigger effect on monthly health care spending and utilization than does being in free care, because people who hit the MDE face high future and past prices. As a result, we estimate that sensitivity to short-lasting price changes is about twice as large as sensitivity to long-lasting changes. These findings help reconcile conflicting estimates of the price elasticity of demand for health care, and suggest that high deductible health plans may be less effective than hoped in controlling health care spending.

Haizhen Lin
Department of Business Economics and Public Policy
Kelley School of Business
Indiana University
1309 East Tenth Street
Bloomington, IN 47405
and NBER
hzlin@indiana.edu

Daniel W. Sacks
Indiana University
dansacks@indiana.edu

1 Introduction

Most health insurance contracts generate nonlinear pricing of health care, as the marginal price of health care typically falls with health care spending due to deductibles and out-of-pocket spending limits. This nonlinear pricing creates an incentive for intertemporal substitution, because patients can minimize their out-of-pocket costs by concentrating spending in years when they have hit the deductible. Little is known, however, about whether patients in fact respond to the dynamic incentives of nonlinear pricing by substituting health care across periods.

Studying intertemporal substitution is important for at least two reasons. First, with the important exceptions of Einav et al. (2015) and Cabral (2016), most of the literature on health care demand and the effect of insurance on spending has neglected intertemporal substitution, and estimated in a wide range of price elasticities, from as small as -0.2 to as large as -1.5 (for example, Manning et al. (1987); Eichner (1998); Zweifel and Manning (2000); Cardon and Hendel (2001); Bajari et al. (2014); Dalton (2014) and Kowalski (2015, 2016))). Most papers assume that health care decisions are made statically on an annual basis, meaning that there is no scope for future prices to affect current demand. Estimated. However, if there is intertemporal substitution, then patients may respond very differently to a temporary price change than to a long-lasting one. Therefore, depending on the sources of price variation used for identification, one might draw dramatically different conclusions regarding price sensitivities. Accounting for intertemporal substitution makes it possible to separate elasticities with respect to permanent or temporary price changes, and may help reconcile the disparate elasticity estimates in the literature.[1]

Second, intertemporal substitution affects the response to high deductible health plans, which are now common in the American health insurance landscape. Regulators, insurers, and policymakers tolerate the weak risk protection of high deductible health plans out of

[1] Intertemporal substitution has been studied in other settings. For example, Hendel and Nevo (2006a,b) show clear evidence of intertemporal substitution and find that failing to account for it leads to biased price sensitivities. The labor supply literature has also noted that different sources of price variation—anticipated or unanticipated, and permanent or temporary wage changes—yield different responses (Blundell and MaCurdy, 1999). In that context, short- and long-lasting wage changes generate differential elasticities because of income effects, not stockpiling of leisure. In the health care context, we expect that income effects are less important, but stocking up on deferrable care is more important.

the hope that they will reduce health care spending. This view implicitly assumes that care foregone in one year because of the high deductible represents a permanent reduction in health care spending. But if patients are deferring needed care, then their spending may be higher in future years, either because deferrable problems become so severe that they must be addressed, regardless of the cost, or because once patients finally do hit the deductible, they stock up on care, retiming deferrable procedures to a year when their price is low. Thus a key question for the effectiveness of high deductible health plans is whether patients intertemporally substitute in their demand for health care. Evidence of intertemporal substitution would suggest that high-deductible health plans may not be as effective as hoped in controlling overall health care spending.

We begin in Section 2 by developing a test of the null hypothesis of no intertemporal substitution in health care demand. The key to our test is that, in the absence of intertemporal substitution, demand depends only on current prices, so people who hit the deductible and face a temporarily low price should spend at the same level as people who permanently face the same low price. The "price" of health care is difficult to define when patients face a nonlinear price schedule, however, because it is not clear whether patients respond to the expected end-of-year price, as a forward looking patient would (Ellis, 1986), or to the spot price of the next dollar of care, as a fully myopic patient would (Aron-Dine et al., 2015; Einav et al., 2015; Brot-Goldberg et al., 2015; Abaluck et al., 2015; Dalton et al., 2015). We sidestep this issue by looking at spending in the last month of the coverage year, when essentially all uncertainty has been resolved, making the spot price and expected end-of-year price nearly the same.

Implementing our test requires addressing two challenges. First, patients select into insurance plans. Patients with high anticipated spending are likely to select more generous plans, confounding any naive comparison between high deductible plans and more comprehensive ones. Second, hitting the deductible is endogenous to spending, making it hard to draw any casual inference regarding how spending respond to prices. We solve both challenges by taking advantage of the RAND Health Insurance Experiment (Newhouse and The Insurance Experiment Group, 1993). As we describe in Section 3, the experiment entailed random assignment to either a free care plan, or to one of several plans with nonlinear cost-sharing, that entailed a fixed coinsurance rate up to a maximum dollar expenditure (MDE). This random assignment solves the selection problem.

To solve the endogeneity problem, instead of examining those who hit the MDE, in Section 4 we compare spending of *all* patients in cost-sharing to all patients in free care. Although overall spending is much higher in free care, we find that by the end of the coverage year, average spending in cost-sharing is slightly higher than in free care, even though most people do not hit the MDE. This implies that spending is much higher among those who actually hit the MDE than it would have been had they faced a permanent price of zero. Quantity response, as measured by episodes of care, appear similar. The data therefore reject the null hypothesis of no intertemporal substitution. The evidence supporting intertemporal substitution is particularly strong for dental care and medically deferrable care, which is straightforward to retime, and weakest for acute care, which cannot be retimed.

We provide in Section 5 a simple quantification of how much health care demand responds to unanticipated and permanent price changes. A one-time, unanticipated price decrease of 0.1 (i.e. a 10 percentage point decrease in the coinsurance rate) increases spending by $17 per month (12% of the mean), and increases the number episodes by 0.036 (7% of the mean). If instead this price change were permanent, then the monthly spending effect would only be $7.1, and the episode effect 0.029. Overall, short-run effects are about twice as large as long-run effects, and the difference is much greater for dental care and for deferable procedures than for acute care. We decompose the divergence between short and long run price sensitivities into an anticipation effect (during sales, people stock up on care health care in anticipation of high future prices) and an offset effect (after a sale, people cut back as there are fewer unmet needs). For spending, we find that the anticipation effect is most important: when people hit the deductible, they spend a great deal more, but this is not offset by particularly low spending in the first months of the next coverage year. But for episodes, we find both effects are important. We find no evidence, however, that the acceleration of deferrable care leads to fewer acute problems in the future. These results suggest that most of the intertemporal substitution therefore reflects retiming of care, rather than stocking up on general health capital.

The divergence between the short and long run price sensitivities imply that different sources of price variation yield different price sensitivities. For example, (quasi-)random assignment to insurance plans generates long lasting price differences, but hitting the deductible generates much shorter price change. To illustrate this, we estimate a series of static models that ignore intertemporal substitution and differ in their identifying variation (all of which, we emphasize, is experimentally induced). Pooling both short- and long-lasting price variation yields a price sensitivity about 8% larger than the long-run sensitivity. Using only short-run price variation yields price sensitivities that are nearly 100% larger than the long-run sensitivity, while including only long-run price variation yields an estimate nearly equal to the long-run price sensitivity. These results reconcile some of the disparate elasticity estimates in the literature. For example, Eichner (1998) and Kowalski (2016) use family members' accidents to instrument for other family members' spending, since if anyone in the family has an accident, everyone is more likely to hit the deductible. These studies find elasticities on the order of -0.5 to -1.5, much larger than the elasticity of -0.2 found in the RAND Health Insurance Experiment (Manning et al., 1987), but closer to the elasticities we find when we use only within-person, short-run variation.

Overall, we find an important role for intertemporal substitution in health care demand. Many studies also find that demand responds to anticipated future policy changes (e.g. Alpert (2016), Brot-Goldberg et al. (2015), Schmid (2016)). We show in an experimental context that deductibles, not just coverage maxima, generate intertemporal substitution, and that anticipatory responses to policy-induced price changes are not an artifact of the policy change, but a common part of the response to nonlinear cost-sharing. This conclusion differs from the original RAND investigators' (Keeler et al., 1982; Keeler and Rolph, 1988), who argued that sales effects and anticipatory spending were unimportant in the Health Insurance Experiment. We offer a reconciliation in Section 6.

2 Modelling health care demand

2.1 Health care demand without intertemporal substitution

We derive testable implications of the hypothesis that health care consumption in one period does not affect the marginal utility of health care consumption in other periods, i.e. that there is no intertemporal substitution in health care demand. This hypothesis implies that health care demand exhibits only pecuniary dynamics: spending in one period may affect the price of health care (or available income) in other periods, but not its marginal utility. The main testable implication will be that people who face the same price of health care this year, but different prices next year, have the same demand.

While this implication is straightforward to derive, the model addresses two conceptual difficulties that arise in the context of health insurance and other nonlinear pricing arrangements. First, it is not obvious what the relevant price is, since patients may not understand the link between current spending and future prices (and so they may overweight the current price). Second, we want to compare spending of people who hit the maximum out-of-pocket limit in a nonlinear insurance plan to people who have free care. This comparison is not clean, though, because high spenders are mechanically more likely to hit the out-of-pocket limit. The model implies that we can avoid the first problem by looking at demand in the final period of the coverage year, when there is little uncertainty, and spending has no further effect on future prices. The model also generates observable conditions under which we can test for no intertemporal substitution by comparing everyone in cost-sharing (whether they have hit the out-of-pocket limit or not) to everyone in free care, therefore getting around the second problem.

Model set-up Patients demand health care h_t each month to maximize utility. Monthly utility depends on h_t, other consumption c_t, and a preference shock ε_t, representing shifts in the marginal utility of health care, for example an illness which necessitates health care spending. We write

$$U_t = u(h_t, c_t, \varepsilon_t)$$

The key assumption that this utility function embodies is that past health care consumption has no direct effect on utility; it affects choices only through the possibly nonlinear budget set. This assumption is commonly and implicitly made in the literature. We further assume that u is convex, although we do not assume that $u' > 0$ for all h (since we observe finite spending at a price of zero). To keep the model simple, we assume that the utility function is quasilinear, so $U_t = u(h_t, \varepsilon_t) + c_t$. This lets us ignore all savings decisions, so that the only source of dynamics is the link between current health spending and future prices. We explain below that allowing for income effects would strengthen our test.

We assume that people face a piecewise linear annual budget set, with the slope and kink points determined by the health insurance contract. Let $C(h)$ give the out-of-pocket cost of h dollars of health care spending. To preview the empirical application, we will assume in particular that either health care is free, or that people face a coinsurance rate of $coins$ up to an out-of-pocket maximum of MDE, so the budget set is piecewise linear with a slope of $coins$ when $h < MDE/coins$ and a slope of 0 above it. People begin the year with income Y. Figure 1 illustrates this budget set.

Demand In the simplest case of forward looking behavior, no uncertainty about ε, and no income effects, demand is easy to characterize. Patients simply choose h_t in each period so that $u'(h_t, \varepsilon_t) = C'(\sum_{\tau=1}^{12} h_\tau) \equiv p$. Nonlinear prices mean there may be multiple solutions to this first order condition, but at any interior solution, the marginal utility of health care spending equals the end-of-year price p, regardless of the overall shape of $C(\cdot)$.

However, the possibility of uncertainty and myopia complicates the analysis, since in either case, patients cannot set their monthly marginal utility equal to p. We introduce uncertainty and myopia with the following annual decision problem:

$$U = \max_{h_t(\cdot)} \sum_{t=1}^{12} \beta^t E[u(h_t, c_t, \varepsilon_t) | \mathcal{I}_t]$$

$$\text{such that } Y = C\left(\sum_t h_t\right) + \sum_t c_t$$

Uncertainty arises because patients only have limited information about future ε_t, represented by an evolving information set $\mathcal{I}_t$. We allow for forward looking behavior when $\beta > 0$, and an extreme form of myopia with $\beta = 0$.

We solve this dynamic optimization problem by backwards induction. For notational simplicity we assume that ε follows a first-order Markov process, so the state variables are accumulated health spending $H_t \equiv \sum_{\tau=1}^{t-1} h_t$ at the beginning of month t, and the preference shock ε_t. Let the function $T(h, H)$ give the required (marginal) out-of-pocket payment for health care spending h when total spending so far is H, and let T be the last period of the coverage year.[2] The Bellman equations are

$$\begin{aligned} V_t(H_t, \varepsilon_t) &= \max_h u(h, \varepsilon_t) - T(h, H_t) + E[V_{t+1}(H_t + h)|\varepsilon_t, H_t, h]. \\ V_T(H_T, \varepsilon_T) &= \max_h u(h, \varepsilon_T) - T(h, H_T) \end{aligned}$$

We denote by $h_t(H_t, \varepsilon_t, T(\cdot))$ the period- and contract-specific policy function that is the solution to this Bellman equation.

This solution depends on β and on the joint distribution of ε across periods. However, inspection of the period T Bellman equation reveals that in the final period, neither discounting nor uncertainty affects demand. This is because, in the final month of the coverage year, there are no meaningful dynamics—current spending does not affect future prices—and all uncertainty is revealed. We therefore focus on demand in the final period of the coverage year.

Testable implications The first order condition for final period health care spending h_T for someone in a cost-sharing plan with accumulated health expenditures H_T is

$$u'(h_T, \varepsilon_T) = T'(h_T, H_T).$$

At an interior solution at the end of the coverage year, people choose health spending so that the marginal out-of-pocket price, $T'(h, H_T)$, equals the marginal utility of health. Corner solutions with $h_T = 0$ are empirically common; these happen when ε_T is such that $u'(0, \varepsilon_T) < T'(0, H_t)$.[3]

Letting $p = T'(h_T, H_T)$ denote the realized end-of-year price, we can therefore write final

[2] $T(h, H) = C(h + H) - C(H)$.

[3] The first order condition also does not hold if an individual chooses consumption to end up exactly at the kink point. However this point is never optimal because the price is decreasing from one line segment to the next.

period demand as

$$h_T = h_T(H_T, \varepsilon_T, T(\cdot)) = h(p, \varepsilon_T). \tag{1}$$

Equation 1 says that two people who have the same end-of-year price will have the same final-period demand, regardless of whether they face the same contract $T(\cdot)$. In particular, a person in a nonlinear cost-sharing plan who hits the maximum dollar expenditure will have the same final period demand as a person who was in free care all along.

Note that by construction, a person who hits the MDE has less available income than a person in free care (exactly the MDE less). If income effects are important, then this means that spending should be *higher* in free care than for people who hit the MDE, as health is a normal good. Precautionary savings motives generate a similar prediction. Allowing for income effects thus strengthens our test, in the sense that they would make it harder to detect higher spending among people who hit the MDE than among people in free care.

We test this implication against an alternative hypothesis that future prices also matter for demand. To do so, we would like to compare patients in free care to patients who have hit the MDE in a cost-sharing plan. These patients face the same current prices, but different future prices: patients in free care will continue to face a price of zero, but patients in the cost-sharing plan will not, since not all patients who hit the MDE in one year will hit it in the next.

We cannot directly test this implication, however. To see this, consider expected demand in free care and in cost-sharing plans among people who hit the MDE:

$$E[h_T|\text{free}] = E[h(0, \varepsilon_T)|\text{ free}]$$
$$E[h_T|\text{cost-sharing, hit}] = E[h(0, \varepsilon_T)|\text{ cost-sharing, hit }]$$

Whether a patient hits the MDE depends on her past and current spending decisions, which depend on past and current realizations of ε_t. As a result, $E[h(0, \varepsilon_T)|\text{cost-sharing, hit}] \neq E[h(\varepsilon_T, 0)|\text{free}]$, even with random assignment of plans and no intertemporal substitution. Conditioning on realized end-of-year prices creates an endogeneity problem.

We avoid this problem by looking at overall expected demand in cost-sharing plans, averaged over people who do and do not hit the MDE. It is helpful to segregate people who, based on their entire history of ε, would or would not hit the MDE. To be precise, define

$$\varepsilon^* = \left\{ (\varepsilon_1, \ldots, \varepsilon_T) : \sum_{t=\tau}^{T} h_\tau(H_\tau, \varepsilon_\tau, T(\cdot)) \geq MDE/coins \right\}.$$

This is the set of ε leading a person who faces an out-of-pocket cost function $T(\cdot)$ to hit the MDE. Expected spending in cost-sharing can be decomposed into the probability weighted average of spending among people who do and do not hit the MDE:

$$E[h_T|CS] = Pr(\varepsilon \in \varepsilon^*)E[h(0,\varepsilon)|\varepsilon \in \varepsilon^*] + (1 - Pr(\varepsilon \in \varepsilon^*))E[h(p,\varepsilon)|\varepsilon \notin \varepsilon^*]. \tag{2}$$

These expectations differ because people who hit the MDE face a different price and a different distribution of ε_T. We may perform the same decomposition in free care, continuing to split the ε by whether people would have hit the MDE in the cost-sharing plan:

$$E[h_T|free] = Pr(\varepsilon \in \varepsilon^*)E[h(0,\varepsilon)|\varepsilon \in \varepsilon^*] + (1 - Pr(\varepsilon \in \varepsilon^*))E[h(0,\varepsilon)|\varepsilon \notin \varepsilon^*]. \tag{3}$$

Comparing Equation 2 and Equation 3 shows that expected spending month T spending is the same among people who would hit the MDE if they were in the cost-sharing plan, regardless of which plan they are actually in. Thus the difference in expected spending in month T between the two plans is

$$(1 - Pr(\varepsilon \in \varepsilon^*))E[h(p,\varepsilon) - h(0,\varepsilon)|\varepsilon \notin \varepsilon^*].$$

This difference is negative as long as two conditions hold: some people do not hit the MDE (so $1 - Pr(\varepsilon \in \varepsilon^*) > 0$) and demand is downward sloping on average for the people who do not hit the MDE, so that the expectation is strictly positive. A sufficient condition for downward sloping demand is that some people who do not hit the MDE nonetheless have positive demand.[4] Thus our main empirical test of no intertemporal substitution is a comparison of demand in free care and and in cost-sharing. In the absence of intertemporal

[4] For these people, the first order condition holds, and concavity of u implies that demand is downward sloping at any interior solution.

substitution, demand is lower in cost-sharing than in free care.

2.2 Alternative hypotheses

The model so far assumes that health care spending in one period has no effect on the marginal utility of health care in future periods. This is a standard assumption, but a strong one, and here we sketch two alternative ways of modelling health care demand that create a link between current health care utilization and future demand.

First is the health capital model, originally developed by Grossman (1972), and econometrically implemented by Gilleskie (1998); Blau and Gilleskie (2008); Khwaja (2010) and Cronin (2016). Under this model, people derive utility from a stock of health H. They may augment this stock by health care utilization, such as visiting a doctor, or by health behaviors, such as better diet or exercise. The stock of health depreciates slowly, so that health care utilization in one period leads to better health in the future. As long as health is durable, health care spending can be shifted from one period to another while keeping health unaffected, and so an anticipated price increase tomorrow may generate a spending response today, and a large spending decline when the price change materializes.

An alternative view is that, even in absence of durable health, some procedures are easy to retime. For example, many tests such as colonoscopies or even annual check-ups can be shifted forward or backwards by a few months with little loss of effectiveness. Patients who anticipate a future price increase may therefore try to move forward such procedures to take care of them when the price is low. As with durable health, if some health care needs are deferable, then spending will rise before an anticipated price change, and decline after it materializes, holding fixed the current price. These models imply that we should see the biggest response to future prices in two kinds of care: easily deferrable care, and care that produces long lasting benefits. On the other hand, we expect not to find an effect of future prices on the demand acute care, which typically does not produce long-lasting benefits, and by definition cannot be easily deferred.

3 Background and data

3.1 Experimental design and randomization

The RAND Health Insurance Experiment, run from November, 1974 to February, 1982, was a randomized field experiment to measure whether more generous health insurance caused higher health care spending.[5] The experiment ran at six different sites, chosen to be broadly representative of the United States, and new families were enrolled over several start dates. Families were selected at random in the site, but the investigators oversampled low income families, and excluded very high income families, so the sample is not representative, nationally or within the sites.[6] At a given site and start date, families were randomly assigned to one of several health insurance plans according to a finite selection model (Morris, 1979), which explicitly balanced a subset of observable characteristics across plans. The plans all covered inpatient and outpatient health care, as well as vision, prescription drugs, medical supplies, and mental health and dental health. Families were also randomly assigned to an enrollment term: three years for 70% of enrollees, and five years for the remainder. In all analyses, we pool the three and five year enrollees, to maximize power.

The plans primarily differed in their coinsurance rates. In the most generous plan, "free care," families faced a coinsurance rate of zero on all services. Three other plans had coinsurance rates of 25%, 50%, and 95%. Figure 1 illustrates the budget set created by these cost-sharing plans. A fifth plan, the "mixed" plan, had 25% coinsurance for medical services and 50% for mental and dental. Patients in these plans were only responsible for cost-sharing up to a maximum dollar expenditure (MDE), which was randomly set to 5, 10, or 15% of family income, but capped at $750 or $1,000.[7] Because the 95% coinsurance plan resembles

[5] Newhouse and The Insurance Experiment Group (1993) provides a detailed overview of the experimental design results of the experiment. Aron-Dine et al. (2013) offer a helpful summary for modern audiences. As Newhouse et al. relate, the initial motivation for the experiment was the widespread presumption in 1970 that national health insurance was imminent, and the only question was how much cost-sharing it should have.

[6] The sample also excludes people aged 62 and older at enrollment, who would eventually obtain insurance through Medicare, as well as some disabled people, institutionalized people, and military families.

[7] This is in nominal dollars. Cost-sharing rules in the HIE were not inflation adjusted over the experiment; $1,000 in 1974 works out to about $4,600 in 2011, and $1,000 in 1982 works out to about $2,300. Note that, because the MDE was tied to family income, it varied from year to year, and families with zero income received de facto free care.

a straight deductible up to a stoploss, it is often called the "family deductible" plan. A final plan, "individual deductible," had a 95% coinsurance rate for outpatient care, but inpatient care was free. In this plan, each individual had an out-of-pocket maximum of $150, but family out-of-pocket spending was capped at $450.[8] In some analyses, to maximize power, we pool all cost-sharing plans together.

Because of their nonlinear cost sharing features, the RAND plans anticipated the design of modern health insurance plans. The family deductible plan, in particular, resembles modern high-deductible health plans, since it has a coinsurance rate of nearly 100% below the MDE, and an MDE that can be as much as 15% of family income. By comparison, 76% of plans on the Health Exchanges in 2014 were classified as "high deductible," meaning their deductible exceeded $1,250. The median silver plan in 2014 had a deductible of $2,500 and a maximum out-of-pocket expenditure of $6,300 (Coe, 2014), or about 12% of median household income (DeNavas-Walt and Proctor, 2015).

3.2 Data and summary statistics

We use the replication files which the original RAND investigators have made publicly available.[9] Our goal is to analyze the effect of free care relative to cost-sharing on health care demand in the final month of the coverage year, so we aggregate spending and utilization from the claims files to the person-month level, and inflate spending to 2011 prices using the monthly CPI-U.[10] In addition to the claims data, we use the demographic file for patient demographic and background information; the eligibility file to record coverage and family structure (to link patients within an insured family); and the episode of care file, to count episodes and to find the date when a patient "hits the MDE," i.e. when her or her family's out-of-pocket spending for the coverage year exceeds the maximum dollar expenditure. We

[8] A separate arm involved random assignment to an HMO, which we do not analyze here.

[9] The files may be downloaded from `http://www.icpsr.umich.edu/icpsrweb/NACDA/studies/06439`.

[10] Each claim in fact has multiple dates, including the date of service and date filed. We date outpatient claims by the date of service, and we date all inpatient claims by the date of the admission. We believe this is consistent with the cost-sharing rules for the experiment, where hospitalizations that span multiple coverage years appear to count towards the coverage year in which they began. For most patients, coverage began on the first of the month, so calendar months and coverage months align. But for some patients, coverage began on the 31st. In these cases, we align calendar and coverage month by shifting all dates forward one day.

use this information to define the end-of-year price as the coinsurance rate for patients who did not hit the MDE that year, and zero otherwise. We define the monthly spot price analogously: it is equal to the coinsurance rate for patients who have not hit the MDE by the beginning of the month, and zero otherwise.

We augment the claims data, which measure spending, with data on episodes of treatment, which measure quantities. Episodes of treatment are groupings of claims reflecting all spending for a particular treatment. Much of the original HIE analysis focused on episodes of treatment (e.g. Keeler and Rolph (1988)). Episodes likely reflect patient decisions rather than physician input, because the decision to seek treatment is likely-patient driven. The episode data are classified into acute, chronic, and well-care, which consists of relatively deferrable procedures, such as examinations or vaccinations. Importantly, the providers themselves make this classification decision, and it reflects the deferrability of treatment, not any information about spending or timing (Keeler et al., 1982). For each month, we record the number of episodes of each type that took place during that month.[11]

We make four restrictions to create our analysis sample. First, following the original RAND investigators, we only include whole coverage years.[12] Second, we exclude all Dayton families from year 1; during this period, dental care was only covered in the free care plan, so we do not see dental spending for the cost-sharing group. Third, we exclude a handful of families with missing information on whether they have reached the MDE. Fourth, we drop the 50% coinsurance plan from our analysis, because Aron-Dine et al. (2013) show that the randomization appears to have failed for this plan. After these exclusions, our analysis sample consists of 4,591 people in 1,820 families, across 214,320 person-months.

Table 1 provides summary statistics by plan type. The first column shows the average of the indicated variable in the free care plan. The remaining columns show the difference in means in each cost-sharing plan, relative to free care. Because plan assignment was random

[11] We omit drug episodes (representing 1.6% of spending) from our analysis, because we cannot reliably date them. Prescriptions which span multiple years are defined as separate episodes for each year. For episodes continuing into a new year, the start date is imputed as the first day of the coverage year. Including these episodes creates a false impression of a surge in new health care on the very first day of the year.

[12] Specifically, this means we drop the first (partial) year of newborns and adopted children, the only late entrants; the final (partial) coverage year for people who attrit; any partial years from suspensions; and all post-death years for people who die. Following the original investigators, we include the final partial year for people who die, and treat post-death spending and utilization in that year as zero.

only for a given site and start date, we follow Aron-Dine et al. (2013) and report means adjusted for site by start date fixed effects. Standard errors, clustered on family, are in parentheses.

The cost sharing plans, unsurprisingly, are less generous than the free care plans. The MDE is about \$1,500 in the individual deductible plan and \$2,400-\$3,000 in the other plans.[13] Between 18 and 46% of people in the cost-sharing plans hit the MDE in a given year. People in the less generous plans are more likely to hit the MDE (because a given amount of spending makes it easier to do so), but end-of-year prices and are increasing with the sticker coinsurance rate of the plan.

These differences in plan generosity translate into differences in spending and episodes of treatment. Total spending (not shown) is about \$180 per person per month in free care, and splits roughly evenly into inpatient, outpatient medical, and outpatient dental. There is almost no spending on mental health care, so we do not analyze it further. Spending is lower in the cost-sharing plans across all categories, and overall by \$30-\$70, or about a third. The difference is largest in dental care and outpatient medical care, and for the least generous plan, the family deductible plan. Episodes show a similar pattern: in free care the average patient has 0.66 total episodes per month, and in cost-sharing patients have about a third fewer. The table also records the fraction of people assigned to each plan who attrited, defined as failing to complete the experiment as scheduled. Attrition is most common in the family plan.

3.3 Balance and validity of randomization

To interpret these differences as the causal effect of plan assignment, we require that insurance plan assignment be uncorrelated with patient health care spending propensity. This assumption can fail either if the randomization itself was unsuccessful, or if differential attrition leads to selection of healthier people in less generous plans. We test for experimental validity in Appendix A; we find that the plans appear balanced on pre-experimental characteristics, including both direct measures of utilization and demographic variables.

[13] The MDE and other cost-sharing parameters are not exactly zero in free care because the reported amount is net of the estimated (and slightly non-zero) site-by-start date fixed effects.

Our tests follow Aron-Dine et al. (2013) closely, but Aron-Dine et al. come to different conclusions about balance after enrollment: they find that the plans appear highly unbalanced, on utilization or other variables, among people who complete the experiment. In Appendix A, we reconcile these findings. The key difference is that Aron-Dine et al. (2013)'s analysis of nonbalance at experiment completion includes the suspect 50% coinsurance plan. As Aron-Dine et al. (2013) show, this plan appears unbalanced even at randomization (i.e. before attrition), and we exclude it from the analysis. When we exclude it, we pass the balance tests, but when we include it, we fail them. Nonetheless, given the differential attrition and refusal rates, we remain cautious about the randomization. Our main concern is differential attrition, since our interest is in the changing time pattern of treatment effects—whether they are growing or shrinking over the coverage year, rather than their absolute level. In robustness tests, we attempt to address differential attrition by controlling for all predetermined variables, interacted flexibly with time dummy variables. These extra controls have no effect on the results. We also find similar results when we restrict attention to a balanced sample, where changing composition cannot explain changing treatment effects.

4 Treatment effects over the coverage year

4.1 Estimating equations

The model in Section 2 shows that we can test for intertemporal substitution by comparing health care demand in cost-sharing and in free care plans in the final month of the coverage year. To do so, we estimate monthly demand in free care and in cost-sharing, adjusting for differences in site-by-start date and for general trends. Specifically, we estimate regressions of the following form:

$$y_{it} = \sum_{year=\{\text{first, middle, last}\}} \sum_{\tau=1}^{\tau=12} (\beta_{\tau}^{year} Free_{i\tau,year} + \gamma_{\tau}^{year} Cost_{i\tau,y}) + \mu_t + \theta_{sem} + \varepsilon_{it}, \tag{4}$$

for several outcomes y of person i in month t. There are three time indices in the regression: t indexes calendar time (e.g. January, 1980), τ indexes coverage months (1-12), and *year* refers to different coverage years (first, middle years, and final).[14] In all specifications, we control for a full set of demeaned calendar time dummies, as well as demeaned site-by-enrollment-month dummies. Our interest is in the coefficients β_τ^{year} and γ_τ^{year}, which measure the average of y_{it} in coverage month τ of a given year, for beneficiaries in free care and in cost-sharing plans, after adjusting for trends and site-by-start date differences. For example, γ_{12}^1 gives the average of y in the cost-sharing plans in the last month of the first coverage year. Because plan assignment is random conditional on site and enrollment month, $\beta_\tau^{year} - \gamma_\tau^{year}$ gives the effect of being in free care, relative to cost-sharing, in relative month τ and year $year$. Our primary interest is in $\beta_{12}^{year} - \gamma_{12}^{year}$, but examining the full set of monthly spending amounts will be informative.

In estimating Equation 4, we pool people assigned to three and five year terms. To maintain power, we pool years 2, 3, and 4 for the five-year enrollees with year 2 of the three-year enrollees; this treats all "middle" years the same. That is, β_1^{middle} is average spending in free care in people in year 2 of a three year term and years 2-4 of a five year term. Likewise we treat year 5 of the five-year enrollees the same as year 3 of the three-year enrollees. Pooling this way lets us highlight beginning-of-experiment and end-of-experiment effects. We have found similar patterns, albeit noisier, when we examine the three and five year enrollees separately.

4.2 Results

Prices Our test is valid only if end-of-year prices are in fact different between the cost-sharing and free care plans. We begin by verifying this.[15] Figure 2 shows the fraction of people who have hit the MDE by the start of each coverage month, as well as the corresponding average spot price. The figure simply plots the estimated values of β and γ from Equation 4. In the free care plan, prices are always zero. In the cost sharing plan, spot prices decline as more and more people hit the MDE. People in cost-sharing plans hit the MDE steadily throughout the coverage year, and as a result prices fall steadily. However, by

[14] Calendar months and coverage months are not collinear because the experiment had staggered start dates, with every calendar month a possible start month.

[15] We also require that some people who do not hit the MDE nonetheless have positive demand in the last month of the coverage year. Average spending in this group is $51 (standard error: $3.8).

the start of the last month of the coverage year, only about a third of people have hit the MDE, so even in the last month of the year, the average spot price is about 0.45. Thus in the absence of intertemporal substitution, we should expect spending and episodes to diverge between the two groups in the final months of the coverage year.[16]

Spending Figure 3 shows the results for spending. We show total spending in Panel A. Because this is noisy, we remove inpatient spending and plot total outpatient spending in Panel B, and then we decompose outpatient spending into medical and dental spending in Panels C and Panel D. Several striking patterns emerge from the figure. Total spending does not show much of a pattern in free care, except for a dramatic surge at the end of the experiment, when it shoots up from its average of around $175 per month, to over $300. The time series of spending looks quite different in the cost-sharing plans. There, spending is flat in the first half of the coverage year, at around $125 per person, and then rises sharply in the second half of the year. By the end of the coverage year, spending in the two groups has converged, and in the final month of the middle year, spending is actually higher in the cost-sharing plan than in free care, in sharp contrast to the predictions of the model without intertemporal substitution.[17]

Total outpatient spending shows a similar pattern, but much more clearly because there is less month-to-month volatility; spending surges in cost-sharing at the end of the coverage year, and eventually rises above spending in free care. Decomposing outpatient spending into medical and dental care shows an interesting pattern in free care: in the first quarter of the experiment, patients spend about $70 per month on outpatient medical care, but over the next few months spending falls, and remains roughly constant until the last month in the experiment. Dental spending in free care is also high at the beginning of the experiment, throughout the first year and especially after the first few months. In the cost-sharing plans, outpatient medical care and dental care both show a pattern of rising spending at the end of the coverage year, although it is more pronounced for dental care than for medical.

[16] The figure does not show the actual end-of-year price, only the beginning-of-month price in the last coverage month. However, on average 35% of people in the cost sharing plans hit the MDE, and the average end-of-year price is 0.43.

[17] To keep the figures readable, we have omitted confidence intervals. However Appendix Figure A.1 and Appendix Figure A.2 plot the monthly treatment effects, $\beta_\tau^{Year} - \gamma_\tau^{Year}$, along with the 95% confidence interval for the difference, so that readers may more easily gauge the magnitude and statistical significance of the effect size over the coverage year.

Episodes of treatment We extend our spending analysis to episodes of treatment in Figure 4. We focus on non-dental well-care, dental care, and acute episodes because these are straightforward to date.[18] Panel A shows the number of well-care episodes by plan and month. Well-care episodes in free care exhibit a clear U-shaped pattern over the experiment: high at the beginning and end, but flat in the middle. In the cost-sharing plans, they rise over the coverage year, and by the end of the first and middle coverage years, the difference between cost-sharing and free care is small and statistically insignificant. Dental care shows a nearly identical pattern, in both free care and in the cost-sharing plans. The high utilization at the end of the coverage year suggests that intertemporal substitution is an important part of demand for these categories of care, which are medically straightforward to retime.

By contrast, acute episodes show a very different pattern, visible in Panel C. The number of acute episodes rises steadily throughout the coverage year, especially in cost-sharing, but without any obvious jump in the last few periods. Instead there is a steady rise in the number of acute episodes in the cost-sharing plans in the second half of the coverage year. But there are always substantially (and significantly) more acute episodes in the free care plan than in the cost-sharing plans. Non-defferrable care shows no evidence for intertemporal substitution. This is an important specification test: it shows that our results are not driven by trends in utilization over the coverage year (that are specific to the cost-sharing plan), nor by providers shifting when they date and file claims.

Summary We summarize three important facts from these figures. First, people hit the MDE smoothly throughout the year, but only about a third ever hit it, so end-of-year prices are much higher in the cost sharing plans than in free care. Second, spending and utilization are typically lower in cost-sharing plans than in the free care plan. Third, spending and utilization exhibit different patterns within a coverage year. In free care, they are roughly flat over the coverage year, but in cost-sharing plans, they rise in the last 1-3 months. By the end of the year spending in the two plans is roughly equal. This pattern is especially prominent for deferrable episodes, and for outpatient spending. Overall the results provide

[18] The remaining categories are non-acute chronic episodes, which are persistent and do not have a clear beginning or end, and hospital episodes. These are difficult to date because they often include maternity care, and the RAND investigators dated most maternal care at the beginning of the coverage year, when it might be anticipated.

strong evidence against the null hypothesis of no intertemporal substitution. Instead we see that demand rises ahead of future price increases—at the end of the coverage year in cost-sharing, and at the end of the experiment, in free care. This rise is most pronounced for deferrable care, and least for acute care.

5 Estimating long- and short-run price sensitivities

5.1 Estimating short and long run price sensitivities

The results so far show that intertemporal substitution is an important part of patients' response to nonlinear cost sharing rules. If so, a long-lasting price change might have a different effect than a short lasting one. To quantify these short and long-run price effects, we estimate models of the form

$$y_{it} = \alpha + \beta_0 price_{it} + \beta_{-1} price_{it-1} + \beta_1 price_{it+1} + X_{it}\theta + \varepsilon_{it}. \tag{5}$$

The outcome of person i in month t, y_{it} depends not only on the price in month t, but also on the lagged price $price_{it-1}$ and the expected lead price $price_{it+1}$. Additional controls always include fixed effects for site-by-start-date, and for date.[19]

This specification separates short and long-run price effects in a simple way. β_0 measures the immediate effect of a one-time, unanticipated price change; β_{-1} measures any offsets in the next period from high or low spending in a given period; and β_1 measures how much demand rises or falls in anticipation of an expected future price change. A long-run price change affects current, past, and future prices, so the per-period effect of a long-run price change is $\beta_0 + \beta_{-1} + \beta_1$. To estimate Equation 5 and recover the short and long-run price sensitivities requires that we address several challenges.

Measuring prices We measure $price_{it}, price_{it-1}$ and $price_{it+1}$ with the category-specific current, lagged, and lead spot prices. Relating demand to spot prices is consistent with the

[19] It is possible to microfound this specification using a health capital model. If the demand for health care derives from the desire to build up a stock of health capital, then the level of health capital likely depends on past and future prices. In principle the entire history of prices matters, not just a single lag. However we show the results are robust to using a larger number of leads and lags.

evidence in Section 4; the rising spending over the coverage year suggests that spot prices are relevant for patients.[20] A further issue is that different categories of care—outpatient medical, dental, and inpatient—have different coinsurance rates, depending on the assigned plan. We address this issue by estimating category-specific regressions where the price in each regression is the relevant one for that category.[21] By aggregating category-specific coefficients across all categories of care, we can obtain an estimate of the price sensitivity of overall demand. That is, letting β_τ^y represent the effect of $price_{i,t+\tau}$ (for $\tau = -1, 0, 1$) on y, we obtain the total spending response as $\beta_\tau^{total} = \beta_\tau^{OutMedical} + \beta_\tau^{OutDental} + \beta_\tau^{Inpatient}$. We obtain the total episode response analogously.

Price endogeneity Measuring prices using realized spot prices raises a new challenge: such prices are mechanical functions of lagged spending, given the insurance plan. If there is any autocorrelation in ε_{it}, then prices will be correlated with the error term and OLS estimates will be biased. We solve this problem by instrumenting for price, using interactions of plan assignment with coverage date and coverage year. We have 144 binary instruments, created from the complete interaction among months (1-12) coverage year (first, middle, last), and plan (25% coinsurance, mixed, family deductible, or individual deductible). These instruments identify price sensitivities by relating plan- and month-specific treatment effects over the entire course of the experiment to month-specific differences in prices. These instruments are valid as long as the only reason that demand differs across the treatment arms in a given coverage month and year is that $price_{it-1}, price_{it}$ and $price_{it+1}$ differ.

Imputing prices in the first and last period A final challenge is that for all patients, $price_{it-1}$ is missing in month 1, and $price_{it+1}$ is missing in the final coverage month. We do not want to exclude these months, however, because they provide valuable information about intertemporal substitution. Instead, we impute prices outside the experiment period as a constant $\bar{p}$, and we include dummy variables indicating imputed values of $price_{it-1}$ and $price_{it+1}$.[22]

[20] We remain agnostic whether this is because of partial myopia or uncertainty about the probability of hitting the MDE, however.

[21] We assume that there are no cross-category price effects. We have attempted to estimate such effects but lack the power to do so precisely.

[22] Note that the estimates are numerically invariant to the choice of $\bar{p}$ because the instruments induce no variation in it. Our estimates are robust, however, to simply dropping the first and last month.

If the random assignment is valid, this imputation is innocuous for the pre-period prices, since randomization implies that on average all pre-determined variables are equal across treatment arms.[23] Randomization alone does not justify the post-period imputation, however. Instead we must assume that plan assignment did not cause people to change their insurance plan generosity immediately after the experiment ended. We make this stronger assumption so that we can use the spike in spending in the very last month of the experiment to help identify β_{t+1}. Despite these imputations, $price_{it-1}$ and $price_{it+1}$ are still missing for a handful of observations: the first month that newborns enter the insurance experiment, the first month after temporarily suspended participants return, and the last month before a temporary suspension. We drop these observations in estimation.

5.2 Identification and first stage

Identification requires that the instruments induce linearly independent variation among spot prices, future prices, and lagged prices. As we report below, our first stage F-statistics are on the order of 500-1000 in all specifications, suggesting that our instruments are strong indeed. The linearly independent variation in prices comes from turning over of the coverage year, as can be seen in Figure 2. Over the coverage year, current prices and future prices move together until month 12, when they diverge sharply. Thus we identify the differential response to $price_{it}$ and $price_{it+1}$ from the spike in spending at the end of the year.

Likewise, $price_{it-1}$ and $price_{it}$ move together except in the first month of the coverage year and in the month of hitting the MDE. β_{-1} is therefore identified by three distinct sources of variation: high spending in month 1 year 1 as people enter free care; any increase in spending in the cost sharing plans from month 1 to month 2, as lagged price changes from lower to higher; and any particularly large surge in spending in the month when people are particularly likely to hit the MDE.

It might be surprising that we can use the first and last month of the experiment to help identify lead and lag price sensitivities even though we do not observe lead and lag prices in

[23] One concern is that if people had nonlinear cost-sharing plans before the experiment, and people assigned to free care cut back on care before enrolling, then in fact month 1 lagged prices differs by treatment status.

these periods. To understand how they contribute to identification, consider treatment effect in year 1 month 1 relative to year 2 month 1. In year 1, randomization induces variation in $price_{it}$ and $price_{it+1}$ but not in $price_{it-1}$. In year 2, however, $price_{it-1}$ differs as well. Thus the difference in treatment effects between year 1 month 1 and year 2 month 1 helps identify β_{-1}. Similar logic shows how the differential treatment effect at the end of the final year helps identify β_1.

5.3 Short and long run price sensitivities

Spending response Columns (1)-(3) of Table 2 show the estimates of Equation 5. A one-time, unanticipated price increase of 0.1—an increase in the coinsurance rate of 10 percentage points—reduces current spending on outpatient medical care by \$5, outpatient dental by \$12, and inpatient care by -\$0.2, for an overall effect of \$17 (standard error: \$5). Much of this spending represents low spending in anticipation of a future price decrease. The coefficient on the lead of price indicates that half of the increase in outpatient medical, and three quarters of the increase in outpatient dental, is anticipatory. Indeed the long-run spending consequences of a price increase are less than half the short-run consequences for medical care, and about a fifth for dental care.

The overall-long response is \$7, smaller than the short-run effect by \$9.8 (standard error: \$5.7), or about 60% of the short-run effect. For outpatient medical spending, the long-run response is about half the short-run response, and for dental spending, the long-run response is only a sixth the short-run response. The inpatient estimates are too imprecise to compare the short and long run spending responses. For spending, long and short-run price responses diverge mainly because demand rises in anticipation of future price increases, not because of any effect of lagged prices on current demand. Lagged prices do not have a significant positive effect on current spending, and for dental care their effect seems to be negative. This suggests that the high anticipatory spending is not later offset by lower spending after the price change is realized.

Episode response Columns (4)-(8) show the price sensitivity of episodes of care, broken down by well-care (non-dental), dental care, acute, chronic, and inpatient. Overall, a one-time price increase reduces monthly episodes by 0.36 (standard error: 0.03), with most of the response coming from well-care, dental care, and acute care. A permanent price increase has a smaller effect, reducing episodes by 0.29 (standard error: 0.02). Well-care and dental care drive the divergence between the short- and long-lasting price effects. For these categories, the long-run effect is only 30-40% of the short-run effect. For acute, chronic, and inpatient episodes, the long-run and short-run responses are closer.

There is considerable anticipatory demand for well-care and for dental care, with lead price sensitivity of 0.042 and 0.085, showing that episodes rise in the period before an anticipated price increase. Interestingly, we also see that well-care episodes rises after a high price period, holding fixed the current and future price: the coefficient on lag price is 0.023, or about half the lead price coefficient, meaning that about half of the anticipatory demand is later offset by lower demand when the price change materializes, pointing to retiming of office visits, check ups, and screenings. Temporary price changes encourage people to reschedule deferrable care to minimize out of pocket costs. We do not see, however, that this extra utilization ahead of a price increase has any effect on acute episodes, which respond in a negative, small and insignificant way to past prices. Although people get more deferrable care when they hit the MDE, and they do seek treatment for more acute and chronic episodes, this extra care does not translate into fewer future acute or chronic episodes, at least over the time horizon that we are able to consider.[24]

Summary For both spending and episodes of care, we have seen that the overall response to a short-lasting price change is lower than the the response to a long-lasting price change. Overall, in both cases, the short-run response is about twice as large as the long-run response. This divergence is largest for dental spending and for well-care episodes, and nearly zero for inpatient spending and acute medical care. It appears that hitting the MDE lets households retime their deferable care to reduce their out-of-pocket spending. This extra care does not translate into fewer acute episodes or less spending in future periods, however.

[24] We show 1-month effects here, but we have found similar results with up to six months of leads and lags. We lack the power to include many more leads and lags than this.

5.4 Robustness tests

We show in Table 3 the robustness of our results to alternative specification choices. The main threat to identification is the possibility of differential attrition among the different plans. Although the balance tests indicated that differential attrition is not a problem on average over the entire experiment, it is possible that changing sample composition leads to changing spending, and so the time-varying effects of cost-sharing might be explained by differential attrition. We address this concern in two ways. First, in Panel A, we augment our main specification with a full set of interactions between coverage year dummy variables and the available pre-determined variables.[25] These controls adjust for any changes in spending resulting from differential attrition that is correlated with observed predetermined characteristics. The point estimates are largely unchanged; the overall spending response is slightly smaller, mainly driven by a lower inpatient spending response. The short-run effect remains much larger than the long-run effect.

As a second check that changing sample composition does not explain the results, we show in panel B the results of estimating Equation 5 when we limit the sample to people who participate in the experiment for their assigned enrollment term. This reduces the sample size by about 15,000 person-months, as we drop all people who attrit, who were ever suspended, and the newborns who entered after enrollment. These restrictions reduce the estimated long-run price sensitivities; for overall spending it falls to -58.3. Again this decline is mainly due to a fall in inpatient spending sensitivity.

We conclude from these robustness checks that changing sample composition is unlikely to account for the results. An alternative concern with our results is that we rely on arbitrary lead and lag specifications to identify long and short-run price responses. As a robustness check, we show in Panel C that none of the results are sensitive to the exact specification of how lag and lead prices enter demand. We do this by including three lags and leads of price instead of one; the results are similar when we include other numbers instead. The results are highly similar, as are results from another specification (not shown) where we included six

[25] These variables are listed in Appendix Table A.1. The predetermined variables are often missing, and in such cases we set their value to -1, and include a dummy variable indicating missing, also interacted with coverage year dummies.

leads and lags. Finally, in Panel D, we verify that our treatment of the first and last month of the experiment—when we must impute $price_{it-1}$ or $price_{it+1}$ does not substantially affect the results. When we drop these months, our point estimates remain largely unchanged.

5.5 Consequences of ignoring intertemporal substitution

We have shown that short and long-lasting price changes can generate substantially different spending responses. To gauge the economic importance of this difference, we assess how failing to account for intertemporal substitution could bias estimates of long-run price sensitivity. This parameter is of interest because it governs the long-run effect of changes in coinsurance rates.

Table 4 shows estimated price sensitivities obtained from models that neglect intertemporal substitution. In the first column, we present our baseline estimates of 70.7. In the next column, we re-estimate Equation 5, but omit p_{t-1} and p_{t+1}. These specifications are otherwise identical to our main estimates; they have use the sample controls, controls, and instruments. The price sensitivity is larger than our baseline by about 8%. This difference arises because we now mix long-lasting price variation (coming from across plans) and short-lasting variation (coming from predictable within-person changes in the spot price over the coverage year). We show in columns (3) and (4) the consequences of just using short-lasting or long-lasting price variation. In column (3) we add fixed effects to the analysis, and the price sensitivity doubles to -142. With the fixed effects, all the price variation is within-person, and so it is short-lasting; the estimate is close to the short-run estimates in Table 2. On the other hand, when we use only the long-lasting variation induced by plan assignment (obtained by using as instruments plan dummies rather than plan-month dummies), we get a price sensitivity that is much closer to the baseline. The bias from neglecting intertemporal substitution is most severe when most of the price variation is short-lasting. Finally we show in column (5) that when we aggregate the data to the annual level and use the end-of-year price as the independent variable (and plan assignment as the instrument), price sensitivity is about 40% higher. This specification recovers Marshallian demand exactly if people have perfect foresight about future health care needs over the coverage year and if there is no intertemporal substitution. It closely resembles those used in the literature on estimating price sensitivity of demand for health care, where it is common to model annual spending as a function of the end of year price (e.g. Cardon and Hendel (2001)). Thus, estimates of price sensitivity that neglect intertemporal substitution are sensitive to the type of variation used for identification. Long-lasting price variation yields a price sensitivity that is closer to the long-run sensitivity; high-frequency variation yields an estimate closer to the short-run sensitivity.

6 Reconciliation with original HIE findings

We have argued that intertemporal substitution is an important part of how patients respond to nonlinear cost-sharing, causing them to stock up on health care when it goes on "sale," with especially large anticipatory responses. Neglecting these dynamics can lead to biased estimates of the long-run effect of cost-sharing on utilization. The original RAND investigators, however, argued that intertemporal substitution was not an important part of the response to cost-sharing, and found little evidence for anticipatory effects. Here we reconcile these different conclusions.

The original investigators considered the possibility of an over-response to sales, and an early technical report, Keeler et al. (1982), used the first three years of data from the pilot site in Dayton. They looked for intertemporal substitution by "looking at the experience on the free plan at the start and end of the experiment, and by studying what happens to families in the months just following the time they satisfy their deductible," with a focus on dental care and deferrable outpatient medical care (p. 48).

Their analysis of free care found a surge in spending in the first few months of the experiment and at the end. They conclude that these "transient effects ... were very minor, representing no more than a doubling for the first quarter." This is roughly consistent with our findings, as well, although whether this is "very minor" is less clear. To study transient demand in the nonlinear cost-sharing plan, Keeler et al. look at spending after people hit the MDE. They break up the the year into three periods—before hitting MDE, the three months after hitting, and the remainder—and call the period just after hitting the the MDE the "sale" period. They find that, for most types of care, spending in the "sale" period is similar to spending in the post-sale period, and conclude that hitting the MDE does not generate a transient demand response. In a later analysis, Keeler and Rolph (1988) looked for anticipatory effects by looking at whether episodes became more common just before people hit the MDE. They found no such effect, and concluded that anticipatory effects were absente, likely because people could not easily predict when they would hit the MDE.

The key difference between our analysis and the original investigators is in the timing of when we look for intertemporal substitution and anticipatory responses. They focus on the period around hitting the MDE, before and after. We focus on the end of the coverage year. As Keeler and Rolph acknowledge, it is likely difficult to detect anticipatory effects or pent up demand by focusing on fine timing around hitting the MDE. Households may not know exactly when they hit the MDE (as has been pointed out by the original RAND investigators), and may not appreciate the link between their current and future spending (Einav et al., 2015; Dalton et al., 2015; Abaluck et al., 2015). On the other hand, by the end of the coverage year, most families who hit the MDE will have seen a bill which makes clear their financial position, and it is not hard to understand that in the future, prices will be higher. Indeed, providers may help make this clear. Thus the myopia or limited understanding of the insurance contracts may have made it difficult for the original investigators to identify intertemporal substitution; by looking at the end of the year, we avoid this difficulty.

7 Conclusion

Studying data from the RAND Health Insurance Experiment, we found striking patterns of health care spending over the insurance coverage year. In most months, spending is lower in cost-sharing plans than in the free care plan. But in the last 1-3 months of the coverage year, spending rises quickly in cost-sharing relative to free care, and by the end of the year spending in the two plans is roughly equal. On the other hand, spending in free care is roughly flat over the coverage year, but it is particularly high early in the first coverage year, and it spikes dramatically at the end of the experiment.

These patterns are inconsistent with the standard model of demand for health care, which assumes away intertemporal substitution. Instead they suggest that patients can retime their care, especially for medically deferrable procedures and dental care, to reduce their out-of-pocket expenses in the face of nonlinear cost-sharing rules. To quantify the importance of intertemporal substitution, we estimate how health care spending responds to lag and lead prices as well as current prices. The estimates suggest that short-run moral hazard—the response to a one time, unanticipated price change—is substantially larger than the long-run response.

These results have important implications for health care spending and insurance design. First, they help reconcile some of the disparate estimates of the price elasticity of health care demand, since they imply that health care spending is more responsive to temporary price changes—for example, hitting the deductible—than to permanent price changes, for example, from insurance plan changes. Second, they suggest that high deductible health plans may not be as effective as hoped in controlling health care spending. These plans can reduce health care spending as long as patients do not hit the deductible. But in years when patients do hit it—as they eventually will—the large short-run response means that spending will make up for lost time, as patients stock up on care.

The results also suggest avenues for future research. Although we believe that the results from the RAND Health Insurance Experiment provide strong evidence for intertemporal substitution, the data are now more than thirty years old, and it is unclear how closely they apply to the current health care landscape. If anything, we expect that there are more opportunities for intertemporal substitution now than there were in the past, for at least the following two reasons. First, there are now many more elective and preventive procedures possible than in the past, and many of these are likely straightforward to retime by at least a few months. Second, consumers nowadays are likely more aware of the prices they face, due to the increasing availability of information regarding one's insurance coverage and medical bills (Lieber, 2016). An important question for future research, then, is whether the behavior we have documented here shows up in modern health care plans, as well as whether the excess spending response to hitting the deductible represents high or low value care. Finally, another important question is how alternative contracts—such as a rolling-window for the deductible, or multiyear deductibles—affect spending and welfare.

References

Abaluck, Jason, Jonathan Gruber, and Ashley Swanson, "Prescription Drug Use Under Medicare Part D: A Linaer Model of Nonlinear Budget Sets," February 2015. NBER Working Paper No. 20976.

Alpert, Abby, "The Anticpatory Effects of Medicare Part D on Drug Utilization," *Journal of Health Economics*, 2016, *49*, 27–45.

Aron-Dine, Aviv, Liran Einav, Amy Finkelstein, and Mark Cullen, "Moral Hazard in Health Insurance: Do Dynamic Incentives Matter," *Review of Economics and Statistics*, 2015, *97* (4), 725–741.

_ , _ , and _ , "The RAND Health Insurance Experiment, Three Decands Later," *Journal of Economic Perspectives*, 2013, *27* (1), 197–222.

Bajari, Patrick, Christina Dalton, Han Hong, and Ahmed Khwaja, "Moral hazard, adverse selection, and Health Expenditures: A semiparametric analysis," *The RAND Journal of Economics*, 2014, *45* (4), 747–763.

Blau, David M. and Donna B. Gilleskie, "The Role of Retiree Health Insurance in the Employment Behavior of Older Men," *International Economic Review*, 2008, *49* (2), 475–514.

Blundell, Richard and Thomas MaCurdy, "Labor Supply: A Review of Alternative Approaches," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3A, Elsevier, 1999, pp. 1560–1695.

Brot-Goldberg, Zarek C., Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad, "What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantitites, and Spending Dynamics," November 2015. NBER Working Paper No. 21632.

Cabral, Marika, "Claim Timing and Ex Post Adverse Selection," *Review of Economic Studies*, 2016, *Forthcoming.*

Cardon, James H. and Igal Hendel, "Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey," *The RAND Journal of Economics*, 2001, *32* (3), 408–427.

Coe, Erica Hutchins, "Exchange Product Benefit Design: Consumer Responsibility and Value Consciousness," March 2014. McKinsey on Healthcare, available at `http://healthcare.mckinsey.com/exchange-product-benefit-design-consumer-responsibility-and-value-consciousness`.

Cronin, Christopher J., "Insurance-Induced Moral Hazard: A Dynamic Model of Within-Year Medical Care Decision Making Under Uncertainty," April 2016. Available at `http://christophercronin.weebly.com/uploads/3/9/0/4/39042047/insyearr.pdf`.

Dalton, Christina M., "Estimating demand elasticities using nonlinear pricing," *International Journal of Industrial Organization*, 2014, *37*, 178–191.

_ , Gautam Gowrisankaran, and Robert Town, "Myopia and Complex Dynamic Incentives: Evidence from Medicare Part D," September 2015.

DeNavas-Walt, Carmen and Bernadette D. Proctor, "Income and Poverty in the United States: 2014," 2015. U.S. Census Bureau, Current Population Reports, P60-252.

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

Einav, Liran, Amy Finkelstein, and Paul Schrimpf, "The Response of Drug Expenditure to Non-Linear Contract Design: Evidence from Medicare Part D," *Quarterly Journal of Economics*, 2015, *130* (2), 841–899.

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

Gilleskie, Donna B., "A dynamic stochastic model of medical care use and work absence," *Econometrica*, 1998, *66* (1), 1–45.

Grossman, Michael, "On the Concept of Health Capital and the Demand for Health," *Journal of Political Economy*, 1972, *80* (2), 1255–1282.

Hendel, Igal and Aviv Nevo, "Measuring the Implications of Sales and Consumer Inventory Behavior," *Econometrica*, 2006, *74* (6), 1637–1673.

_ and _ , "Sales and Consumer Inventory," *The RAND Journal of Economics*, 2006, *37* (3), 543–561.

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

_ , _ , Naihua Dunn, Janet Hanley, and Jr. Willard G. Manning, *The Demand for Episodes of Medical Treatment: Interim Results from the Health Insurance Experiment*, RAND Corporation (Pub. No. R-2829-HHS), 1982.

Khwaja, Ahmed, "Estimating willingness to pay for Medicare using a dynamic life-cycle model of demand for health insurance," *Journal of Econometrics*, 2010, *156*, 130–157.

Kleibergen, Frank and Richard Paap, "Generalized reduced rank tests using the singular value decomposition," *Journal of Econometrics*, 2006, *133* (1), 97–126.

Kowalski, Amanda E., "Estimating the tradeoff between risk protection and moral hazard with a nonlianer budget set model of health insurance," *International Journal of Industrial Organization*, 2015, *43*, 122–135.

_ , "Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care," *Journal of Business & Economic Statistics*, 2016, *34* (1), 107–117.

Lieber, Ethan M.J., "Does it Pay to Know Prices in Health Care?," *American Economic Journal: Economic Policy*, 2016, *forthcoming.*

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

Morris, Carl, "A finite selection model for experimental design of the health insurance study," *Journal of Econometrics*, 1979, *11* (1), 43–61.

Newhouse, Joseph P. and The Insurance Experiment Group, *Free for All? Lessons from the RAND Health Insurance Experiment*, Harvard Univeristy Press, 1993.

Schmid, Christian P.R., "Forward-looking Behavior in Health Insurance: Empirical Evidence from Switzerland," June 2016. Dartmouth Atlas Project.

Zweifel, Peter and Willard G. Manning, "Moral Hazard and Consumer Incentives in Health Care," in Anthony J. Culyer and Joseph P. Newhouse, eds., *Handbook of Health Economics, Vol. 1 Part A*, Elsevier, 2000, pp. 409 – 459.

Table 1: Summary statistics

Plan:	Free	25% coins	Mixed	Family Deductible	Individual Deductible
	(1)	(2)	(3)	(4)	(5)
	Mean	Difference in means, relative to free care			
Hit MDE (yearly)	1.00	−0.82	−0.75	−0.63	−0.54
	(0.00)	(0.02)	(0.03)	(0.02)	(0.02)
End-of-year price	0.00	0.21	0.19	0.59	0.51
	(0.00)	(0.01)	(0.01)	(0.02)	(0.02)
Maximum dollar expenditure	−11	2599	2387	2984	1508
	(10)	(58)	(61)	(67)	(20)
Medical spending	60.4	−17.2	−14.1	−27.3	−15.3
	(2.0)	(3.8)	(4.5)	(3.0)	(3.1)
Dental spending	49.6	−14.7	−17.6	−22.5	−16.5
	(2.0)	(3.4)	(3.9)	(2.8)	(3.0)
Mental spending	0.0	−0.0	−0.0	−0.0	0.0
	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Inpatient spending	67.8	−23.6	−1.5	−18.4	−6.7
	(5.5)	(9.6)	(12.6)	(8.1)	(8.6)
Deferrable medical episodes	0.08	−0.02	−0.01	−0.03	−0.02
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Acute episodes	0.23	−0.07	−0.06	−0.10	−0.07
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Dental episodes	0.18	−0.04	−0.05	−0.06	−0.06
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Chronic episodes	0.16	−0.05	−0.06	−0.07	−0.05
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Hospital episodes	0.04	−0.01	−0.01	−0.01	−0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Attrit	0.035	−0.005	0.011	0.058	0.034
	(0.006)	(0.017)	(0.018)	(0.019)	(0.014)
# Families	629	218	167	365	441
# People	1,677	403	499	838	1,174
# Person-months	78,348	25,236	20,424	41,424	48,888

Notes: Table shows the average of the indicated variable in the free care plan, and the difference relative to free care in the cost sharing plans, both adjusted for site-by-start-date differences (see text for details). Robust standard errors, clustered on family, in parentheses. Spending amounts are measured in 2011 dollars, and the unit of observation is a person month, except for attrition, where the unit of observation is a person. The average MDE is not exactly zero in free care because of the adjustment for site by start month differences. Sample consists of all person-months in the RAND fee-for-service plans, excluding the 50% coinsurance plan, excluding partial years cut short because of attrition, suspension, or birth, and excluding Dayton year 1, where only the free care plan covered dental services.

Table 2: Effect of current, past, and future prices on health care spending

Outcome	Spending			# Episodes				
Category	Medical (1)	Dental (2)	Inpatient (3)	Well-care (4)	Dental (5)	Acute (6)	Chronic (7)	Inpatient (8)
Price	-49.4	-116.9	-1.8	-0.095	-0.175	-0.076	-0.032	0.020
	(5.9)	(15.8)	(51.6)	(0.011)	(0.019)	(0.015)	(0.010)	(0.011)
Lag price	-0.7	-16.6	-10.6	0.023	0.014	-0.031	-0.002	-0.012
	(3.6)	(5.0)	(27.4)	(0.007)	(0.013)	(0.011)	(0.010)	(0.008)
Lead Price	25.3	110.4	-9.9	0.042	0.085	0.008	-0.033	-0.019
	(5.8)	(17.0)	(51.9)	(0.009)	(0.017)	(0.014)	(0.011)	(0.009)
Long-run effect	-24.8	-23.1	-22.4	-0.030	-0.076	-0.099	-0.067	-0.011
	(3.2)	(3.1)	(10.1)	(0.004)	(0.007)	(0.013)	(0.013)	(0.005)
Long minus short	24.6	93.8	-20.6	0.065	0.099	-0.023	-0.035	-0.031
	(6.4)	(16.4)	(53.7)	(0.011)	(0.020)	(0.017)	(0.015)	(0.011)
F-statistic	692.8	917.2	10618.8	692.8	917.2	692.8	692.8	10618.8
Mean dep. var.	48.3	38.1	59.8	0.064	0.146	0.183	0.124	0.035
Mean price	0.35	0.38	0.19	0.35	0.38	0.38	0.35	0.19

Notes: Table shows estimates from a regression of monthly spending or number of episodes in the indicated category on that category's spot, lag, and lead price. Additional controls include a set of dummies for date and site-by-start-date, plus dummies for year 1 and final month (when lag and lead price are imputed). We instrument for prices using a set of dummies for plan assignment interacted with year by coverage month. We report the Kleibergen and Paap (2006) F-statistic. The sample is defined as in the notes to Table 1 but additionally excludes observations missing lead or lag price. It consists of 213,730 person-months in 1,820 families. Robust standard errors, clustered on family, are in parentheses.

Table 3: Robustness of price sensitivity estimates

Outcome	Spending			# Episodes				
Category	Medical	Dental	Inpatient	Well-care	Dental	Acute	Chronic	Inpatient
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Control for predetermined variables interacted with time							
Short-run effect	-52.4	-111.3	4.6	-0.095	-0.171	-0.078	-0.047	0.020
	(6.1)	(15.7)	(52.9)	(0.011)	(0.020)	(0.016)	(0.012)	(0.012)
Long-run effect	-24.2	-23.6	-16.5	-0.029	-0.077	-0.096	-0.064	-0.008
	(2.8)	(3.1)	(9.3)	(0.004)	(0.007)	(0.012)	(0.011)	(0.005)
Long − short	28.2	87.7	-21.1	0.065	0.094	-0.017	-0.017	-0.029
	(6.5)	(16.4)	(54.5)	(0.011)	(0.020)	(0.017)	(0.015)	(0.011)
	Panel B: Restrict to continuously enrolled sample							
Short-run effect	-45.0	-128.7	10.7	-0.092	-0.169	-0.067	-0.033	0.023
	(6.3)	(18.2)	(53.1)	(0.011)	(0.021)	(0.015)	(0.011)	(0.012)
Long-run effect	-22.3	-23.8	-12.3	-0.028	-0.081	-0.091	-0.055	-0.011
	(3.7)	(3.7)	(11.0)	(0.005)	(0.008)	(0.014)	(0.014)	(0.006)
Long − short	22.8	104.9	-22.9	0.064	0.088	-0.021	-0.022	-0.034
	(6.8)	(19.0)	(55.2)	(0.012)	(0.022)	(0.018)	(0.015)	(0.012)
	Panel C: Include three lags/leads of price							
Short-run effect	-49.4	-117.9	-17.6	-0.099	-0.180	-0.071	-0.031	-0.011
	(6.0)	(15.9)	(12.0)	(0.011)	(0.020)	(0.015)	(0.011)	(0.006)
Long-run effect	-24.8	-21.8	-24.3	-0.027	-0.074	-0.101	-0.067	-0.011
	(3.3)	(3.3)	(10.9)	(0.005)	(0.007)	(0.013)	(0.014)	(0.006)
Long −short	24.6	96.1	-6.7	0.072	0.106	-0.030	-0.036	-0.000
	(6.7)	(16.8)	(12.0)	(0.012)	(0.021)	(0.018)	(0.016)	(0.005)
	Panel D: Omit first/last month							
Short-run effect	-39.8	-111.8	-49.6	-0.078	-0.147	-0.069	-0.052	0.042
	(6.5)	(17.3)	(74.7)	(0.011)	(0.021)	(0.017)	(0.012)	(0.015)
Long-run effect	-24.6	-22.8	-23.8	-0.030	-0.074	-0.099	-0.068	-0.011
	(3.2)	(3.1)	(10.3)	(0.004)	(0.007)	(0.013)	(0.013)	(0.005)
Long − short	15.2	89.1	25.8	0.048	0.072	-0.030	-0.016	-0.053
	(7.3)	(18.3)	(77.4)	(0.012)	(0.022)	(0.020)	(0.017)	(0.015)

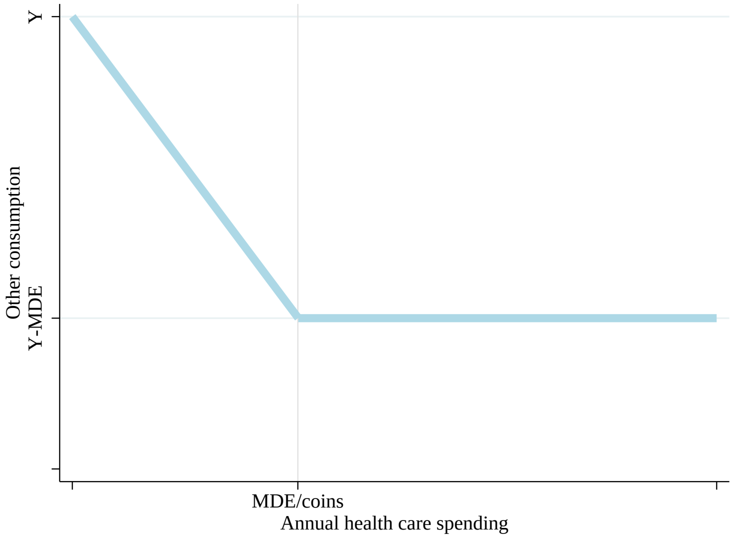
Notes: Robustness of results in Table 2; see notes there. Short-run effect is the coefficient on spot-price, and long-run effect is the sum of all leads and lags. In Panel A, we control for a set of fixed effects for interactions between between coverage month, year, and the predeterminend variables. In Panel B, we limit the sample to people who are continuously enrolled in the experiment for the assigned number of months. In Panel C, we control for the first three lags and leads of price. Robust standard errors, clustered on family, are in parentheses.

Table 4: Consequences of ignoring intertemporal substitution

Specification	Baseline (1)	Spot only (2)	Fixed effects (3)	Narrow variation (4)	Annual (5)
Long-run	-70.3	-75.6	-142.6	-72.4	-99.0
price sensitivity	(11.0)	(10.6)	(52.2)	(11.1)	(14.1)
Price	p_{t-1}, p_t, p_{t+1}	p_t	p_t	p_t	EOY p
Instruments	Plan-Month	Plan-Month	Plan-Month	Plan	Plan
Frequency	Monthly	Monthly	Monthly	Monthly	annual
Person fixed effects	No	No	Yes	No	No

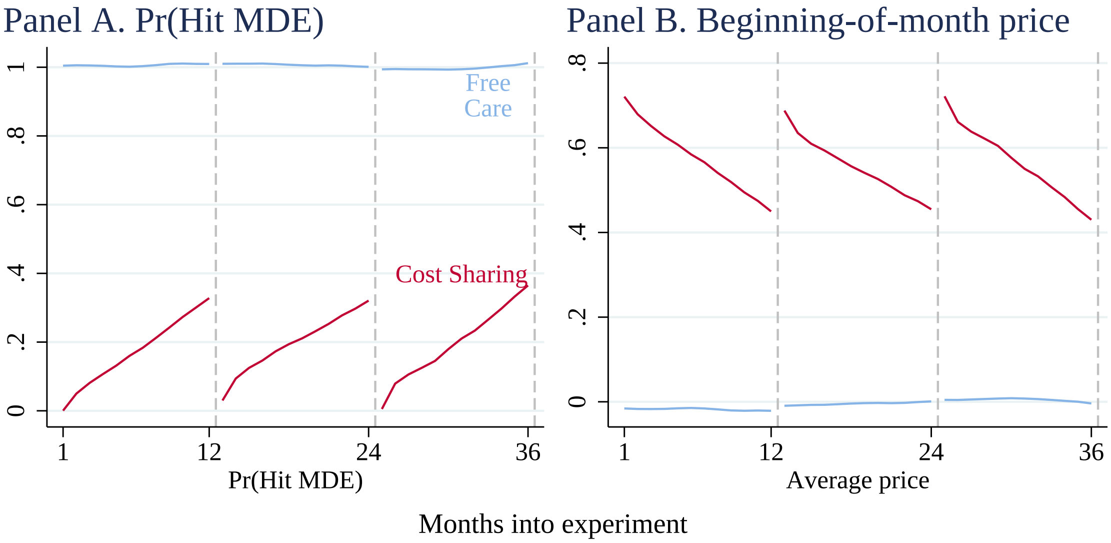
Notes: Table shows estimated long-run price sensitivity of total monthly spending, obtained from specifications that differ in the variation they use, the level of aggregation, or the way they measure price. Column (1) is the baseline result presented in Table 2. In columns (2)-(6), we ignore intertemporal substitution in various ways. All specifications also control for a trend (date fixed effects with the monthly data, or year × start date fixed effects in the annual data), and the monthly specifications include dummies for first and last month of the experiment. Robust standard errors, clustered on family, are in parentheses.

Figure 1: Annual budget set for plans with non-linear cost-sharing



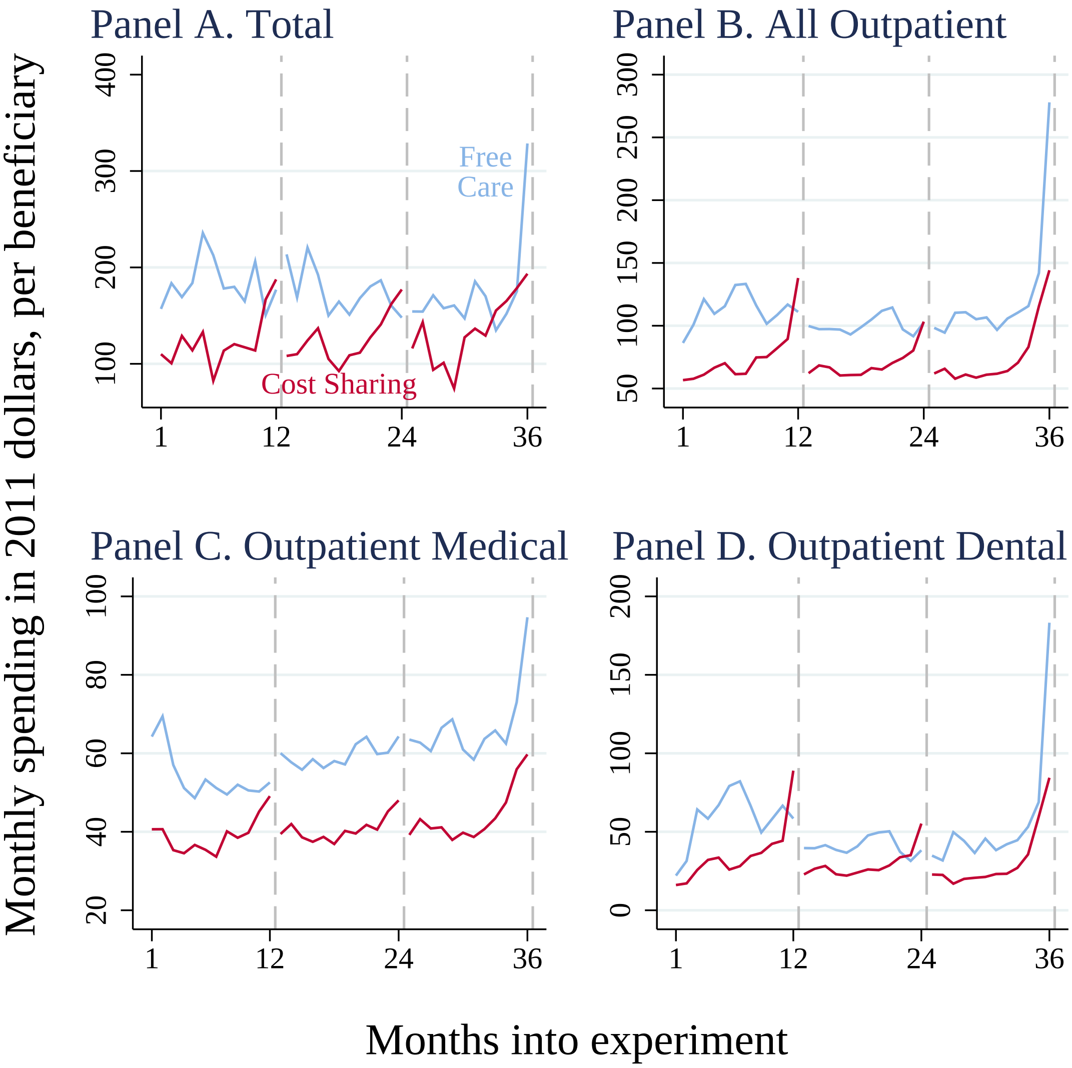
Notes: Figure shows the annual budget set created by the cost-sharing plans in the RAND Health Insurance Experiment (except for the individual deductible plan and mixed plan). Patients pay a coinsurance rate up to a maximum dollar expenditure, above which they do not pay on the margin for health care. See text for further details on the plans.

Figure 2: Prices by experiment month, free care vs. cost sharing



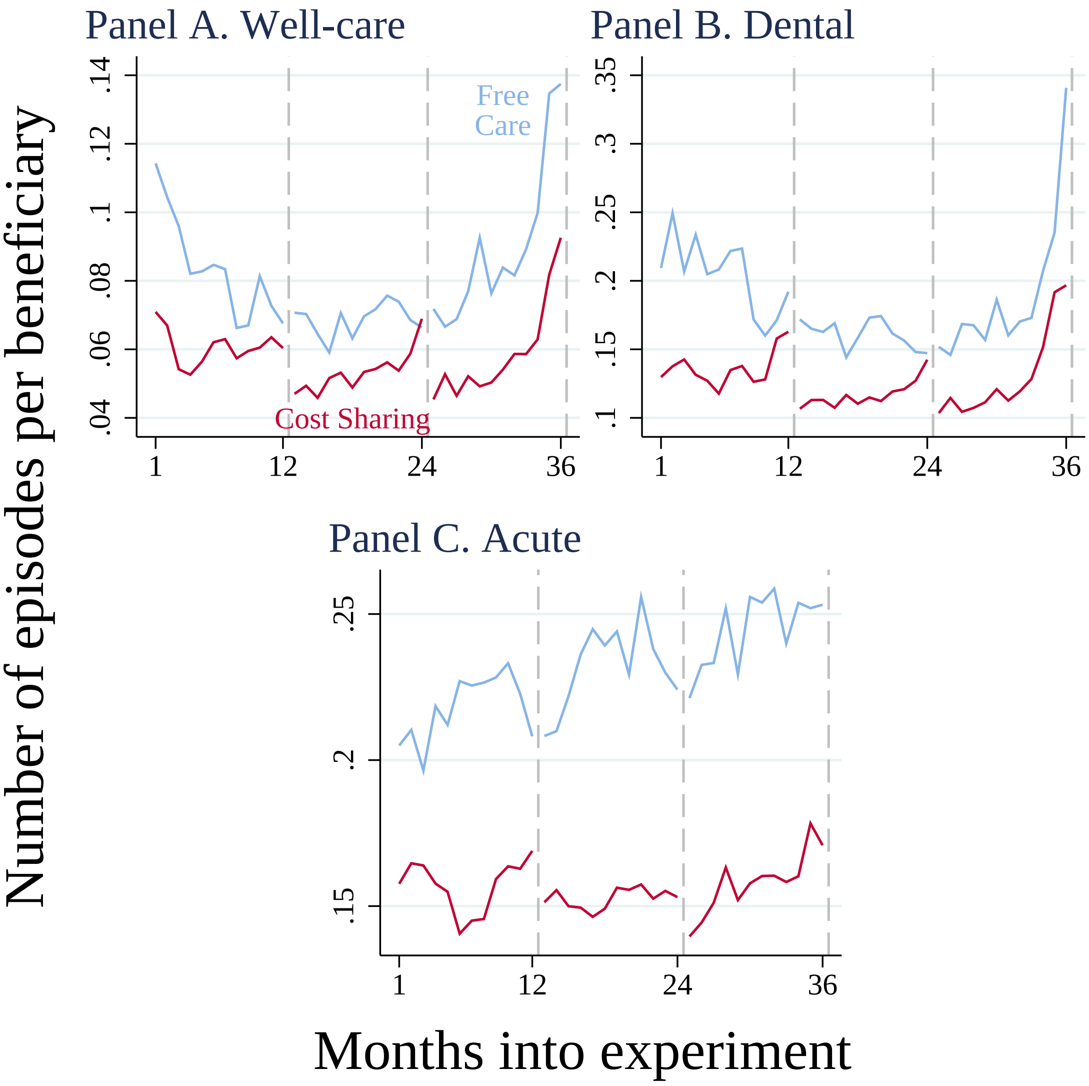
Notes: Figure shows, in Panel A, the probability that a given beneficiary has the the maximum dollar expenditure for the coverage year by the start of the month, and in Panel B the average beginning of month price, equal to zero for beneficiaries who have hit the MDE, and equal to the assigned coinsurance rate for those who have not. Months 1-12 are from the first year of coverage, 25-36 the last year of coverage, and months 13-24 pool all middle months. Averages are regression adjusted for date fixed effects and site-by-start date fixed effects using Equation 4.

Figure 3: Spending by experiment month, free care vs. cost sharing



Notes: Figure shows average spending in free care (in gray) and in cost-sharing plans (in black), in each month until end of coverage, for the indicated categories. Months 1-12 are from the first year of coverage, 25-36 the last year of coverage, and months 13-24 pool all middle months. Total spending is the sum of inpatient and outpatient spending, and outpatient spending decomposes into medical, dental, and mental care. Spending averages are regression adjusted for date fixed effects and site-by-start date fixed effects using Equation 4.

Figure 4: Number of episodes of treatment by experiment month, free care vs. cost sharing



Notes: Figure shows average number of episodes of treatment in free care (in gray) and in cost-sharing plans (in black), in each month until end of coverage, for the indicated categories. Months 1-12 are from the first year of coverage, 25-36 the last year of coverage, and months 13-24 pool all middle months. Episode counts are regression adjusted for date fixed effects and site-by-start date fixed effects using Equation 4.

A Balance tests

We test the validity of random plan assignment by looking at whether, in our analysis sample, plans differ in predetermined characteristics. The RAND investigators collected a host of information from potential study participants prior to randomization. This information was used to make sure that participants met eligibility criteria, and to assign participants to plans (in hopes of achieving balance). Following Aron-Dine et al. (2013), we divide pre-period variables into ones that directly measure health care utilization, and into all others. We test for balance with regressions of the following form:

$$y_i = \sum_p \beta_p 1\left\{plan = p\right\} + \mu_{SSD} + \varepsilon_i, \tag{A.1}$$

Here y_{it} is some pre-determined variable, and we include dummies for each of our plans. We expect that assignment is random only conditional on site by enrollment month, so we also include a set of demeaned site-by-enrollment month fixed effects. We omit the constant from this regression, so β_p is the average value of y in plan p, adjusting for differences in site and enrollment month. If plan assignment is random, we expect that $\beta_1 = \beta_2 = ...\beta_P$.

We estimate this regression using the pooled panel data, even though the outcome does not vary over time. Because our panel is not balanced, this approach gives more weight to people who remain in the panel longer. To the extent that differential attrition may bias our results, we would expect to find differences in predetermined characteristics, as attriters would be differentially underrepresented in some plans. Our test sample also excludes newborns—the only people who joined the experiment after it began—who are missing all predetermined variables.

Appendix Table A.1 shows the results; in Panel A we report the coefficients for the utilization variables, and in Panel B the coefficients for non-utilization variables. We also report the p-value of the hypothesis that the plan coefficients are jointly for each outcome, and at the bottom of each panel, that the coefficients are jointly equal across all outcomes. The table shows that the utilization variables are well-balanced: although there are some small differences in the probability of having a doctor and in the number of medical exams, they do not point to more utilization in the free care plan, and we fail to reject the null hypothesis that utilization is jointly equal. The results for the non-utilization variables are similar. The predetermined characteristics are balanced across our treatment groups.

This conclusion differs from Aron-Dine et al. (2013), who find that although plans appear mostly balanced at the moment enrollment is *offered*, differential refusal and attrition leads to unbalanced plans by the time the experiment concludes. There are several minor differences between our approach and theirs—they use a cross-section of people, we use a monthly panel, for example—but we show here that a key difference is the presence of the 50% coinsurance plan, which we exclude from our analysis and tests (because Aron-Dine et al. note that offers for that plan appear non-random), but Aron-Dine et al. include in their test for balance at experiment completion. To show this, in Appendix Table A.2 we repeat our balance analysis, but we include the 50% coinsurance plan, yielding the same set of plans as in Aron-Dine et al. (2013). In contrast to our earlier results, we now find statistically significant differences across the plans in the utilization and in the non-utilization variables.

Table A.1: Balance in pre-randomization variables

	Plan average					p-value of test
	Free	25% Coins	Mixed	Fam deduct	Ind deduct	for joint equality
	Panel A: Pre-period utilization variables					
Hospitalized	0.10	0.10	0.10	0.09	0.10	0.96
Missing hospitalized	0.03	0.02	0.01	0.02	0.02	0.18
Has doctor	0.98	0.98	0.97	0.97	0.97	0.98
Missing doctor	0.24	0.24	0.25	0.23	0.24	0.33
Had medical exam	0.50	0.53	0.47	0.50	0.43	0.04
Missing exam	0.04	0.03	0.04	0.04	0.03	0.80
# Medical visits	4.86	4.40	4.91	4.31	5.11	0.08
Missing visits	0.20	0.18	0.15	0.19	0.19	0.34
ln Medical spending	3.88	3.91	3.79	3.75	3.89	0.21
Missing spending	0.46	0.47	0.42	0.48	0.46	0.59
# Routine dental exams	0.73	0.70	0.71	0.70	0.72	0.94
Missing routine dental exams	0.39	0.33	0.38	0.35	0.42	0.06
# Special dental exams	0.54	0.50	0.57	0.54	0.54	0.66
Missing special exams	0.39	0.33	0.38	0.35	0.42	0.06
Jointly equal						0.14
	Panel B: Other predetermined variables					
Female	0.51	0.51	0.52	0.52	0.53	0.77
Age	24.40	23.99	23.99	24.26	25.08	0.69
White	0.45	0.46	0.45	0.44	0.50	0.13
Missing race	0.48	0.45	0.47	0.46	0.42	0.02
High school	0.21	0.22	0.23	0.22	0.25	0.14
More than HS	0.18	0.21	0.17	0.19	0.19	0.58
Missing education	0.41	0.42	0.43	0.41	0.37	0.10
From city	0.17	0.16	0.17	0.15	0.18	0.62
From suburb	0.07	0.06	0.05	0.07	0.06	0.62
From town	0.23	0.23	0.22	0.23	0.24	0.95
Missing backgrnd	0.40	0.41	0.42	0.40	0.37	0.16
Income ($thousands)	9.23	9.27	9.24	9.25	9.28	0.84
Income^2	85.61	86.32	85.79	85.98	86.51	0.82
Worked	0.85	0.89	0.83	0.86	0.86	0.61
Missing work	0.01	0.01	0.01	0.01	0.01	0.87
Any insurance	0.85	0.89	0.84	0.88	0.89	0.46
Missing insurance	0.06	0.04	0.03	0.05	0.04	0.08
Employer insurance	0.76	0.74	0.75	0.78	0.80	0.78
Missing employer insurance	0.23	0.20	0.20	0.24	0.22	0.10
Private insurance	0.14	0.17	0.14	0.16	0.16	0.92
Missing private insurance	0.23	0.20	0.20	0.23	0.22	0.12
Public insurance	0.09	0.08	0.08	0.07	0.07	0.67
Missing Public insurance	0.06	0.04	0.03	0.05	0.04	0.13
Excellent health	0.48	0.50	0.49	0.48	0.48	0.95
Good health	0.36	0.36	0.40	0.38	0.38	0.78
Missing health	0.06	0.04	0.03	0.05	0.05	0.17
Any pain	0.50	0.50	0.52	0.53	0.51	0.88
Missing pain	0.06	0.04	0.03	0.05	0.04	0.28
Any worry	0.40	0.42	0.41	0.38	0.37	0.61
Missing worry	0.06	0.04	0.03	0.05	0.05	0.16
Jointly equal						0.658
All equal						0.13

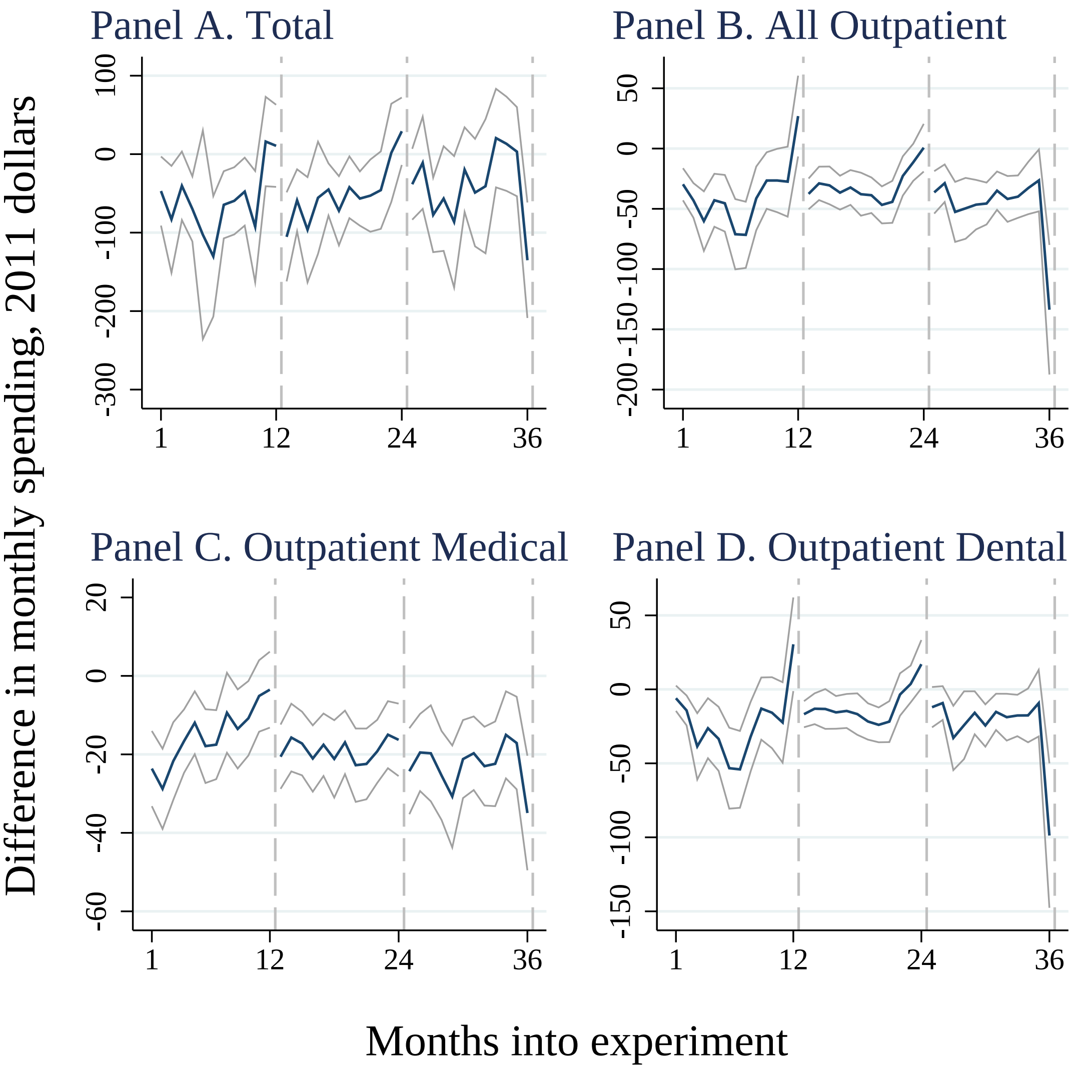
Notes: Table shows tests for balance across different plans by reporting average values of predetermined variables across the different plans. The first five columns show the estimated coefficients from a regression of the indicated variable on dummies for plan assignment, as well as demeaned site-by-start-date fixed effects (but no constant). The sample excludes the 50% coinsurance plan. The final column reports the p-value of the hypothesis that coefficients are jointly equal across plans.

Table A.2: Reconciliation with Aron-Dine et al.'s balance test

	Plan average						p-value of test
	Free	25% Coins	Mixed	50% Coins	Fam. deduct	Ind. deduct	for joint equality
	Panel A: Pre-period utilization variables						
Hospitalized	0.1	0.1	0.1	0.08	0.09	0.1	0.71
Missing hospitalized	0.03	0.02	0.01	0.02	0.02	0.02	0.24
Has doctor	0.97	0.97	0.97	0.98	0.97	0.97	0.96
Missing doctor	0.23	0.23	0.24	0.21	0.22	0.22	0
Had medical exam	0.5	0.52	0.47	0.52	0.5	0.43	0.05
Missing exam	0.04	0.02	0.04	0.03	0.04	0.03	0.74
# Medical visits	4.89	4.43	4.96	3.86	4.37	5.15	0.01
Missing visits	0.2	0.18	0.15	0.17	0.19	0.2	0.43
ln Medical spending	3.87	3.89	3.8	3.73	3.74	3.89	0.22
Missing spending	0.46	0.47	0.42	0.41	0.47	0.46	0.31
# Routine dental exams	0.73	0.71	0.71	0.77	0.7	0.72	0.79
Missing routine dental exams	0.39	0.33	0.38	0.38	0.35	0.42	0.12
# Special dental exams	0.54	0.5	0.57	0.61	0.54	0.54	0.36
Missing special exams	0.39	0.33	0.38	0.38	0.35	0.42	0.12
Jointly equal							0.00
	Panel B: Other predetermined variables						
Female	0.51	0.51	0.52	0.51	0.51	0.53	0.81
Age	24.41	23.97	23.96	24.31	24.27	25.06	0.81
White	0.45	0.46	0.45	0.48	0.44	0.5	0.17
Missing race	0.48	0.45	0.47	0.45	0.46	0.42	0.05
High school	0.21	0.22	0.23	0.22	0.22	0.25	0.23
More than HS	0.18	0.21	0.17	0.19	0.19	0.19	0.8
Missing education	0.41	0.42	0.43	0.4	0.41	0.37	0.15
From city	0.17	0.16	0.17	0.12	0.15	0.18	0.1
From suburb	0.07	0.06	0.05	0.07	0.07	0.06	0.7
From town	0.23	0.23	0.22	0.29	0.23	0.24	0.22
Missing backgrnd	0.41	0.41	0.42	0.39	0.4	0.37	0.24
Income ($thousands)	9.24	9.28	9.25	9.3	9.26	9.28	0.85
Income^2	85.7	86.42	85.86	86.76	86.07	86.57	0.84
Worked	0.86	0.9	0.83	0.94	0.86	0.86	0.01
Missing work	0.01	0.01	0.01	0	0.01	0.01	0.11
Any insurance	0.86	0.89	0.84	0.89	0.88	0.89	0.66
Missing insurance	0.06	0.03	0.03	0.04	0.05	0.04	0.1
Employer insurance	0.61	0.59	0.6	0.64	0.62	0.64	0.89
Missing employer insurance	0.25	0.22	0.21	0.23	0.25	0.23	0.16
Private insurance	0.14	0.16	0.14	0.11	0.16	0.16	0.9
Missing private insurance	0.25	0.22	0.22	0.22	0.25	0.24	0.04
Public insurance	0.09	0.08	0.07	0.06	0.07	0.07	0.66
Missing Public insurance	0.06	0.03	0.03	0.04	0.05	0.04	0.15
Excellent health	0.48	0.51	0.49	0.54	0.48	0.48	0.78
Good health	0.36	0.36	0.4	0.32	0.38	0.38	0.5
Missing health	0.06	0.04	0.03	0.04	0.05	0.05	0.26
Any pain	0.5	0.5	0.52	0.47	0.52	0.51	0.77
Missing pain	0.06	0.04	0.03	0.04	0.05	0.04	0.4
Any worry	0.4	0.41	0.41	0.38	0.38	0.37	0.72
Missing worry	0.06	0.04	0.03	0.04	0.05	0.05	0.25
Jointly equal							0.06
All equal							0.00

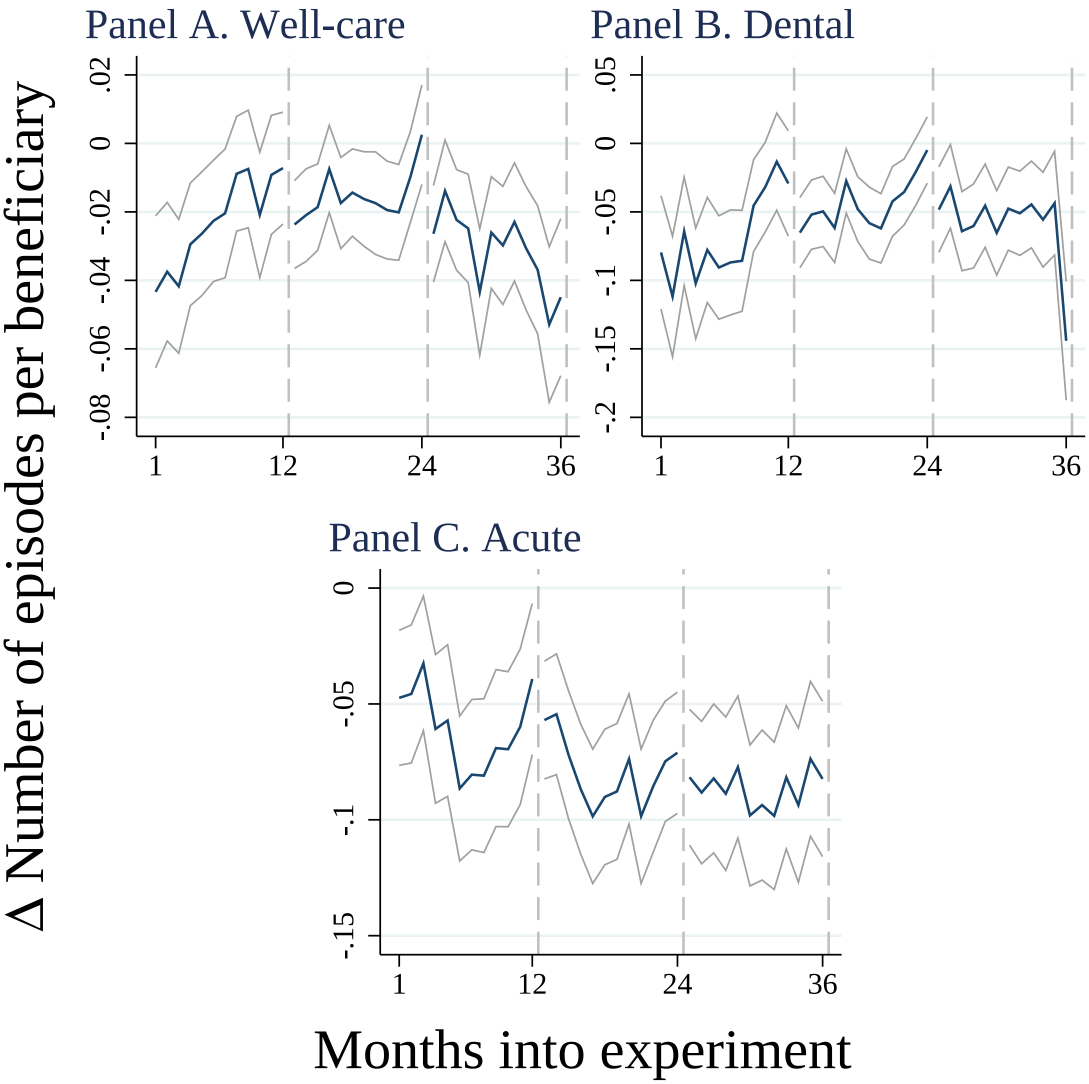
Notes: Table reconciles our balance tests in Table A.1 with the balance tests reported in Aron-Dine et al. (2013), by showing that when we follow Aron-Dine et al.'s classification of plans and include the 50% coinsurance plan, we fail the balance tests (as they do). The first six columns show the estimated coefficients from a regression of the indicated variable on dummies for plan assignment, as well as demeaned site-by-start-date fixed effects (but no constant). The final column reports the p-values of the hypothesis that coefficients on the plan dummies are all equal

Figure A.1: Month-specific treatment effects for spending



Notes: Figure shows the coverage-month specific effect of being in a cost-sharing plan relative to the free care plan, on spending in the indicated category, corresponding to $\gamma_{\tau}^{Year} - \beta_{\tau}^{Year}$ in Equation 4. The light gray lines give 95% confidence intervals, constructed from standard errors clustered on family.

Figure A.2: Month-specific treatment effects for episodes of care



Notes: Figure shows the coverage-month specific effect of being in a cost-sharing plan relative to the free care plan, on episodes of the indicated type, corresponding to $\gamma_{\tau}^{Year} - \beta_{\tau}^{Year}$ in Equation 4. The light gray lines give 95% confidence intervals, constructed from standard errors clustered on family.