Intertemporal Substitution in Health Care Demand: Evidence from the RAND Health Insurance Experiment*

Haizhen Lin and Daniel W. Sacks

January, 2019

Abstract

Nonlinear cost-sharing in health insurance encourages intertemporal substitution because patients can reduce their out-of-pocket costs by concentrating spending in years when they hit the deductible. We develop a test for intertemporal substitution and apply it to 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 health care demand than does being in free care, because people who hit the MDE face low current prices but high future prices, and so stock up on health care. As a result, short-lasting price changes induce nearly twice as big a response as do long-lasting changes. These findings help reconcile disparate estimates of the price elasticity of demand for health care in the existing literature. Failing to account for intertemporal substitution can lead researchers to overstate cost savings from high deductible health plans by 20 percent or more.

JEL codes: D12, G22
Key words: Health insurance, Intertemporal substitution, moral hazard, dynamic incentives

*Lin: Indiana University, Kelley School of Business, and NBER, email: hzlin@indiana.edu; Sacks: Indiana University, Kelley School of Business, email: dansacks@indiana.edu. We thank the Editor, two helpful referees, our discussant Chris Cronin, as well as Guy Arie, Anthony Defusco, Mark Duggan, Johannes Haushofer, Anita Mukherjee, Maria Polyakova, 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 not only for their research and insights but also for providing publicly available, well documented replication data. Jordan Keener provided excellent research assistance.

High deductible health plans have become common. In 2016, 29 percent of workers with employer-sponsored insurance were in such plans (Kaiser Family Foundation and Health Research and Educational Trust, 2017). On the Health Insurance Exchanges, three-quarters of plans offered in 2014 had deductibles of at least $1,250 for single coverage (Coe, 2014). High deductible plans offer a tradeoff of worse risk protection against possible reductions in health care spending. Such possible reduction in spending implicitly assumes that care foregone in one year because of the high deductible represents a permanent reduction in spending.[1] 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, 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. Therefore, a key question for the effectiveness of high-deductible health plans is whether patients intertemporally substitute in their demand for health care.

In this paper, we develop a test for intertemporal substitution in health care demand. The essence of this test is that, in the absence of intertemporal substitution, people who hit the deductible and therefore face a *temporarily* low price should spend at the same level as people who *permanently* face the same low price, all else equal. But with intertemporal substitution, the person facing the temporary price cut would have higher spending. We then test for intertemporal substitution by looking for an excess response to hitting the deductible, relative to facing a persistently low price.

Testing for an excess response to hitting the deductible requires addressing several challenges. First, patients who hit the deductible are likely to be high spenders. We avoid this problem by comparing all patients in cost-sharing to all patients in free care, therefore getting around the endogeneity problem. Second, high spending patients are likely to select plans with less cost-sharing, making it difficult to identify the spending effect of different plans. We solve this problem by using data from the RAND Health Insurance Experiment (Newhouse and The Insurance Experiment Group, 1993). The experiment entailed random assignment to either a free care plan, or to one of several plans with nonlinear cost-sharing,

[1] High deductible plans could also reduce spending by encouraging consumers to search for low prices. But the literature has found no or little evidence of consumer price shopping (Brot-Goldberg et al., 2015; Lieber, 2016).

with a fixed coinsurance rate up to a maximum dollar expenditure (MDE). Last, it is not obvious what the relevant price is. Theory suggests that people should respond to their expected end-of-year price, i.e. the expected out-of-pocket price on the last dollar of health care spending, given current information (Ellis, 1986). Recent evidence, however, suggests that people respond heavily to the "spot price" of health care, i.e. the out-of-pocket price of the next dollar of health care spending. We sidestep this issue by focusing on spending in the end of a coverage year, when the spot price and the expected end of yea price are essentially the same.

We find clear evidence for intertemporal substitution. At the beginning of a coverage year, we find that spending is substantially higher in the free care plan, as expected. By the end of a coverage year, however, the situation reverses: average spending in the cost-sharing plans is slightly higher than in free care. Thus, the one third of people who do hit the MDE (in the data) have a much larger spending increase than do people permanently assigned a price of zero. We interpret this extra spending as intertemporal substitution because it shows up most clearly in the month immediately before a price change (i.e., at the end of a coverage year), and because it is clearest for dental care and other deferrable procedures, and smallest in acute care, which is likely difficult to retime. As further evidence of intertemporal substitution, we find that patients in free care have particularly high spending in the last month of the experiment, after which they will return to less generous insurance.

One implication of intertemporal substitution is that people respond differently to temporary and long lasting price changes. We provide a simple quantification using a reduced-form dynamic specification. We find that a temporary decrease of 10 percentage points in the coinsurance rate increases spending by 12 percent of the mean level, and increases utilization (as measured by episodes of care) by 7 percent. A permanent price change has less than half as large an effect. This specification also allows us to decompose the divergence between short- and long-run price sensitivities into an anticipation effect (during sales, people stock up on 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: after 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.

For utilization, 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.

Another implication of intertemporal substitution is that divergent price elasticity estimates in the literature could be explained by different sources of price variation used for identification. 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). Price sensitivities estimated from within-person (short-run) variation are twice as large as those that rely purely on cross-plan (long-run) variation. These static models can also give misleading spending implications under nonlinear cost-sharing plans. In particular, intertemporal substitution can work to undermine cost savings that otherwise might be achieved under a high deductible plan. To quantify the magnitude of bias, we simulate spending under linear and nonlinear high deductible health plans. We find models that neglect dynamics overstate savings associated with moving from a generous plan to a high deductible plan; the bias can be 20 percent or more.

Our results contribute to a growing literature on dynamic decision making in the face of nonlinear cost-sharing in insurance. Much of this literature has focused on the rationality of decision making. Keeler et al. (1977) is one of the first papers that model dynamic decision making of medical demand under a nonlinear contract. Ellis (1986) shows that a rational, forward-looking person should not respond to the current or "spot" price of the next dollar of health care, but only her expected end-of-year price of the last dollar of care. Recent literature has found, however, that people respond heavily to spot prices, implicitly discounting future savings at high rates (Abaluck et al., 2015; Brot-Goldberg et al., 2015; Dalton et al., 2015; Einav et al., 2015; Sacks et al., 2017), although Aron-Dine et al. (2015) show that people do respond to the dynamic incentives of nonlinear contracts. The little evidence of forward-looking behavior in the existing literature might be explained by people having difficulty in anticipating whether they will hit the deductible (or using this information optimally).[2] However, by focusing on the last month of a coverage year, we

[2] The original RAND investigators found little intertemporal substitution in the experiment (Keeler et al.,

show that once people do hit the deductible, they react to the fact that their future prices will be relatively high. Our results therefore suggest that future prices also matter. Note that we are not the first to identify intertemporal substitution in the demand for health care. For example, Einav et al. (2015) and Cabral (2016) study coverage maxima under Medicare Part D and dental insurance, respectively. Different from their approach, we offer a test of intertemporal substitution, which could be implemented using aggregate plan-level data on health care demand, given exogenous variation in plan assignment.[3] We also show that deductibles, not just coverage maxima, generate intertemporal substitution.

Our results also contribute to the large literature on moral hazard effects of health insurance. Existing literature has largely neglected intertemporal substitution, and estimated a wide range of price elasticities, from as small as -0.2 to as large as -1.5 (Manning et al. (1987); Eichner (1998); Zweifel and Manning (2000); Cardon and Hendel (2001); Bajari et al. (2014); Dalton (2014) and Kowalski (2015, 2016)). Our results help reconcile these disparate estimates, because intertemporal substitution implies that a short-lasting price change, for example from hitting the deductible, generates a much larger response than a permanent price change, for example from different plan assignment. Indeed, some of the highest estimated price elasticities are identified from exogenous variation in hitting the deductible (Eichner, 1998; Kowalski, 2016).

Finally, our paper contributes to the literature in Marketing and Industrial Organization that studies how consumers respond to sales. These sales are a valuable source of price variation, but if intertemporal substitution is possible, then the response to sales will be different than the response to a long-run price change, which is the usual object of interest. Hendel and Nevo (2006b) provide clear evidence for intertemporal substitution for storable groceries. Erdem et al. (2003), Hendel and Nevo (2006a), and Hartmann (2006) use structural models to estimate long-run price elasticities from sales. We contribute to this literature by showing that nonlinear pricing rules generate sales-like effects, and showing that here too intertemporal substitution affects estimated price sensitivities.

1982); we reconcile this finding in Appendix E.

[3] Our identification relies on random plan assignment in the RAND experiment. But the test can be applied to non-experimental settings with exogenous variation in plan assignment. For example, data from multiple employers which introduce high deductible health plans over time could be used.

Table 1: Spending in free care and family deductible plan, selected months of coverage year

	Average outpatient spending		Difference	% who
	Free care	Deductible plan	Free−Deductible	hit deductible
	(1)	(2)	(3)	(4)
First month of year	95.2	48.6	46.7	0.04
	(5.7)	(4.7)	(7.6)	(0.01)
Last month of year	98.8	103.6	-4.9	0.37
	(7.2)	(14.9)	(16.6)	(0.03)

The sample consists of 2,945 people in 945 families, in the free care plan and the family deductible plan of the RAND Health Insurance Experiment; it excludes the first and last year of the experiment. Columns (1) and (2) show average outpatient spending in the indicated plan and month, column (3) shows the difference; spending amounts are adjusted for date and site-by-start date fixed effects as described in Section 3. Column (4) shows the fraction of people in the family deductible plan who hit the deductible (i.e. the maximum dollar expenditure) by the indicated month. Robust standard errors, clustered on family, are in parentheses.

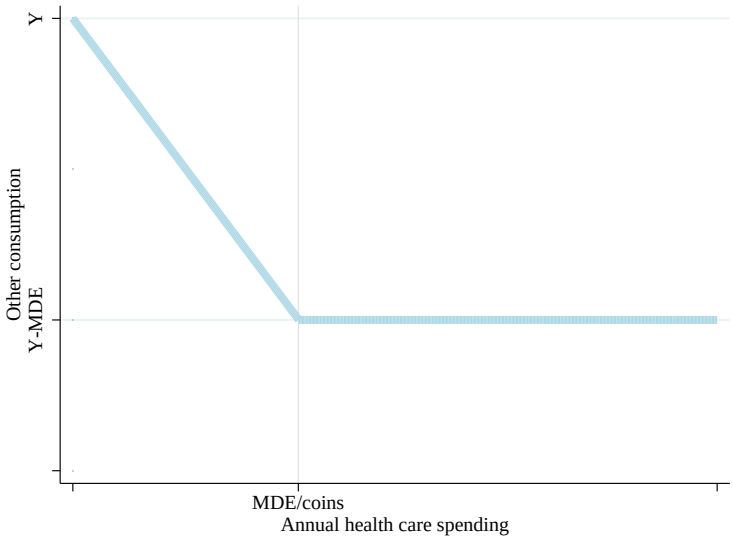
This paper is structured as follows. The next section presents the model. Section 2 describes the experiment and data. Section 3 shows the results of our test, for spending and utilization. Section 4 presents estimates of long- and short-run elasticities, and Section 5 offers implications for insurance design based on a simulation. The final section concludes.

1 A test for intertemporal substitution

1.1 Testing for intertemporal substitution

Our test for intertemporal substitution is based on the the pattern of spending over the coverage year. We illustrate the logic of our test—and why we find intertemporal substitution in the RAND Health Insurance Experiment—in Table 1. The table shows average spending for people in two plans from the Experiment: the free care plan, which had no cost-sharing, and the family deductible plan, which had a 95% coinsurance rate, but capped out-of-pocket spending at a maximum dollar expenditure (MDE) of roughly $3,000 (in 2011 dollars). Families whose spending reaches the MDE face a marginal price of zero for subsequent care.

Figure 1: Annual budget set for plans with nonlinear 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 1 illustrates the budget set.[4]

In the first month of a coverage year, health care spending is more than twice as high in the free care plan as in the cost-sharing plan. This is expected, as health care is heavily subsidized in the free care plan. By the end of a coverage year, however, spending in the two plans is nearly equal—in fact, slightly higher in the deductible plan. One possible explanation for this convergence in demand is a convergence in prices; perhaps most families hit the MDE by the end of year, meaning that the realized price is nearly zero in the deductible plan as well as in free care. But in fact the last column shows that only 37% of families hit the MDE. Thus the average price is much higher in the deductible plan than in the free care plan.

Our explanation for the convergence of spending, despite the non-convergence of prices, is intertemporal substitution. With intertemporal substitution, people delay health care in high-price periods, waiting until they face a low price to stock up on care. Hitting the MDE causes a much bigger increase in demand than does facing a permanent price of zero,

[4] See Section 2 for much more information on the structure of the experiment and the data. Table 1 excludes observations from the first and last year of the experiment, so that beginning- and end-of-experiment effects in free care do not contaminate the results.

because families who hit the MDE could have delayed spending until doing so, and because such families know that they will face higher prices in the future (unlike families in free care).[5]

In Appendix A, we present a formal model of health care demand with nonlinear cost-sharing. The key result from the model is that, in the absence of intertemporal substitution, spending in a free care plan in the final month of a coverage year will be larger than spending in a cost-sharing plan, as long as not everyone in cost-sharing has hit the maximum dollar expenditure. The key assumption required for this result, beyond intertemporal substitution, is (quasi-)random assignment to different plan types. This assumption is satisfied in our empirical setting because the HIE entailed random assignment.

An advantage of our test is that it does not require us to take a stand on how people form expectations about health care prices. The result (that in the absence of intertemporal substitution, spending in a free care plan in the final month of the coverage year will be larger than spending in a cost-sharing plan) holds under perfect foresight, rational (but imperfect) expectations, or complete myopia. This is useful because it is otherwise unclear whether health care demand depends on the expected end-of-year price (as it would for a rational consumer), on the current spot price, or on some combination of the two. In addition, calculating the expected end-of-year price is itself challenging since it depends on people's information sets. We do not need to take a stand on expectations because in the final month of the coverage year, expectations are irrelevant (in the absence of intertemporal substitution); nearly everyone knows their end-of-year price, and it is essentially the same as the spot price, except for those whose next purchase will cause them to hit the MDE.

We note that our test is underpowered in three senses. First, intertemporal substitution implies that hitting the MDE has a bigger per-month effect on spending than does being in free care. It would be natural to compare spending of people who hit the MDE to people in free care. The problem with that approach is that hitting the MDE is endogenous in the sense that it requires high spending. We therefore compare spending of all people in free care to spending of all people in the cost sharing plans, only some of whom hit the MDE.

[5] In practice about 40 percent of enrollees who hit the MDE in a cost-sharing plan in one year do not hit it in the next year.

Because we average over people who do and not hit the MDE, our test might fail to detect the true presence of intertemporal substitution. Second, income effects work against finding an effect of hitting the MDE. People who hit the MDE face the same prices as people in free care, but their disposable income is lower by exactly the amount of MDE. Because health care is likely a normal good, this pushes down health care demand, and may cause the test to fail to detect intertemporal substitution. Third, people who do not hit the MDE in a given year may want to reduce care towards the end of that year and use more care next year, when they have a chance of hitting the MDE. This substitution reduces end-of-year spending in the cost sharing plan, making it harder for us to detect intertemporal substitution. As we end up detecting intertemporal substitution, the weakness of our test only reinforces our conclusion that intertemporal substitution is a meaningful part of health care demand.

1.2 Microfoundations and implications of intertemporal substitution

Our test is a test of the null hypothesis of no intertemporal substitution. It 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 alternatives 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. A forward-looking person can therefore reduce health care spending, without hurting health, by substituting care to periods with low relative prices. An anticipated price increase tomorrow may therefore 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.

Both the "stock of health" and the "retiming care" model have testable implications. First, they imply an excess response to hitting the deductible, relative to free care. This response should be especially strong at the end of a coverage year, when there is a large, looming price increase. These models also 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. Both these models therefore imply that hitting the MDE should have a larger per-month effect than being in free care. Thus our test for intertemporal substitution—looking for excess spending in the cost-sharing plan at the end of a coverage year—is both an affirmative test of models generating such substitution, as well as a test against the null hypothesis of no intertemporal substitution.

2 Background and data

2.1 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.[6] The experiment ran at six different sites, chosen to

[6] Newhouse and The Insurance Experiment Group (1993) provides a detailed overview of the experimental design results of the experiment. Aron-Dine et al. (2013) offers 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

be broadly representative of the United States, and new families were enrolled over several different 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.[7] 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.[8] Because the 95% coinsurance plan resembles 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.[9] 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

should have.

[7] 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.

[8] 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.

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

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).

2.2 Data and summary statistics

We use the replication files which the original RAND investigators have made publicly available.[10] 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.[11] 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 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

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

[11] 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.

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.[12]

An advantage of episodes of care, relative to spending, is that they are designed to accounts for the lumpiness of healthcare. This lumpiness arises because health needs can require a series of treatments, which makes it hard for patients simply to spend more upon hitting the MDE. However, if some procedures or episodes are deferrable, then we expect them to be put off until the price is low. Looking at episodes let us detect this type of substitution.

We make four restrictions to create our analysis sample. First, following the original RAND investigators, we only include whole coverage years.[13] 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 2 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 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.

[12] 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.

[13] 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.

Table 2: Summary statistics

Plan:	Free	25% coins	Mixed	Family Deductible	Individual Deductible
	(1)	(2)	(3)	(4)	(5)
	Mean	Difference in means, relative to free care			
Porb(Hit MDE) (yearly)	1.00	−0.82	−0.75	−0.63	−0.54
	(0.00)	(0.02)	(0.03)	(0.02)	(0.02)
Average 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, are 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.

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. 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, but end-of-year prices are increasing with the sticker coinsurance rate of the plan. The reported average end-of-year price is the coinsurance rate times the fraction of people who do not hit the MDE by the end of the coverage year. It is therefore the average, realized end-of-year price.[14]

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.

2.3 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 C; we find that the plans appear balanced on pre-experimental characteristics, including both direct measures of utilization and demographic variables.

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 unbal-

[14] Under rational expectations, of course, the average realized end-of-year price must equal the average expected end-of-year price, so this can also be interpreted as the average expected end-of-year price.

anced, on utilization or other variables, among people who complete the experiment. In Appendix C, 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.

3 Treatment effects over the coverage year

3.1 Estimating equations

In Section 1 we argue 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 a coverage year. To do so, we estimate monthly demand in these plans, adjusting for differences in site-by-start date and for general trends. Specifically, we estimate regressions of the following form:

$$y_{it} = \sum_{year=\{\text{first, middle, last}\}} \left[\sum_{\tau=1}^{\tau=12} (\beta_{\tau}^{year} Free_{i\tau,year} + \gamma_{\tau}^{year} Cost_{i\tau,year}) \right] + \mu_t + \theta_{sem} + \varepsilon_{it}, \quad (1)$$

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).[15] 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

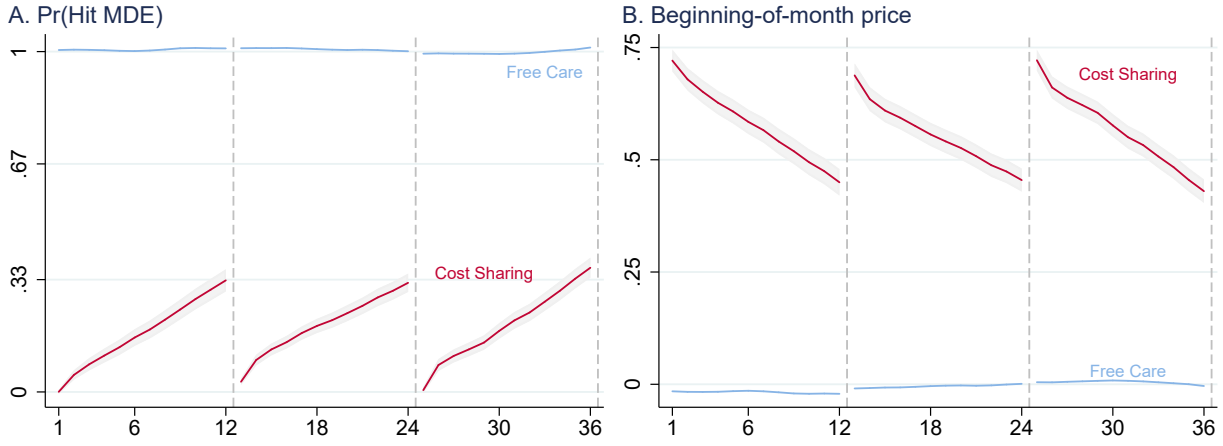
[15] Calendar months and coverage months are not collinear because the experiment had staggered start dates, with every calendar month a possible start month.

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}^{first} 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 $year$. In particular, we are interested in $\beta_{12}^{year} - \gamma_{12}^{year}$, which measures the difference in outcomes between free care and cost-sharing plans, for the last month of coverage year $year$.

In all specifications, we control for a full set of demeaned calendar time dummies, μ_t. These are dummy variables indicating each month the experiment ran, for example "December 1978." Our estimates therefore are not due to a "December" effect or other seasonality in health care demand. Such controls are possible because the HIE involved staggered start dates; families entered the experiments in waves, beginning in November 1974, with new families entering every month until February, 1979. Each family's coverage year ends 12 months after it enters the sample, so cover years end in every calendar month. Because plan assignment was only random conditional on these different start month (and conditional on enrollment site), we also control for a full set of start site by enrollment month dummies, θ_{sem} (these are also demeaned).

In estimating Equation 1, our sample includes people assigned to both 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.

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



Notes: Figure shows the probability that a given beneficiary hits the MDE for the coverage year by the start of the month, and the average beginning of month price. 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 1. To show both the level and the statistical significance of the difference, the shaded region shows the 95% confidence interval for the *difference*, centered on the cost-sharing means.

3.2 Results

Prices We begin by verifying that end-of-year prices are in fact different between the cost-sharing and free care plans.[16] 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.[17] 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 a coverage year, and as a result prices fall steadily. However, by the start of the last month of a coverage year, only about a third of people have hit the MDE, so even in the last month, 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 a coverage year.[18]

[16] 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).

[17] We plot the average spot price rather than some measure of the expected end-of-year price, because the average expected end-of-year price cannot change within the coverage year under rational expectations.

[18] 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, very similar to the average fraction of people who have hit the MDE

Spending Figure 3 shows the results for spending. This figure plots the estimated values of β and γ from Equation 1. To assess the precision and statistical significance of our findings, we also plot as a shaded area the 95% confidence interval for the difference in monthly treatment effects, $\beta_\tau^{Year} - \gamma_\tau^{Year}$. We center this confidence interval around the γs. When the shaded area excludes the estimated β_τ^{Year}, the monthly difference between the free care and cost-sharing plans is statistically significant.

Panel A shows the results for total spending and several striking patterns emerge. 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 a coverage year, at around $125 per person, and then rises sharply in the second half of the year. By the end of a coverage year, spending in the two groups has converged, and in the final month of the middle year, spending is actually much higher in the cost-sharing plan than in free care, in sharp contrast to the predictions of the model without intertemporal substitution.

Because the results for total spending are noisy, we also remove inpatient spending and plot total outpatient spending in Panel B. Total outpatient spending shows a similar pattern. Spending surges in cost-sharing at the end of a coverage year, and eventually rises above spending in free care except for the final year of the experiment. Further decomposing outpatient spending into medical (Panel C) and dental care (Panel D) 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 of 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 a coverage year, although it is more pronounced for dental care than for medical.

Overall, we find the difference in spending between free care and cost sharing is large and

by the start of the last month, and similar to the average price at the beginning of the last month of a coverage year.

statistically significant early in a coverage year, and becomes small or zero and insignificant close to the end of a coverage year. We see this pattern for both total spending (although noisy) and all outpatient spending. For outpatient care, these results seem to be driven largely by outpatient dental care.

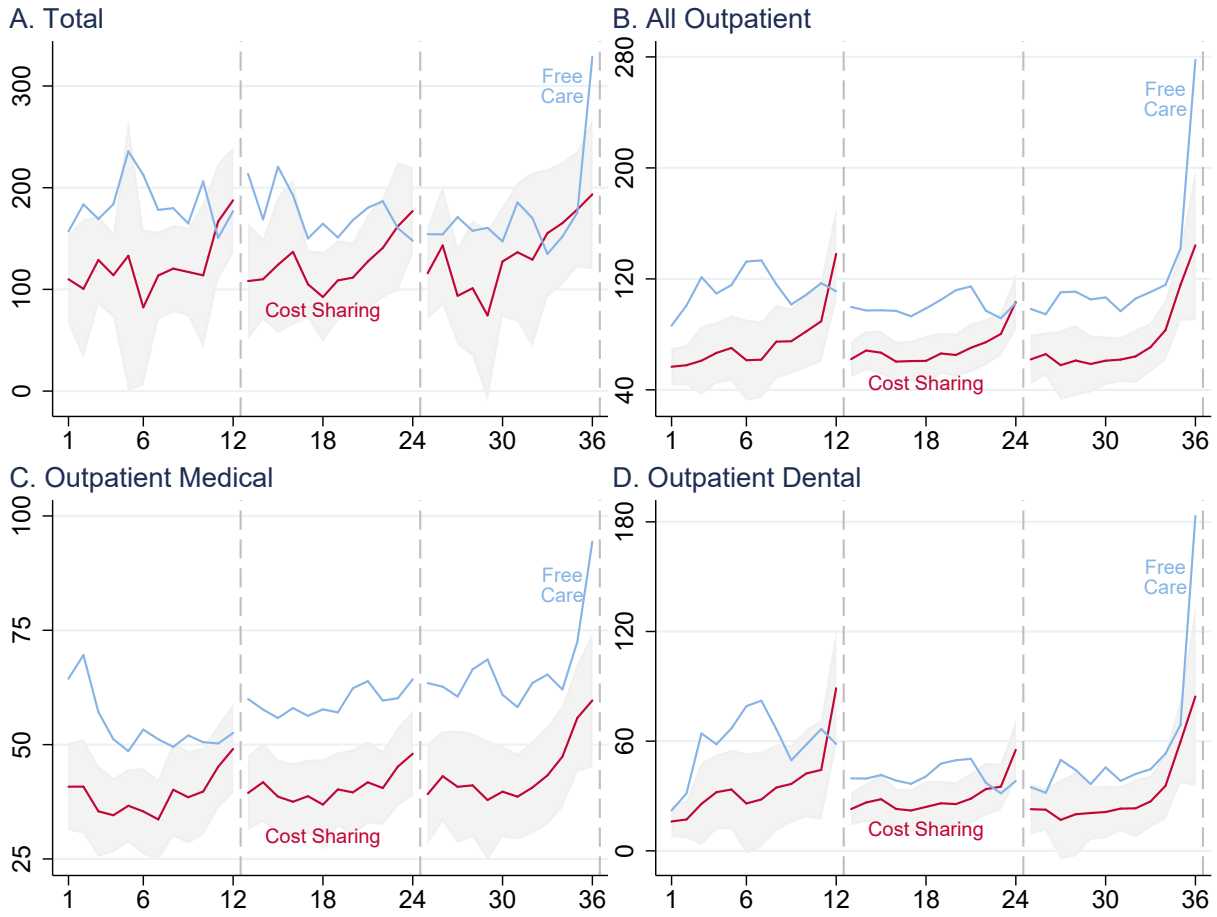
Episodes of treatment We extend our spending analysis to episodes of treatment in Figure 4. We focus on non-dental well-care, dental care, and acute episodes because these are straightforward to date.[19] Panel A shows the number of well-care episodes by plan and month. Well-care episodes in free care 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 a coverage year suggests that intertemporal substitution is an important part of demand for these categories of care, which are medically straightforward to retime. Although the lumpiness of medical care likely makes it difficult for patients to precisely retime their medical spending, these episode results show that patients do indeed initiate new bundles of care at the end of the coverage year, for deferrable care.

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 a coverage year (that are specific to the cost-sharing plan), nor by providers shifting when they date and file claims.

Results conditioning on actually hitting the MDE So far we have avoided looking

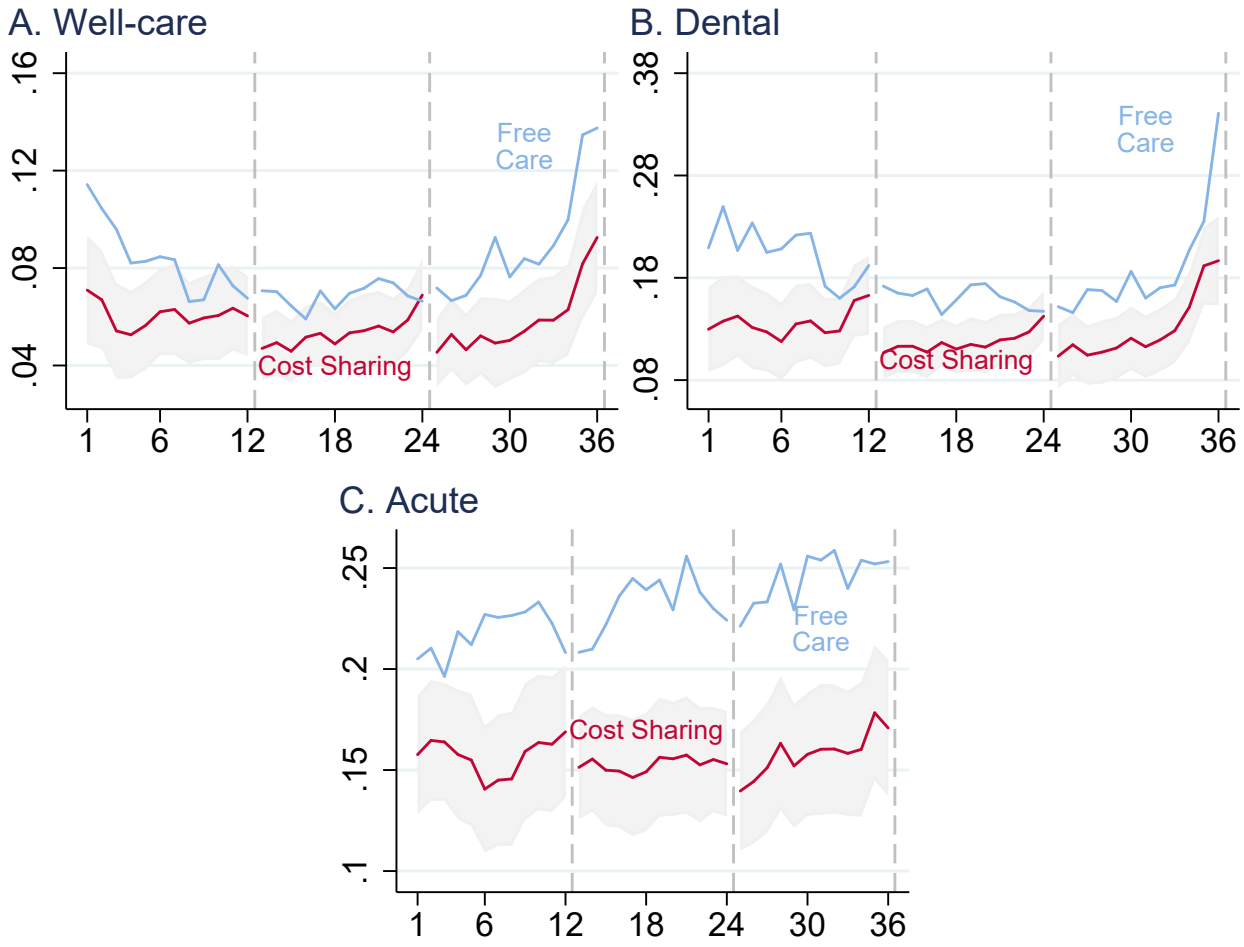
[19] The remaining categories are non-acute chronic episodes, which are persistent and do not have a clear beginning or end, and hospital episodes. These are difficult to date because they often include maternity care, and the RAND investigators dated most maternal care at the beginning of the coverage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

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



Notes: Figure shows average monthly spending per beneficiary in free care and in cost-sharing plans, 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 1. To assess both the level of spending and the statistical significance of the difference, the shaded region shows the 95% confidence interval for the *difference* in spending, but centered on the cost-sharing means.

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 and in cost-sharing plans, 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 1. To assess both the level of spending and the statistical significance of the difference, the shaded region shows the 95% confidence interval for the *difference* in spending, but centered on the cost-sharing means.

at spending or episodes conditional on hitting the MDE because hitting the MDE is mechanically related to these variables. However we show in Appendix Figure B.1 and Appendix Figure B.2 the trends in spending and episodes, broken out by hitting the MDE ever in a coverage year. Although Figure 2 shows that people hit the MDE steadily throughout a coverage year, people who hit the MDE have a surge in spending at the end of a coverage year. Spending is flat throughout a coverage year for people who do not hit the MDE. These results are of course highly consistent with intertemporal substitution.

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 a coverage 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 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 a 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.

4 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. One implication of these results is that 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

$$y_{it} = \alpha + \beta_0 price_{it} + \beta_{-1} price_{it-1} + \beta_1 price_{it+1} + X_{it}\theta + \mu_t + \theta_{sem} + \varepsilon_{it}. \quad (2)$$

Additional controls include fixed effects for site-by-start-date, and for date, and indicators for the first and last month of the experiment, collected in the vector X_{it}. In this specification, 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 $price_{it+1}$.[20]

This treatment results in two main advantages. First, since characterizing a nonlinear contract by a single price could produce misleading inferences (Aron-Dine et al. (2015)) allowing for lead and lag prices would allow us to better captures how consumers respond to a nonlinear budget set. Second, this specification allows us to separate short- and long-run price effects in a simple way. To see this, consider a one-unit price change lasting for only one period, say period t. In period t, the direct effect of such a price change is β_0. Since this coefficient measures the instantaneous impact of a price change, it is the price sensitivity we would estimate if we ignored dynamics and identified price sensitivity from a single-period, exogenous, and unanticipated price change. We therefore take β_0 as the short-run response. If there is intertemporal substitution, in addition to the instantaneous response, we should expect spending to respond to lagged and lead prices. For example, an anticipated increase in price in period $t+1$ would encourage more spending in period t (i.e., β_1 measures anticipatory response). And a decrease in price realized in period $t-1$ would likely lead to a lower spending in period t if there is any offsetting in spending (i.e., a positive β_{-1} measures the magnitude of offset response). As a result, we measure the long-run price sensitivity as $\beta_{-1} + \beta_0 + \beta_1$, because this is both the long-run effect of an anticipated price change, and the steady-state effect of a permanent one unit price change. Estimating Equation 2 let us recover short- and long-run price sensitivities. To estimate this equation we must address several challenges.

Measuring prices We measure $price_{it}, price_{it-1}$ and $price_{it+1}$ with the current, lagged, and lead spot prices. These prices take on a limited number of values: they are always zero in free care, and either zero or the sticker coinsurance amount in the cost-sharing plans. Relating demand to spot prices is consistent with the evidence in Section 3; the rising

[20] 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.

spending over the coverage year suggests that spot prices are relevant for patients.[21] 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.[22] 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. For example, in the last month of the middle year of the experiment, $price_{it}$ differs by about 0.45 between free care and cost-sharing, but $price_{it+1}$ differs by about 0.75 (see Figure 2). Loosely speaking, sensitivity to $price_{it+1}$ is therefore identified in part by the excess spending in the cost-sharing plans in this month. 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

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

[22] 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.

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 $price_{it+1}$.[23]

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.[24] 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_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 the estimation.

4.1 Identification and the first stage

Identification requires that the instruments induce linearly independent variation among spot prices, future prices, and lagged prices. The linearly independent variation in prices comes from turning over of a 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 a coverage year.

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

[23] In practice we set $\bar{p} = 0.5$. 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.

[24] 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.

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 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.

Formal tests establish the strength of our instruments. Across all specifications, we find Kleibergen and Paap (2006) F-statistics in excess of 600, and Sanderson and Windmeijer (2016) partial F-statistics of at least 500. These exceed conventional cutoffs. The Kleibergen and Paap (2006) F-statistic is a test of underidentification and the Sanderson and Windmeijer (2016) F statistics are tests of weak identification. We reject both null hypotheses, suggesting that our instruments are strong.[25]

4.2 Short and long run price sensitivities

Spending response Columns (1)-(3) of Table 3 show the estimates of Equation 2. 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 $4.94, outpatient dental by $11.69, and inpatient care by $0.18, for an overall effect of about $17 (standard error: $5).

The estimated long-run response is $7 given the same increase in the coinsurance rate (10%), which is calculated by summing up β_{-1}, β_0, and β_1 across the three categories of care. We sum up across the three categories of care because they are mutually exclusive. The long-run response is smaller than the short-run effect by about $10 (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

[25] We reject in the sense that the F-statistics exceed the conventional cutoff of 10. In fact the critical values with heteroscedastic and clustered standard errors are unknown (Sanderson and Windmeijer, 2016).

Table 3: Effect of current, past, and future prices on health care spending and episodes of care

Outcome	Spending			# Episodes				
Category	Medical	Dental	Inpatient	Well-care	Dental	Acute	Chronic	Inpatient
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Price (p_t)	-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 (p_{t-1})	-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 (p_{t+1})	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	-24.7	-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)
Mean dep. var.	47.3	41.3	54.1	0.061	0.153	0.179	0.113	0.035
Mean price	0.35	0.38	0.17	0.35	0.38	0.35	0.35	0.17

Notes: Table shows coefficients 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. The sample is defined as in the notes to Table 2 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.

is only a sixth the short-run response. The inpatient estimates are too imprecise to compare the short- and long-run spending responses.

It is interesting to note that long- and short-run price responses diverge mainly because demand rises in anticipation of future price increases. The high anticipatory spending is not later offset by lower spending after the price change is realized. That is, the coefficient on p_{t-1} is fairly small relative to the coefficient on p_t (and sometimes statistically insignificant). The coefficient on lagged price for dental care spending is significant but negative; if high past spending were offsetting, we would expect the coefficient on p_{t-1} to be positive. By contrast the coefficient on p_{t+1} is large and statistically significant, between 50 and 95 percent as large as the coefficient on p_t. We therefore conclude anticipatory effects are driving the difference in long- and short-run responses for spending.

Episode response Columns (4)-(8) show price sensitivities of episodes of care, broken down by well-care (non-dental), dental care, acute, chronic, and inpatient. Overall, a one-time price increase from zero to one 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. The coefficient on p_{t+1} is 0.042 for well-care and 0.085 for dental care, about half of the coefficient on p_t. These estimates show that episodes rise in anticipation of high future prices, consistent with our results for spending. Interestingly for well-care episodes the coefficient on p_{t-1} is 0.023, which amounts to about a quarter of the coefficient on p_t and about half the coefficient on p_{t+1} . These estimates suggest that about half of the anticipatory demand, therefore, is later offset by lower demand when price changes materialize. This offset is likely driven by retimed 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. For acute episodes, the coefficient on p_{t-1} is wrong-signed (-0.031),

meaning that acute spending episodes are not lower in periods following low 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.[26] These results suggest that most of the intertemporal substitution therefore reflects retiming of care, rather than stocking up on general health capital.

Robustness and summary of results We show in Appendix D how we address several concerns related to our main specification of Equation 2, including differential attrition among different plans, arbitrary choice of lead and lag specifications, and imputation of prices for the first and the last month of the experiment. All these results, reported in Appendix Table D.1 are largely similar to our main results. To summarize, for both spending and episodes of care, 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.

4.3 Reconciling disparate estimates of price sensitivities

We have shown that short- and long-lasting price changes can generate substantially different spending responses. To gauge the economic importance of this difference for researchers estimating price sensitivities and to reconcile disparate estimates in the existing literature, we show how price sensitivities estimated with the RAND data vary according to the source of variation used. To do so, we estimate the following regression of monthly spending or episodes of care on the spot price of care:

$$y_{it} = \beta_0 + \beta_1 price_{it} + X_{it}\theta + \mu_t + \theta_{sem} + \varepsilon_{it}. \tag{3}$$

The controls always include calendar time dummies and site-by-start-date fixed effects.

[26] 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.

Table 4: Price sensitivity estimates differ according to the type of variation used in estimation

Variation:	Across-plan (1)	Within-plan (2)	All (3)
	Panel A: Spending		
Outpatient medical	-25.7	-30.4	-26.0
	(3.3)	(6.3)	(3.1)
Dental	-25.4	-82.6	-27.8
	(3.2)	(12.3)	(3.0)
Inpatient	-21.2	-29.7	-21.7
	(10.1)	(50.4)	(9.7)
Total	-72.3	-142.6	-75.5
	(11.1)	(52.2)	(10.6)
	Panel B: Episodes		
Well-care	-0.031	-0.068	-0.033
	(0.005)	(0.011)	(0.004)
Dental	-0.078	-0.138	-0.081
	(0.007)	(0.021)	(0.007)
Acute	-0.101	-0.02	-0.098
	(0.013)	(0.017)	(0.012)
Instruments	Plan Dummies	Plan-Month Dummies	Plan-Month Dummies
Person fixed effects	No	Yes	No

Notes: Table shows the coefficient of price from a regression of the indicated outcome on price, plus a set of fixed effects for site-by-start date and calendar month. Each cell shows a different regression result. In each column, we use a different set of instruments, as indicated. Robust standard errors, clustered on family, are in parentheses.

We consider three sets of instruments and controls which help to isolate different sources of price variation: long-run, short-run, or a mix of both. To isolate long-run price variation, we instrument for $price_{it}$ using plan assignment. Given that plan assignment varies across people but not over time, these instruments therefore induce long-lasting price changes. To identify responses to short-run price variation, we instrument for $price_{it}$ using a full set of interactions between plan assignment dummies and coverage month dummies. These instruments reflect both long-lasting, cross-person price variation (coming from differences in plan assignment), and short-run, within person variation (coming from predictable within-plan, over-time price changes variation). To further isolate only the short-run differences, we add person fixed effects to the regression, so that the only variation in these instruments (conditional on the controls) is the within-person, short-run changes in spot prices occurring throughout the coverage year. In a final specification, we keep the plan-by-month dummies as instruments, but we drop the person fixed effects as controls. This specification uses all of the price variation induced by the experiment, and so the estimated responses reflect a mix of short- and long-lasting price changes.

Table 4 shows the estimates. In column (1), we isolate long-run price variation (reflecting the long-run effect of different plan assignments), and we obtain price sensitivities that are nearly identical to the long-run estimates in Table 3. In column (2) we include person-fixed effects and use within-plan, over-time price changes, we find price sensitivities that are much larger: twenty percent larger for outpatient medical spending, more than three times as large for dental spending, and twice as large for total spending. For episodes, the effects are similar: well-care and dental episodes are twice as responsive to within-person price changes. Not all categories of health care demand respond differently. Hospitalizations and acute care do not respond to temporary price changes, possible because these categories are hardest to retime. Using the full price variation induced by experiments in column (3), we obtain price sensitivities that are roughly 10 percent larger than the long-run estimates reported in Table 3. These estimates reflect a mix of short- and long-run responses. They are close to the long-run response, because most of the price variation induced by the experiment is long-run variation.

Thus, estimates of price sensitivity that neglect intertemporal substitution are sensi-

tive 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. These results help reconcile heterogeneous elasticity estimates reported in the literature; papers using short-run variation tend to find larger elasticities than papers using long-lasting variation. For example, Eichner (1998) and Kowalski (2016) identify price elasticities using an instrumental variables strategy based on family member injuries, which increase the likelihood that the family hits the deductible, and change the price of health care for uninjured family members. This is considered a short-run price shock since families with an injury are likely face low prices only within the coverage year. Eichner and Kowalski find large elasticities -0.7 or larger. By contrast, studies that use quasi-exogenous variation in plan assignment tend to find smaller elasticities. To the extent that plan assignment lasts multiple years, it is likely that these studies are picking up long-run effects. For example, the RAND HIE reported an elasticity of -0.2 (Manning et al., 1987); more recent work using structural methods and inferring moral hazard from cross-plan variation in utilization also estimates similar (albeit somewhat lower) responses (Cardon and Hendel, 2001; Bajari et al., 2014).[27]

4.4 Reconciliation with the original HIE results

The original RAND HIE investigators concluded that there was little evidence of intertemporal substitution in the RAND data (Keeler et al., 1982). We reconcile our results in Appendix E. The key to our reconciliation is that Keeler et al. (1982) test for intertemporal substitution by looking what happens just before and after people hit the MDE, rather than looking at the end of a coverage year as we do.

So how does intertemporal substitution affect the overall implications of the RAND HIE? We first discuss its implication for the estimation of long-run price elasticity. In the

[27] Not all variation in estimates can be accounted for in this way. Dalton (2014) and Kowalski (2015) use nonlinear budget set methods to estimate moral hazard, and find relatively low elasticities. Dalton (2014) identifies price sensitivity exclusively off of behavior around kink points in a single insurance plan, and finds elasticities of -0.26 to -0.09. These elasticities are driven by people whose health care demand shocks put them in the neighborhood of the kink points, so they are likely short run. Kowalski (2015) reports very small elasticities. Her estimates are identified off of both behavior around the kink points and plan choice, so they reflect a mix of long-lasting and short-lasting price variation.

RAND setting, intertemporal substitution biases the estimate of the long-run price sensitivity mainly because it leads to excess spending in free care for the first and the last month of the experiment.[28] These months represent a small share of overall spending, and therefore they do not enormously contribute to the estimated treatment effect. As a result, failing to account for intertemporal substitution makes only a small difference to the estimated price elasticity in the RAND setting.[29]

An arguably more important implication relates to predicting spending under different insurance plans in the presence of intertemporal substitution. If the goal is to quantify spending across linear plans with different generosity, a single price elasticity would suffice since price is fixed under a linear contract. However, in the case of a non-linear contract, intertemporal substitution leads to dynamic responses so characterizing a health plan using a single price could produce misleading inferences. Actually previous literature such as Aron-Dine et al. (2015) has highlighted the importance of explicitly account for the entire non-linear budget set rather than a single price. In the next section, we conduct counterfactuals to quantify how our model outperforms a static model which fails to account for dynamic response.

5 Implications for insurance design

We find failing to account for intertemporal substitution would lead to erroneous conclusions regarding how medical spending response to price changes. In particular, intertermporal substitution can work to undermine cost savings that otherwise might be achieved when switching consumers across plans, offering important policy implications. To gauge the extent to which intertemporal substitution could undo cost savings and illustrate the implications for insurance reform, we offer a series of simulations which allow us to calculate spending under different plan designs.

[28] The spike in spending in the cost-sharing plans at the end of a coverage year reflects a steady-state consequence of cost sharing, and so it should be counted towards long-run spending.

[29] To see this, compare the long-run estimates in Table 3 to the estimates in column (1) of Table 4. The estimates in Table 3 account for intertemporal substitution. The estimates in column (1) of Table 4 do not account for intertemporal substitution, but they are similar to those reported by the RAND HIE investigators, because they are identified off of cross-plan differences (e.g. Manning et al. (1987)).

Table 5: Simulated effects of alternative insurance contracts

Plan	Platinum (1)	Bronze (2)	HDHP (3)
	Monthly spending	Δ Spending vs. platinum	
Model:			
Dynamic	208.67	-8.96	-40.70
Static, identified via long-run variation	209.57	-9.37	-48.70
Static, identified via short-run variation	229.89	-21.50	-114.17

Notes: Column 1 shows average simulated spending under a platinum insurance plan with a constant 10 percent coinsurance rate. Columns (2) and (3) show the simulated change in average spending relative to the platinum plan, under a bronze plan (with a constant 37 percent coinsurance rate), or a HDHP (with a $1,250 deductible and 10 percent coinsurance above the deductible). See Appendix F for more details.

We start from a generous "platinum" insurance plan with a constant coinsurance rate of 10 percent, and we consider the effect of moving to a HDHP, with a deductible of $1,250 and 10 percent cost sharing above the deductible. This deductible is about the 90th percentile of annual spending in the RAND data; we explain our choice of these parameters further in Appendix F. As a benchmark, we also consider the effect of moving to a "bronze" plan with a constant 37 percent coinsurance rate. We chose this coinsurance rate so that the "bronze" plan is roughly equally generous as the HDHP.[30]

For the purpose of comparison, we simulate spending using the dynamic model in Equation 2, and using the static model identified from long-run price variation and from short-run price variation (the variation underlying Table 4, columns (1) and (2)).[31] Appendix F provides details of our simulation approach.

Table 5 shows spending under the platinum plan in column (1), the change in spending under the bronze plan in column (2), and the change in spending under the HDHP in column (3). We focus our discussion on changes in spending when moving across plans. We first consider moving between linear contracts (from the platinum to the bronze plan). Column (2) shows that the dynamic model and the static (long-run) model give nearly

[30] Specifically, we simulated total and out-of-pocket spending under the HDHP; out-of-pocket spending is 37 percent of total spending in the HDHP.

[31] We simulate spending for adults only, as the insurance plans all have individual deductibles which children are unlikely to meet.

identical predictions for the changes in spending. This is because intertemporal substitution is muted under a linear contract and these two models generate nearly identical long-run price elasticities. As a result, moving from one liner contract to another would result in a nearly identical effect under these two models. Also unsurprisingly, the static (short-run) model has produced much larger predictions of cost savings, consistent with the fact that the estimated short-run price elasticity is about double the size of the long-run elasticity.

By contrast, we find that accounting for dynamics is crucial when focusing on insurance contracts with nonlinear cost-sharing, which induce non-trivial dynamics in prices. When moving from the linear platinum plan to the nonlinear HDHP, the dynamic model predicts a reduced spending of about $40, while the static model predicts larger spending reductions, $48 per month for the model identified off long-run price changes and $114 per month for the model identified off short-run variation. We draw two implications here. First, relying on short-run price sensitivity tends to overestimate cost savings and the bias is exaggerated under a nonlinear contract ($114 vs $40 as compared to $21 vs $9). Second and more importantly, relying on a single price elasticity alone (such as the long-run price sensitivity) could generate biased prediction of spending under a nonlinear contract. This is because summarizing price responsiveness for a nonlinear contract requires more than a single price elasticity, as has been pointed out by Aron-Dine et al. (2015). In our simulation, the static (long-run) model tends to overstate savings from an HDHP by about 20 percent. Such bias is driven by ignoring dynamics under the static model. Our reduced-form dynamic model has the advantage of tractability, and offers an improvement and a direct comparison to a static model.

While we think our counterfactuals offer important policy implications, it is important to discuss internal and external validity. To assess internal validity, we use the dynamic model and the static model (using all variation) to simulate behavior in the RAND setting. Specifically we simulate spending in the free care and family deducible plans, with which had the largest enrollment in the data. The simulation of monthly spending follow the same procedure described above, except that when simulating behavior in a given plan, we exclude people in that plan, to improve the credibility of the fit. We find the dynamic model outperforms the static one in explaining spending patterns close to the end of a coverage

year when we find strongest evidence for intertemporal substitution. The dynamic model is also preferred by comparing model fit measured using root-mean-square errors.

Regarding external validity, there are reasons that we should worry about extrapolating the RAND results out of sample. Today's health care environment differs in fundamental ways from the one in which the RAND experiment took place. For example, managed care has become more prominent and a lot more emphasis has been put on preventive care. Additionally, spending on prescription drugs has risen rapidly for the past decades. We therefore acknowledge that the results from the RAND estimates might not directly apply to our current setting, and we need to interpret our counterfactuals with caution. In particular, it is important not to emphasize too much about the models' prediction regarding the total amount of spending. Instead, we think our prediction regarding differences in cost savings from switching plans under different model specifications, which is the focus of our earlier discussion of the counterfactuals, is more informative. Despite these caveats, we believe that intertemporal substitution is likely to play a larger role nowadays than in the early 1980s. This is because there are many more elective and deferrable procedures (e.g., imaging and diagnostic tests) available now, and a lot of those procedures could move a few months earlier or later without great harm. There has also been an increased use of IT tools for consumers to know how far they are from their deductible. These features likely make it easier to track the exact spot price and offer more opportunities for intertemporal substitution. If these effects are large, our results could be interpreted as a lower bound of the estimated cost savings undo due to intertemporal substitution.

6 Conclusion

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. Our estimates suggest that moral hazard in the short run—the response to a one time, unanticipated price change—is substantially larger than in the long-run. Neglecting intertemporal substitution can also lead to biased estimates of the long-run effect of cost-sharing on utilization.

These results have 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 in the existing literature, 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, our results 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. We illustrate this problem quantitatively by simulating the effect of moving from a platinum plan to a high deductible health plan, under dynamic or static models of health care demand. Relative to the dynamic model, the static model overstates spending reductions by roughly 20 percent. Third, our results suggest that some prices may be more salient than others. We find that, for people in cost-sharing plans, spending and utilization rise dramatically at the end of a coverage year, in anticipation of next year's higher prices. By contrast, the original RAND investigators did not find any surge in utilization in the period immediately after hitting the MDE (Keeler et al., 1982), suggesting that hitting the deductible may not be immediately salient, but the "turning over" of the coverage year is. This finding is consistent with recent findings emphasizing that consumer decision marking responds much more to salient characteristics (Chetty et al., 2009; Bordalo et al., 2012; Dalton and Zhong, 2018).

These 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 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. A recent paper by Diamond et al. (2018) provides evidence that enrollees in California's Health Insurance Marketplace strategically time their health care use and enrollment decisions, using care when covered and then dropping out. 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 the extent of intertemporal substitution 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.

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.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience Theory of Choice Under Risk," *Quarterly Journal of Economics*, 2012, *127*, 1243–1285.

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.

Chetty, Raj, Adam Looney, and Kory Kroft, "Salience and Taxation: Theory and Evidence," *American Economic Review*, 2009, *99*, 1145–1177.

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.

Dalton, Christina Marsh and Yi Zhong, "Initial Learning and Eventual Substitution: A Behavior Study in Medicare Part D," May 2018. Unpublished working paper.

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.

Diamond, Rebecca, Michael J. Dickstein, TImothy McQuade, and Petra Persson, "Take-Up, Drop-Out and Spending in ACA Marketplaces," May 2018. NBER working Paper No. 24668.

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.

Erdem, Tulin, Susumu Imai, and Michael P. Keane, "Brand and Quantity Choice Dynamics under Price Uncertainty," *Quantitative Marketing and Economics*, 2003, *1* (1), 5–64.

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.

Hartmann, Wesley R., "Intertemporal effects of consumption and their implications for demand elasticity estimates," *Quantitative Marketing and Economics*, 2006, *4* (4), 325– 329.

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.

Kaiser Family Foundation and Health Research and Educational Trust, "Employer Health Benefits, 2016 Annual Survey," 2017.

Keeler, E.B., J.P. Newhouse, and C.E. Phelps, "Deductibles and the Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule under Uncertaitny," *Econometrica*, 1977, *45* (3), 641–656.

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 Willard G. Manning, Jr., *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.

Sacks, Naomi, James F. Burgess, Howard J. Cabral, and Steven D. Pizer, "Myopic and Forward Looking Behavior in Branded Oral Anti-Diabetic Medication Consumption: An Example from Medicare Part D," *Health Economics*, 2017, *27*, 753–764.

Sanderson, Eleanor and Frank Windmeijer, "A Weak Instrument F-Test in Linear IV Models with Multiple Endogenous Regressors," *Journal of Econometrics*, 2016, *190*, 212–221.

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.

For Online Publication

A Formal model of health care demand without intertemporal substitution

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 relaxing quasilinearity 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.

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:

$$\begin{aligned} U = &\max_{h_t(\cdot)} \sum_{t=1}^{12} \beta^t E[u(h_t, c_t, \varepsilon_t)|\mathcal{I}_t] \\ &\text{such that } Y = C\left(\sum_t h_t\right) + \sum_t c_t \end{aligned}$$

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 $\tau(h, H)$ give the required (marginal) out-of-pocket payment for monthly health care spending h when total spending at the beginning of the month is H, and let T be the last period of the coverage year. That is, $\tau(h, H) = C(h+H) - C(H)$. The "spot price" is the marginal price of the next dollar of health care, $\tau'(h, H)$. For example, consider a person who has not hit the MDE by the beginning of the month and whose spending will not cause her to hit the MDE by the end of the month. Then $\tau(h, H) = coins \times h$ and $\tau'(h, H) = coins$, the spot price.

The Bellman equations are

$$V_t(H_t, \varepsilon_t) = \max_h u(h, \varepsilon_t) - \tau(h, H_t) + \beta E[V_{t+1}(H_t + h)|\varepsilon_t, H_t, h].$$
$$V_T(H_T, \varepsilon_T) = \max_h u(h, \varepsilon_T) - \tau(h, H_T)$$

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

A.1 Testable implications

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.

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

$$u'(h_T, \varepsilon_T) = \tau'(h_T, H_T).$$

At an interior solution at the end of the coverage year, people choose health spending so that the marginal out-of-pocket price, $\tau'(h, H_T)$, equals the marginal utility of healthcare dollar. Corner solutions with $h_T = 0$ are empirically common; these happen when ε_T is such that $u'(0, \varepsilon_T) < \tau'(0, H_t)$.[32]

Letting $p = \tau'(h_T, H_T)$ denote the realized end-of-year price, we can therefore write final period demand as

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

Equation A.1 says that two people who have the same end-of-year price will have the same final-period demand (holding fixed their health shocks ε), regardless of whether they face the same contract $\tau(\cdot)$. In particular, a person in a nonlinear cost-sharing plan who hits the

[32] 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.

maximum dollar expenditure will have the same final period demand as a person who is in free care all along, assuming their health is the same. In general it is likely that people in worse health are more likely to select into free care. Our application avoids this problem by using data with random plan assignment.

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 in a cost-sharing plan who have hit the MDE . 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 why, 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, $E[h(0,\varepsilon_T)|\text{cost-sharing, hit}] \neq E[h(0,\varepsilon_T)|\text{free}]$, 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, as in the example in Section 1. 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

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

This is the set of ε leading a person who faces an out-of-pocket cost function $\tau(\cdot)$ 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:

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

These expectations differ because people who hit the MDE face a different price and a different distribution of ε. 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:

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

Comparing Equation A.2 and Equation A.3 shows that expected 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.[33] Thus our main empirical test of no intertemporal substitution is a comparison of demand in free care and in cost-sharing. In the absence of intertemporal substitution, demand is lower in cost-sharing than in free care.

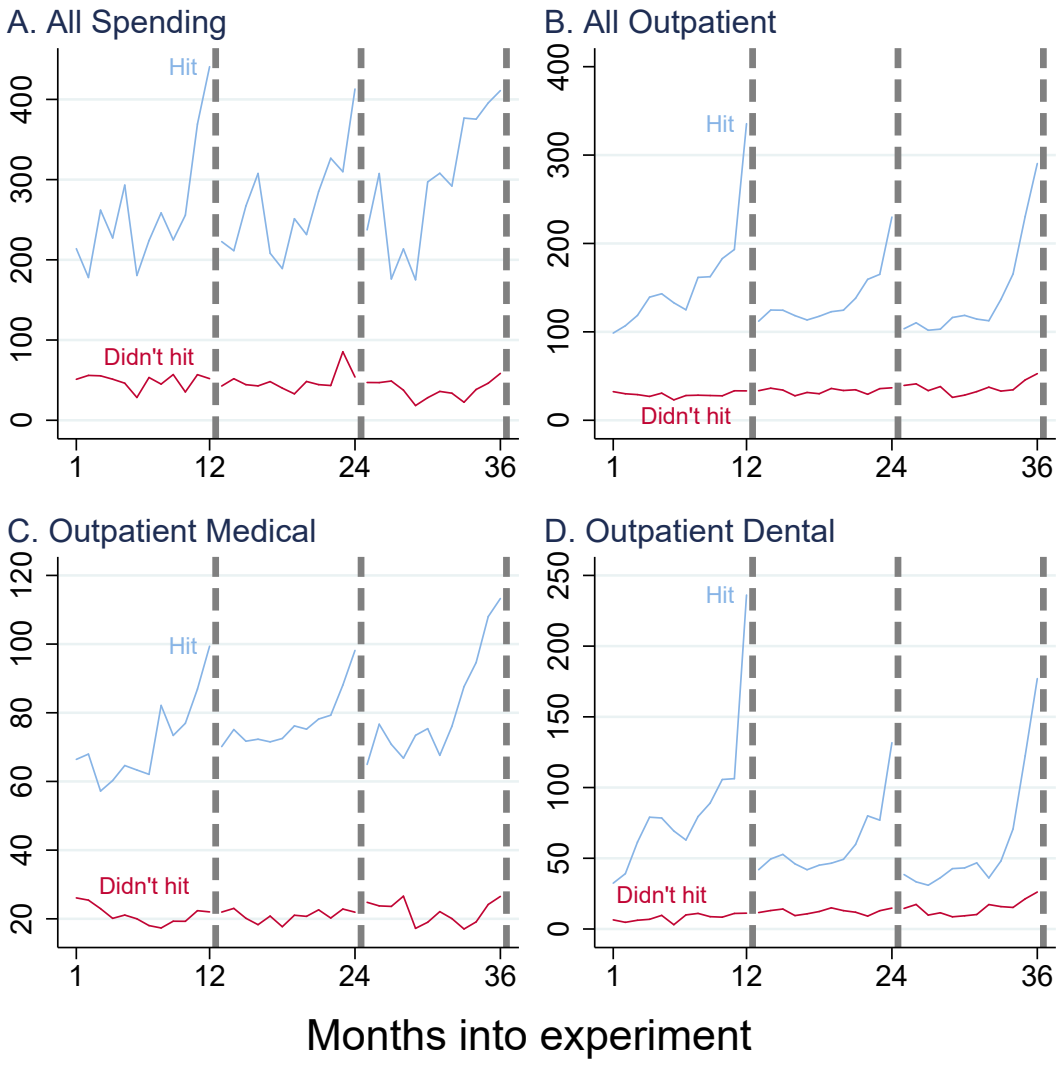
Income effects Note that by construction, a person who hits the MDE has less available income than a person in free care (exactly the MDE less). If income effects are important, then spending should be *higher* in free care than for people who hit the MDE, as health is a normal good. Precautionary savings 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.

Cross-year anticipatory effects We focus on dynamics arising from the fact that people who hit the MDE this year will face higher prices next year. However a further source of dynamics is that people who know they will not hit the MDE this year may want to retime some of their spending for next year, when they might hit. Such behavior again makes it harder for us to detect an end-of-year surge in spending in the cost-sharing group, reducing the power of our test. As we find clear evidence for intertemporal substitution, this concern only strengthens our results.

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

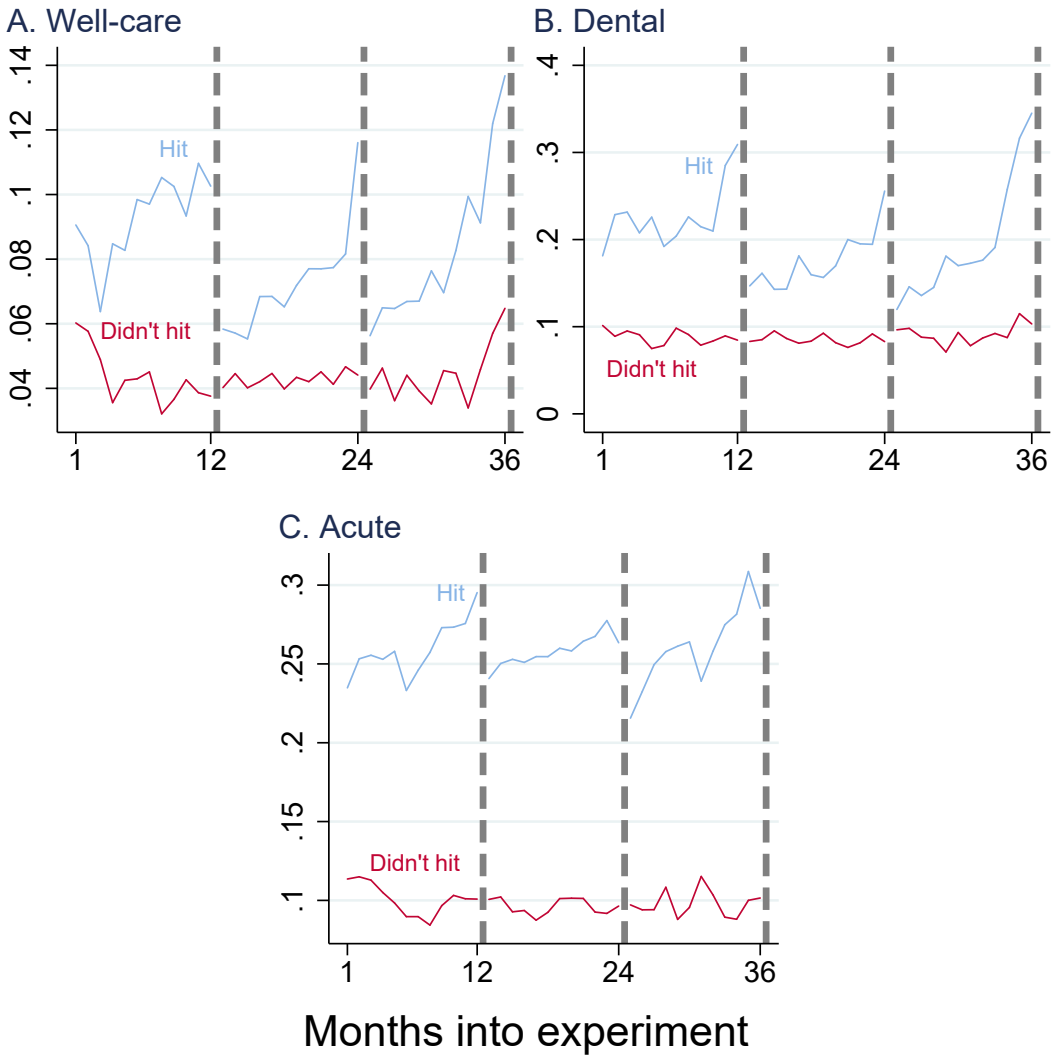
B Appendix Figures

Figure B.1: Spending by experiment month, hit MDE vs. did not hit MDE



Notes: Figure shows average spending for those who hit the MDE in a given coverage year (in gray) and for those who did not(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 1.

Figure B.2: Number of episodes of treatment by experiment month, hit MDE vs. did not hit MDE



Notes: Figure shows average number of episodes of treatment for those who hit the MDE (in gray) in a given coverage year and for those who did not (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 1.

C Balance tests

We test the validity of random plan assignment by looking at whether predetermined characteristics differ in our analysis sample. 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:

$$y_i = \sum_p \beta_p 1\left\{\text{p}lan = p\right\} + \mu_{SSD} + \varepsilon_i, \tag{C.4}$$

Here y_i 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 C.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 equal for each outcome, and at the bottom of each panel, that the coefficients are jointly equal across all outcomes. This 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 C.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 C.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 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 C.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	0.10	0.10	0.10	0.08	0.09	0.10	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	0.00
Had medical exam	0.50	0.52	0.47	0.52	0.50	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	0.20	0.18	0.15	0.17	0.19	0.20	0.43
ln (Medical spending)	3.87	3.89	3.80	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	0.70	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	0.50	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	0.50	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	0.80
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	0.10
From suburb	0.07	0.06	0.05	0.07	0.07	0.06	0.70
From town	0.23	0.23	0.22	0.29	0.23	0.24	0.22
Missing background	0.41	0.41	0.42	0.39	0.40	0.37	0.24
Income ($thousands)	9.24	9.28	9.25	9.30	9.26	9.28	0.85
Income^2	85.70	86.42	85.86	86.76	86.07	86.57	0.84
Worked	0.86	0.90	0.83	0.94	0.86	0.86	0.01
Missing work	0.01	0.01	0.01	0.00	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	0.10
Employer insurance	0.61	0.59	0.60	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	0.90
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	0.40	0.32	0.38	0.38	0.50
Missing health	0.06	0.04	0.03	0.04	0.05	0.05	0.26
Any pain	0.50	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	0.40
Any worry	0.40	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 C.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.

D Robustness of results

We show in Table D.1 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.[34] 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 2 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 our 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 highly similar, as are results from another specification (not shown) where we included six 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.

[34] These variables are listed in Appendix Table C.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.

Table D.1: 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 3; 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 three lags and leads of price. In Panel D, we drop the first and the last month of the experiment. Robust standard errors, clustered on family, are in parentheses.

E Reconciliation with the original HIE findings

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 (page 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. (1982) 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 three-month period just after hitting the MDE the "sale" period. Three month is chosen because the initial surge of demand fell to the normal rate after the first 12 weeks of the experiment for the free care plan. 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) examine pre-MDE and post-MDE periods. They find episodes rate (as compared to the free care) is largely smaller than one during post-MDE period. They therefore conclude that people are myopic or unaffected by price changes within the year. They also looked for anticipatory effects by examining whether episodes became more common just before people hit the MDE. They found no such effect, and concluded that anticipatory effects were absent, likely because people could not easily predict when they would hit the MDE.

The key difference between our analysis and the original investigators is 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, 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 myopia or limited understanding of the insurance contracts may have made it difficult for the original investigators to identify intertemporal substitution. However, by looking at the end of the coverage year, we avoid this difficulty.

F Details of simulation procedure

Our goal is to illustrate the importance of accounting for demand dynamics in simulating health care spending. These dynamics arise because, under nonlinear contracts, price is not constant. We therefore focus on the effect of moving to a nonlinear contract. In particular we simulate the effect of moving from a generous "platinum" plan to a high deductible health plan. This simulation has some practical relevance as many plans in recent years have made this switch.

Selecting plan parameters We start with a platinum plan, defined as plan with a constant 10 percent coinsurance rate for all services. We think of this as an example of a fairly generous employer-sponsored plan. Next we consider a high-deductible health plan. Exact definitions of such plans vary. We consider a plan with a deductible of $1,250, 100 percent coinsurance below the deductible, and 10 percent coinsurance above the deductible. Two considerations guided our choice of deductible. On the one hand, we wanted a deductible that could be hit by people in the RAND data; too high a deductible is simply uninsurance. On the other hand, we wanted a deductible within the range of current high deductible plans. Our deductible of $1,250 strikes a balance between these considerations; it falls just above the 90th percentile of annual spending in the middle years of the experiment, and $1,250 is the cutoff used by Coe (2014) to define HDHP. This cutoff is low relative to average deductibles. For example, among HDHPs offered by employers in 2016, the average deductible as $2,200 (Kaiser Family Foundation and Health Research and Educational Trust, 2017). We consider individual, not family deductibles, to avoid the complications of cross-family member spillovers. For comparison, we also consider the effect of moving to a less generous plan with linear cost-sharing. To cleanly isolate the effect of nonlinearity (as opposed to generosity), we considered a plan with an average coinsurance rate equal to the average coinsurance rate in the HDHP, which turns out to be 0.37. (That is, simulated out-of-pocket spending under the HDHP equals 37 percent of simulated total spending.) We refer to this as the bronze plan.

Parameterizing health care demand For the dynamic model, we simulate demand using Equation 2, which relates spending in a given month to the price of care in that month, the previous month, and the next month. In the RAND plans, different categories of care have different prices, so we estimated Equation 2 category-by-category. To keep the simulation simple, we assume here that price sensitivity does not vary by category of care. We therefore estimated Equation 2 pooling all categories of spending. The results are in Appendix Table F.1, column (1). We also simulate spending under static models (Equation 3), identified via long-run or short-run variation; estimates of the static models, pooling all spending category, are in columns (2) and (3) of Appendix Table F.1. The estimted price sensitivities of total spending are quite similar to the total spending price sensitivity implied by aggregating the category-specific sensitivities in Table 3 and Table 4.

Simulating spending under the dynamic model For a given sequence of covariates, prices, and error terms, simulated spending is trivial to obtain using Equation 2. We use the empirical distribution of covariates and error terms, which we obtain as the errors from the estimated insurance demand equations. The main difficulty is obtaining prices under a given insurance contract. In linear contracts these prices are simply the constant coinsurance rate (10 percent in the platinum plan or 37 percent in the bronze plan).

Table F.1: Effect of current, past, and future prices on health care spending, pooling all categories

Specification	Dynamic	Static Long-run variation	Static Short-run variation
	(1)	(2)	(3)
Price	-181.3	-62.4	-143.3
	(35.7)	(11.6)	(34.3)
Lag price	-8.8		
	(20.2)		
Lag price	128.9		
	(36.8)		
Instruments	Plan-Month Dummies	Plan Dummies	Plan-Month Dummies
Person fixed effects	No	No	Yes

Notes: Column (1) shows coefficients from a regression of monthly spending the spot, lag, and lead price. Columns (2) and (3) show coefficients from regressions which include only the spot 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. The sample is defined as in the notes to Table 2 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.

In nonlinear contracts, however, prices depend on the accumulated spending decisions. Given a deductible D, a price p^b below the deductible, and p^a above it, in any coverage month t, the spot price is

$$p_{it} = p^b + (p^a - p^b)1\left\{\sum_{\tau=1}^{\tau=t-1} spend_{i\tau} \geq D\right\}. \tag{F.5}$$

Since spending depends on lagged and lead prices as well as current prices, simulating spending in a given year requires some way of getting prices in the past year (for the lagged price in $t = 1$) and in the next year (for the lead price in $t = 12$).

Given any contract (D, p^b, p^a), we simulate average spending as follows. First, to get lagged prices in the first coverage month, we start with an arbitrary draw of p_0; we assume that $p_0 = p^a$ with probability $\bar{p}$ an p^b otherwise. Second, we assume that no one hits the deductible in period 1, so that $p_{i1} = p^b$. To determine when people hit the deductible, if at all, we use the following algorithm:

1. Set $\tau = 2$.

2. Assume that i hits the deductible in month τ. Update current, lag, and lead prices accordingly.

3. Find simulated spending in each month using Equation 2, using the actual values of X_{it} and $\hat{\varepsilon}_{it}$.) Let S_{it}^{τ} be accumulated predicted spending in month t assuming i hit the deductible in month τ.

4. Using Equation F.5 and S_t^{τ}, find out if person i in fact hits the deductible in month τ. If they do, stop. If they don't, increment τ by 1 and go to step 2.

This procedure finds prices, month of hitting the deductible, and spending for a given draw of p_0. In steady state, the fraction of people who hit the deductible by month 12 must equal the fraction of people in month 1 who hit the deductible last month. We therefore iterate the above algorithm until $\bar{p} = pr(p_{12} = p^a)$. In practice this happens quickly (after one iteration) because lagged prices do not matter much for spending.

Simulating spending under the static model Simulating spending under the static model is simpler, as spending does not depend on past or future prices. We simulate spending of person i in coverage month 1 using Equation 3, assuming that she has not hit the deductible in month 1. For each month $\tau > 1$, we calculate accumulated spending up through $\tau - 1$, calculate whether i has hit the deductible, and update prices accordingly.