What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics *

Zarek C. Brot-Goldberg[a]
Amitabh Chandra[b]
Benjamin R. Handel[c]
Jonathan T. Kolstad[d]

February 22, 2017

Abstract

Measuring consumer responsiveness to medical care prices is a central issue in health economics and a key ingredient in the optimal design and regulation of health insurance markets. We leverage a natural experiment at a large self-insured firm that required all of its employees to switch from an insurance plan that provided free health care to a non-linear, high deductible plan. The switch caused a spending reduction between 11.8%-13.8% of total firm-wide health spending. We decompose this spending reduction into the components of (i) consumer price shopping (ii) quantity reductions and (iii) quantity substitutions and find that spending reductions are entirely due to outright reductions in quantity. We find no evidence of consumers learning to price shop after two years in high-deductible coverage. Consumers reduce quantities across the spectrum of health care services, including potentially valuable care (e.g. preventive services) and potentially wasteful care (e.g. imaging services). To better understand these changes, we study how consumers respond to the complex structure of the high-deductible contract. Consumers respond heavily to spot prices at the time of care, reducing their spending by 42% when under the deductible, conditional on their true expected end-of-year price and their prior year end-of-year marginal price. There is no evidence of learning to respond to the true shadow price in the second year post-switch.

JEL Codes: I13, G22, D81

* *c*: Corresponding Author: handel@berkeley.edu, 521 Evans Hall, Department of Economics, University of California Berkeley, Berkeley, CA 94720, 510-643-0708. We thank Eva Lyubich and Ishita Chordia for excellent research assistance. We thank Angela Fertig, Martin Gaynor, and Gautam Gowrisankaran for insightful discussions. We thank Aaron Schwartz and co-authors for sharing their claims-based classification of low-value health care procedures. We thank seminar participants for their comments provided at AEA Annual Meetings 2016, Analysis Group, ASHE 2016, Bates White, Berkeley-NHH Industrial Organization Conference, Chicago Harris, Chicago Booth, Erasmus, European Health Econometrics Workshop, Georgia State, Harvard, Hebrew University, Microsoft Research, Lund University, MIT, NBER Insurance, NBER Health Care, North Carolina, Northwestern, Notre Dame, Ohio State, Penn State, Queens University, Southern Denmark University, Stanford, Texas A & M, UCLA, UCSD, Universidad de Los Andes and the University of British Columbia. We thank Microsoft Research for their support of this work.

I Introduction

Spending on health care services in the United States has grown rapidly over the past 50 years, increasing from 5.0% of GDP in 1960 to 17.5% in 2014 (CMS, 2015). As health care spending has risen, policymakers, large employers, and insurers have grappled with the problem of how to limit growth in health care spending without substantially reducing the quality of care consumed. One approach to addressing cost growth is to rely on demand side incentives by exposing consumers with insurance to a greater portion of the full price for health care services. Both public programs, such as Medicare and state-based insurance exchanges, and employers have moved towards a reliance on demand side incentives. For example, in 2014, 41% of consumers with employer provided coverage had individual deductibles greater than $1,000, up from 22% in 2009 (Kaiser Family Foundation, 2015a). Moreover, the share of employers offering only high-deductible coverage increased markedly from 7% in 2012 to 24% for 2016 (Towers Watson, 2015).

Assessing the appropriate combination of supply side policies, which aim to directly restrict the technologies and services consumers can access, and demand side policies depends on how consumers respond to cost-sharing. Accordingly, consumer responsiveness to medical care prices has been studied in great detail in large scale randomized control trials, notably in the RAND Health Insurance Experiment (Newhouse and the Insurance Experiment Group, 1993), the Oregon Health Insurance Experiment (Finkelstein, Taubman, Wright, Bernstein, Gruber, Newhouse, Allen, Baicker, and The Oregon Health Study Group, 2012) and, more recently, in quasi-experimental studies of high-deductible plans. The bulk of the evidence suggests higher prices reduce spending. However, there is limited evidence on precisely how these spending reductions are achieved. Consequently many employers and regulators worry that increased consumer cost-sharing is a relatively blunt instrument in the sense that (i) it may cause consumers to cut back on needed (as well as wasteful) services (Baicker, Mullainathan, and Schwartzstein (2015); Haviland, Marquis, McDevitt, and Sood (2012)) and (ii) consumers may not appropriately understand the price incentives embedded in their insurance contracts (Anastov and Baker (2014); Handel and Kolstad (2015)).

In this paper we use a new proprietary dataset from a large self-insured firm to better understand precisely how and why consumers reduce medical spending when faced with higher cost-sharing. Originally, almost all of the employees at the firm were enrolled in a generous insurance option with no cost-sharing (i.e. completely free medical care) and a broad set of providers and covered services. During and after the treatment year, which we refer to as t_0, the firm discontinued this option, moving all of its employees

enrolled in that plan into a non-linear high-deductible insurance plan that, for the population on average, paid 76% of total employee expenditures in t_0.[1] Importantly, this high-deductible plan gave access to the same providers and medical services as the prior free option leaving only variation in financial features. Additionally, employees received an up front lump sum subsidy post-switch into their Health Savings Accounts (HSA), similar in value to the population average of out-of-pocket payments in that plan.[2] With this context in mind, we observe detailed administrative data, spanning a window of six consecutive years (four years pre-switch, two years post-switch) in the time window 2006-2015, with individual-level line by line health claims providing granular information on medical spending, medical diagnoses, and patient-provider relationships. We also observe employee and dependent demographic and employment characteristics as well as the linked benefit decisions of HSA elections and 401(k) contributions. Employees at the firm are relatively high income (median income $125,000-$150,000), well-educated, and technologically savvy. In this sense, our environment presents close to a best-case scenario for the ability of consumers to (i) use technology in support of health care decisions and (ii) understand complex aspects of insurance contracts.

The required firm-wide change from free health care to high-deductible insurance constituted both a substantial increase in average employee cost-sharing and a meaningful change in the structure and complexity of that cost-sharing. We use this natural experiment, together with the detailed data described, to assess several aspects of how consumers respond to increased cost-sharing. First, we develop a time-series framework to understand how spending changed, in aggregate and for heterogeneous groups and services. In doing so, we account for both medical spending trends and consumer spending in anticipation of the required plan switch. We find that the required switch to high-deductible care caused an immediate spending reduction of between 11.1-15.4%, with the bounds reflecting a range of assumptions on anticipatory spending. Spending was reduced by 12.5% comparing t_{-1} to t_1, implying that this reduction persists in the second year post-switch. These numbers are broadly consistent with other recent work quantifying the impact of high-deductible coverage on total medical spending: see, e.g., Haviland, Eisenberg, Mehrora, Huckfeldt, and Sood (2016), Lo Sasso, Helmchen, and Kaestner (2010), and Buntin, Haviland, McDevitt, and Sood (2011) for specific examples and Cutler (2015) for a brief overview. In addition to this in-sample time-series analysis, we conduct several difference-in-differences specifications that compare spending trends in our

[1] We refer to the year of the change as t_0, the year after the change as t_1, and the years before as t_{-1}, t_{-2}, etc. To preserve the anonymity of the firm, we cannot give an exact employee count, but can note that the total number of employees (employees plus dependents) is larger than 35,000 (105,000).

[2] These funds are similar in spirit to a straight income transfer that compensates employees, on average, for these increased out-of-pocket payments. This transfer mirrors the experimental design used to address income effects in the RAND HIE (Newhouse and the Insurance Experiment Group, 1993).

primary sample to those of two potential control groups. Both specifications find results that are similar to our time-series results.

Our primary goal is to understand the mechanisms behind these spending reductions, including both how and why they occur. To investigate how consumers reduce spending, we leverage the granular data on medical procedures and patient-provider relationships to decompose the total reduction in medical spending into (i) price shopping for cheaper providers (ii) outright quantity reductions and (iii) quantity substitutions to lower-cost procedures. We perform this analysis in the spirit of Oaxaca (1973) and Blinder (1973), and also control for supply-side price responses. In this mutually exclusive and exhaustive decomposition of prices and quantities, our price shopping measure accounts for **within-procedure** shifts down the distribution of prices, while our quantity substitution measures accounts for shifts across types of procedures.

From a policy standpoint, understanding whether spending reductions are achieved through prices versus quantities is crucial. A primary argument for HDHPs is that, given appropriate financial incentives, consumers will price shop, i.e. search for cheaper providers offering a given service without compromising much on quality [Lieber (2015) and Bundorf (2012)]. In turn, providers may lower prices to reflect increasing consumer price sensitivity. Whether or not price shopping actually occurs is an empirical question that depends upon a range of factors, including consumers' provider preferences, information about prices, and search effort.[3] While enhanced consumer price shopping is almost always thought of as an efficient way to achieve spending reductions, recent research suggests that quantity reductions or substitutions may be positive or negative for welfare, depending on exactly how they occur (Baicker, Mullainathan, and Schwartzstein (2015); Chandra, Gruber, and McKnight (2007)). A model with rational and fully-informed consumers predicts that all quantity reductions are welfare improving, since consumers would value the foregone care at less than the total cost. Conversely, if consumers lack information or face other constraints, they may reduce valuable services as well as wasteful services, potentially leading to a net welfare loss.

We find no evidence of price shopping in the first year post switch. We find no evidence of an increase in price shopping in the second year post-switch; consumers are not learning to shop based on price. Instead, we find that essentially all spending reductions between t_{-1} and t_0 are achieved through outright quantity reductions (-17.9%) whereby consumers receive less medical care. These quantity reductions persist over

[3] In our setting consumers were provided a comprehensive price shopping tool that allowed them to search for doctors providing particular services by price as well as other features (e.g. location). Recent work by Lieber (2015) and Whaley (2015) finds that most consumers do not actively engage with price shopping platforms similar to the current state-of-the-art but that those who do substitute to cheaper providers for the services they search for. In a mid-t_0 survey we implemented at our firm, we find that approximately 33% of consumers have heard of the price shopping tool, 22% have logged in at least once, and 4% characterize themselves as active users.

time. Consumer substitutions across types of care plays a limited role in reduced spending (-2.2%) from t_{-1} to t_0. These results occur in the context of consistent (and low) provider price changes over the whole sample period. Importantly, the results of this decomposition are almost identical for the sickest quartile of the population, categorized using ex ante diagnoses and a well-known predictive health algorithm. For these sicker consumers, it is especially interesting to understand exactly what services they abandon, and why they choose to do so when they can readily expect to pass the deductible during the year.

Given that consumer quantity reductions are the key to total spending reductions in our setting, we next investigate service-specific reductions to shed more light on the types of care consumers forego. Our first approach decomposes the spending changes for each of the top 30 procedures by total spending across each two-year pair. Consumers reduce quantities across the board rather than targeting specific kinds of services. There is no similarly distinct change for price shopping or provider price changes across these procedures. Our second approach seeks to specifically classify services into those that are likely to be low-value versus those likely to be high value. For low-value care we follow Schwartz, Landon, Elshaug, Chernew, and McWilliams (2014), who synthesize clinical recommendations from national medical agencies to define a specific set of undesirable treatments. For high-value care, we focus on preventive care, mental health care, physical therapy, and drugs for diabetes, cholesterol, depression, and hypertension. All of the results for low and high value care mark large departures from pre and post period trends and suggest that consumers meaningfully reduce both types of care, calling into question whether quantity reductions overall are net welfare increasing or decreasing.

These findings help motivate the last major part of our analysis, which seeks to better understand why consumers who are predictably sick and well-off reduce spending during the year, despite the fact that their true shadow price of care should be close to zero in the HDHP. A range of recent evidence across different contexts with non-linear contracts suggests that, instead of responding to the true shadow price implied by a contract, consumers often respond to simpler to understand prices such as the *spot prices* paid for current purchases or their prior contract period's final marginal price.[4] If consumers respond to their spot prices, which are always weakly higher than their true shadow prices in the HDHP contract throughout the year, then they will under-consume care relative to what a fully rational dynamically optimizing consumer would do, potentially explaining our observed spending reductions.

[4]See, e.g., Einav, Finkelstein, and Schrimpf (2015), Dalton, Gowrisankaran, and Town (2015) and Abaluck, Gruber, and Swanson (2015) in Medicare Part D, Aron-Dine, Einav, Finkelstein, and Cullen (2015) in a large employer health insurance context, Ito (2014) in electricity markets, Nevo, Turner, and Williams (2016) in broadband markets, and Grubb and Osborne (2015) in cellular phone markets.

Our data and setting provides a unique opportunity to understand how consumers respond to non-linear contracts because we observe a large population of consumers who are required to move from completely free health care to the non-linear, high-deductible contract with different, potentially complex, price signals. We perform descriptive and regression analyses that shed light on which contract price signals consumers respond to. We model three high-deductible contract price signals for each family in each month: (i) the spot price, or price paid when seeking care (ii) a consumer's end-of-year marginal price from the prior year and (iii) a consumer's true shadow price of care, i.e. their expected end-of-year marginal price. Given these price signals, we compare incremental spending at different points in the calendar year for consumers in t_0 and t_1 to that of equivalent matched consumers at the same points in time during the years prior to t_0. We match consumers in the post-period and pre-period using a quantile-based approach that conditions on ex ante health status, demographics, and year-to-date spending.

Strikingly, we find that nearly all incremental spending reductions in high-deductible care are achieved in months where consumers began those months under the deductible (90% or larger in t_0 and t_1). When we condition on consumers' true shadow prices, we continue to find that consumers substantially reduce spending when under the deductible. 25% of all reductions come from the sickest quartile of consumers in months that they begin under the deductible, with 49% coming from the sickest half of consumers when they are pre-deductible. This is true even though, throughout the year, the sickest quartile of consumers can expect to pass the deductible with near certainty and the out-of-pocket maximum in many cases. We find no evidence that consumers learn to respond to their shadow price in the second-year post-switch. We discuss potential mechanisms for this spot price bias, including myopia, limited information, and liquidity constraints.

We bring these pieces together in a regression analysis that, in addition to controlling for our three price measures, also controls for spending persistence, demographics, and health status in a granular manner. We find results the mirror our descriptive analysis: consumers reduce spending when under the deductible by 42.2%, conditional on other price measures, relative to similar consumers in pre-period years. While we find no evidence that consumers respond more heavily to shadow prices, or less heavily to spot prices, in the second year post-switch, we do find evidence that consumers more heavily respond to their prior year end-of-year marginal price in t_1. This suggests that consumers may learn to respond to their end-of-year prices, but may do so based on what happened in the previous year, rather than forming new expectations for the current year.

The rest of the paper proceeds as follows. Section 2 describes our empirical setting and data. Section 3 presents our treatment effect analysis of the overall medical spending response to the required HDHP switch. Section 4 presents our decomposition of these spending reductions into (i) consumer price shopping (ii) consumer quantity reductions and (iii) consumer quantity substitutions and studies behavior for a range of services and consumer types. Section 5 presents our analysis of consumers responding to non-linear contract prices, and Section 6 concludes.

II Data and Setting

We analyze administrative data from a large self-insured firm over six consecutive years during the time window between 2006 and 2015. These six years include the year the policy took effect, which we denote t_0, the next year after, which we denote t_1, and the four years prior, which we denote t_{-4} through t_{-1}. Our dataset includes three major components. First, we observe each individual's enrollment in a health insurance plan for each month over the course of these six years, including their choice of plan and level of coverage. Second, we observe the universe of line-item health care claims incurred by all employees and their dependents, including the total payment made both by the insurer and the employee as well as detailed codes indicating the diagnosis, procedure, and service location associated with the claim. In the course of our analysis, we use these detailed medical data together with the Johns Hopkins ACG software to measure predicted health status for the upcoming year.[5] Finally, we observe rich demographic data, encompassing not only standard demographics such as age and gender, but also detailed job characteristics and income, as well as the employee's participation in and contributions to health savings accounts (HSA), flexible spending accounts (FSA), and 401(k) savings vehicles. These data are similar in content to other detailed data sets used recently in the health insurance literature, such as those in, e.g., Einav, Finkelstein, and Cullen (2010), Einav, Finkelstein, Ryan, Schrimpf, and Cullen (2013), Handel (2013), or Carlin and Town (2009). The data we use here have a particular advantage for studying moral hazard in health care utilization due to a policy change that occurred during our sample period, which we discuss in detail below.

[TABLE I ABOUT HERE]

[5]This score reflects the type of diagnoses that an individual had in the past year, along with their age and gender, rather than relying on past expenditures alone. See e.g. Handel (2013), Handel and Kolstad (2015) or Carlin and Town (2009) for a more in depth explanation of predictive ACG measures and their use in economics research. See http://acg.jhsph.org/index.php/the-acg-system-advantage/predictive-models for further technical details.

The first column of Table I presents summary statistics for the entire sample of employees and dependents enrolled in insurance at the firm. Though we cannot reveal the precise number of overall employees, to preserve firm anonymity, we can say that the number of employees is between 35,000-60,000 and the total number of employees and dependents is between 105,000-200,000. 51.2% of all employees and dependents are male, and employees are high income (91.7% $\geq$ \$100,000 per year) relative to the general population. The employees are relatively young (12.0% $\leq$ 29 years, 83.2% between 30 and 54), though we have substantial coverage of the age range 0-65 once dependents are taken into account. 23.5% of employees have insurance that only covers themselves, 20.0% cover one dependent and 56.5% cover two or more. Mean total medical expenditures (including payments by the insurer and the employee) for an individual in the plan (an employee or their dependent) were \$5,020 in t_{-1}.[6]

While the sample of employees and dependents differs from the U.S. population as a whole, it is at least partially representative of other large firms nationwide, many of which are in the process of transitioning their health benefits programs in similar manners [see Towers Watson (2015)]. Employees at the firm are relatively high income, and are almost exclusively college educated and technologically-savvy. The majority of employees live in or near a major urban area, implying they have access to a wide range of medical providers. These employees represent close to best-case scenario in terms of (i) ability to use technology to shop for care (ii) ability to pay for necessary health care and (iii) ability to understand and respond to complex non-linear insurance contracts.

Policy Change. From t_{-4} through t_{-1}, employees at the firm had two primary insurance options. Table II lists features of the two plans, side by side. The first was a popular broad network PPO plan with unusually generous first-dollar coverage. This plan had no up front premium and no employee cost-sharing for in-network medical services. The second primary option was a high-deductible health plan (HDHP) with the same broad network of providers and same covered services as the PPO. Enrollees in this plan face cost-sharing for medical expenditures, with a deductible, coinsurance arm, and out-of-pocket maximum typical of more generous high-deductible health plans. Despite higher cost sharing, this plan was potentially attractive relative to the PPO because it offered a substantial subsidy to enrollees that was directly deposited into their health savings account that was directly linked to the HDHP. As shown in table I, in t_{-1}, 85.2%

[6]These statistics include permanent (non part-time) employees enrolled in the primary insurance options (PPO or HDHP) the firm offers at t_{-1}. It excludes (i) employees enrolled in an HMO option available in select locations and (ii) employees who decline insurance: these groups total approximately 5% of all employees, stable over time.

of employees (corresponding to 94.3% of firm-wide medical spending) chose the PPO with the remainder choosing the HDHP. Regarding employee plan choice in the pre-period, for this paper it is only important to note that the large majority of employees were enrolled in the PPO prior to the required plan switch that occurred at the firm for t_0.

[TABLE II ABOUT HERE]

In year t_{-3}, the firm announced to its employees that it would discontinue the PPO option as of t_0. This required the vast majority of employees and dependents, who were still enrolled in the PPO in t_{-1}, to switch to the HDHP option for t_0. For these employees, this policy change represented a substantial and exogenous change to the marginal prices they faced for health care services.[7] Moreover, because of the PPO plan structure, the employees that were required to switch into the HDHP had a zero marginal price for medical care prior to the switch, implying that we observe true cost-free demand for health care services as our baseline.[8] The required shift from free care to the HDHP also presents a natural experiment that introduces within-year price dynamics. We explore the nuances of employee responses to these different potential perceived prices in Section V.

Primary Sample. For the majority of our forthcoming analysis, we use the sample of employees who (i) were present at the firm for the whole six years of the sample period (t_{-4} through t_1) and (ii) were enrolled in the PPO prior to the required switch in t_{-1}. We use this sample to ensure that we have a substantial time series of information on the health status of employees we analyze. Column 3 of Table I shows the summary statistics for this primary sample, which can be compared to the full sample of employees present in t_{-1} presented in Column 1. There are 22,719 employees in the primary sample covering 76,759 dependents (approximately 50% of employees and dependents present in the t_{-1} full sample in Column 1). Relative to all employees present, primary sample employees have similar distributions of age and gender, are slightly higher income, and cover slightly more dependents. Taking employees and dependents together, the primary sample and entire firm have similar distributions of age and gender, while those in the primary sample have about 4% higher medical spending on average. Table A1 in Online Appendix A1 presents sum-

[7] Table A22 in Online Appendix A10 presents statistics related to the cost-sharing change faced by the 76,759 employees and dependents in our primary sample (described below) required to move into the HDHP in t_0.

[8] As noted in Table II, there is some very limited cost-sharing for out-of-network providers in the PPO. Since the network is quite comprehensive, in a given year, approximately 5% of consumers consume any care out-of-network, 2.5% of total medical spending is out-of-network, and of this spending almost 100% is paid for by insurance. Since it is so small in magnitude, we don't consider this out-of-network spending in the remainder of the paper.

mary statistics for an alternative sample that includes all employees and dependents present from $t_{-2} - t_0$ who are in the PPO for t_{-2} and t_{-1}. Our main results are essentially unchanged for this alternative sample.

Figure A1 in Online Appendix A1 examines whether there is substantial incremental attrition from the firm after the announcement of the switch to the HDHP (later in year t_{-3}) or after the actual required switch to that plan in t_0. Reassuringly, the figure shows that there is no meaningful change in employee exit at these key points in time, or any other point during our study period. There is some incremental dependent attrition at the implementation date (1 percentage point higher than baseline), but not enough to meaningfully impact our main results. See Online Appendix A1 for additional detail.

III Impact of Cost-Sharing on Spending

We first investigate the impact of the required switch of consumers to the high-deductible plan on total medical spending. We present a series of analyses for our primary sample, including a within-sample time-series analysis and difference-in-differences analyses that compare these time-series patterns to those of relevant comparison groups, both internal and external to the firm.

The left panel in Figure I plots mean monthly spending at the individual level for our primary sample over the six years in our data (Figure A19 in Online Appendix A12 plots median spending over time to remove the effects of very high cost consumers, with similar results). The vertical line in the figure represents December of t_{-1}. The figure clearly illustrates that spending drops after the required switch to the HDHP: the average yearly spending for an individual dropped from \$5222.60 in t_{-1} to \$4446.08 in t_0, a 14.9% drop. Table III presents the year-on-year mean total spending changes over the six years, revealing a sharp break in trend for spending in t_0 relative to prior years and future years.

[FIGURE I ABOUT HERE]

As is typical in health care, the raw spending data show total medical spending increasing steadily over time. We attribute this to two factors. First, our primary sample is a balanced panel where consumers age over the six year period. Second, the price of care typically rises over time due to both price inflation and other factors such as the introduction of new medical technologies. If we fail to account for these factors, we will understate the true impact of the required HDHP switch on medical spending because t_0 spending will be mechanically larger than t_{-1} spending.

To adjust spending for age, we take monthly individual-level spending for January of year t_{-4} and regress it on age and a number of other controls. Within our sample, mean monthly spending increases by \$7.50 for each year someone ages indicating a small effect of aging on the $t_{-1} - t_0$ treatment effect estimates.[9] Additionally, we adjust for medical price inflation using the Consumer Price Index (CPI) for medical care for each month in our sample.[10] This index adjusts for price inflation, but not price increases from technological change, and as a result this adjustment may understate the impact of the required switch to the HDHP on spending reductions. In this section we intentionally use this broader price inflation index so that any equilibrium price effects as a result of the required HDHP switch are still accounted for in our treatment effect estimates, an issue we return to in Section IV.

[TABLE III ABOUT HERE]

The left panel of Figure I also presents the raw spending data adjusted for in-sample aging over time and for medical price inflation. We express the adjusted spending values in January t_{-4} dollars, i.e. in terms of ages and medical prices at year t_{-4}. The figure clearly illustrates the drop in average monthly individual spending following the required HDHP switch. The numbers in Table III show that, once these adjustments are accounted for, average individual spending drops by 18.4% from t_{-1} to t_0. Adjusted spending drops by 15.9% comparing t_{-1} to t_1, implying that the impact of high-deductible insurance on medical spending persists for both years post-switch. We use a block bootstrap method, described in more detail in Online Appendix A3, to compute the standard errors for all of the estimates presented in this section.

The right panel in Figure I investigates the impact of the switch to high-deductible health care as a function of consumer health status. The figure plots spending over time by consumer health status, categorized into quartiles using the ACG predictive index described Section II. Consumers in the sickest quartile are those who, at the beginning of each calendar year, based on the last year of medical diagnoses and spending, are predicted to spend the most for the upcoming calendar year (while the healthiest quartile are those predicted to spend the least).[11]

[9] The relative youthfulness of our sample is a key reason for the low estimated impact of aging: using nonlinear specifications gives similar results.

[10] This comes from the index collected by the Bureau of Labor Statistics. A time series of this index can be found at `http://research.stlouisfed.org/fred2/series/CPIMEDNS` and an index description at `http://www.bls.gov/cpi/cpifact4.htm`.

[11] One key difference between this figure and prior figures in this section is that the sample in each group can switch from year to year: consumers in the top quartile line for t_{-1} are those predicted to be the sickest for t_{-1}, who might not be the same predicted sickest 25% of consumers for t_0. It is crucial to construct the figure this way (rather than fixing health status at a given point in time) to avoid reversion to the mean that occurs when categorizing health at one point in time.

The figure clearly shows that health spending is reduced for the sickest three quartiles, and that the majority of the spending reductions we document come from the sickest quartile of consumers, predicted on an ex ante basis. This is striking for several reasons. First, as we will document in Section V, all of the consumers in the sickest quartile are expected to spend well past the deductible and many of these consumers can expect to pass the out-of-pocket maximum. This implies that the true price change these consumers should expect to face is quite low. Second, because these consumers are predicted ex ante to be in the sickest group, many of them have chronic medical conditions where medical care may have especially high value. In the next section, we show that these consumers reduce consumption of a broad range of medical services, including some that are likely to be wasteful and others that are likely to be of high value.

Anticipatory Spending. While it is clear from Figure I that aggregate spending decreases when the HDHP is introduced in t_0, it is also apparent that consumer spending ramps up at the end of t_{-1} in anticipation of the required plan shift. As discussed in Section II, the t_0 HDHP switch was first announced in October t_{-3} with many regular subsequent related announcements leading up to the actual change in t_0. As a result, the plan switch was a well known and salient event throughout t_{-1}, leading to anticipatory spending by consumers before the switch actually occurred, when health care spending was cheaper. This kind of anticipatory spending is clearly documented in Einav, Finkelstein, and Schrimpf (2015) in the context of Medicare Part D prescription drug insurance and Cabral (2013) in the context of dental insurance.

In our context, quantifying the extent of anticipatory spending is important for obtaining a true impact of the required HDHP shift. Without understanding the extent of such spending our estimates would overstate the true impact of the increase in cost sharing on medical spending since some of the spending that would have occurred in a normal HDHP year would have been shifted to the end of t_{-1}. To quantify excess spending in the second half of the year t_{-1}.We estimate the following specification to predict mean monthly spending:

$$\bar{y}_m = \alpha + \beta m + \lambda_M + \bar{\epsilon}_m$$

We estimate the regression on data from January t_{-4} to December t_{-2}, well in advance of the HDHP switch.[12] m denotes one of the specific 36 months over this timeframe, while M denotes a given month in

[12]It is also possible that some anticipatory spending occurs prior to the second half of t_{-1}. Such spending is highly unlikely to matter for our analysis, since consumers would have to be substituting medical care over six months forward. Figures A2 and A19

the calendar year. $\bar{y}_m$ is mean individual-level spending in our primary sample at the firm in a given month m, β is a linear time trend to account for inflation and aging, λ_M is a calendar month fixed effect to adjust for seasonality, and $\bar{\epsilon}$ is the population level idiosyncratic monthly shock to mean spending.

We determine which months have meaningful anticipatory spending by looking at the months at the end of t_{-1} that have $\bar{y}_m$ that is statistically larger than the predicted value $\widehat{\bar{y}_m}$ from the regression. Online Appendix A2 presents this analysis in detail, and shows that there is clear evidence of excess spending mass in October-December t_{-1} but not prior. Given this, we compute t_{-1} mean excess spending mass as $\Sigma_{t=10}^{12}[\widehat{\bar{y}_m} - \bar{y}_m]$. Predicted mean excess mass for October is \$37.82, for November is \$41.57, and for December is \$85.83, totaling \$165.23 per individual. The 95% confidence interval for this three-month excess mass estimate is [\$113.96, \$216.50], equivalent to 2.6% to 5.0% of mean age and CPI adjusted individual spending in t_{-1}.

To integrate this excess mass estimate into our treatment effect analysis, we need to assess how much would have been spent in t_0 under the HDHP. It is possible that some of the anticipatory spending would not have occurred at all in t_0 once prices were raised and the end of the year in t_{-1} was the final chance for consumers to consume services of low marginal value. Though it seems from Figures I and A2 that most of this excess spending would have occurred in January and February of t_0 if it occurred at all, it is difficult to credibly estimate 'missing mass' in January and February of t_0 with only two years of post-treatment data. Consequently, we allow for the percentage of anticipatory spending that would have been spent in t_0 to vary over the entire range of possible values, from 0% to 100%, and use this approach to bound the treatment effect. Throughout, we assume that any care substituted back into t_{-1} came from t_0, and not afterwards. As a result, no adjustments are required for t_1 as long as population spending is in yearly steady state.

The third column of Table III presents our range of estimates that incorporate anticipatory spending into our time-series analysis. We find that the switch to the HDHP in t_0 decreased total spending by between 11.1% (all anticipatory spending would have been spent in t_0) and 15.1% (no anticipatory spending would have been spent in t_0). The difference between this range, and our 18.4% estimate where anticipatory spending is not accounted for, indicates the importance of measuring such spending when using a pre-post or difference-in-differences design to assess the impact of cost-sharing on health care spending. Under this framework, t_1 spending is reduced by 12.5% relative to t_{-1}. Table III also presents this percentage change in spending as a semi-arc elasticity, for comparison to prior work that reports this statistic as a measure of

in Online Appendix A2 clearly illustrate that claim counts and median monthly spending spike in October-December t_{-1}, but not earlier in t_{-1}.

price responsiveness.[13] The three semi-arc elasticity estimates in Table III range from -0.57 to -0.85, or from about one-quarter to one-third of the RAND study estimates described in Keeler and Rolph (1988). We note that the economic implications of our treatment effect estimates are still substantial while there are many potentially important differences between our setting and the RAND setting. See Online Appendix A4 for more detail on these elasticities and related comparisons.

Early Switcher Difference-In-Differences. In addition to this primary sample time-series analysis, we present three difference-in-differences analyses. The primary purposes of these analyses are to (i) form relevant control groups for our primary sample time-series analyses and (ii) explore the external validity of our time-series results.

The first control group we use are "early switchers," the 15% of consumers who switched to the HDHP in years prior to the required switch at t_0. These consumers are not an exogenous comparison group, since they selected to join the HDHP in t_{-2} (6,225 individuals) and t_{-1} (5,528 individuals).[14] This is clearly seen in the left panel of Figure II, which plots spending for early switchers vs. our primary sample over time, revealing that early switchers spend less than our primary sample on average. We form a weighted early switcher sample that matches early switchers to our primary sample based on health status. We use predictive ACG scores constructed for the beginning of year t_{-1} to weight the early switcher sample, so that their health status distribution is equivalent to that of the primary sample at that point in time. We implement this matching at a granular level, based on ACG score ventiles: see Online Appendix A5 for more details.

[FIGURE II ABOUT HERE]

The difference-in-differences specification compares primary sample spending over the two year period spanning t_{-1} - t_0 to weighted early switcher sample spending. The first column of Table IV presents these estimates, which are bounded between an 11.3% and 15.2% reduction. This range is, reassuringly, quite similar to our primary sample time-series estimate presented in Table III. The lower end of this range is statistically different from a 0% change at the 10% level: the standard errors for this specification is

[13] As discussed in Aron-Dine, Einav, and Finkelstein (2013) and shown in this paper in Section V, describing a non-linear insurance contract by one price for an entire population is a strong oversimplification. We note that while most of the literature uses arc elasticity rather than semi-arc elasticity, when the price change in question starts from zero price, as in our setting, arc elasticity just represents the percent change in quantity irrespective of the price change, and so is not a satisfactory descriptive statistic for price responsiveness. The semi-arc elasticity we report is $\frac{(q_2-q_1)/(q_2+q_1)}{(p_2-p_1)/2}$ while the oft-reported arc-elasticity is $\frac{(q_2-q_1)/(q_2+q_1)}{(p_2-p_1)/(p_2+p_1)}$.

[14] We restrict the early switcher sample to consumers present for all six years, t_{-4} to t_1, similar to our primary sample. As with the primary sample, robustness checks that relax this balanced panel restriction yield similar results.

higher than for the other presented in this section because of the relatively small size of the early switcher sample (approx. 12,000). See Online Appendix A5 for additional figures and details on this early switcher difference-in-differences specification.

[TABLE IV ABOUT HERE]

Truven Control Difference-in-Differences. It is useful to have a broader comparison group for our primary sample time-series analysis, to ensure that there were not specific regional spending trends over the time period t_{-1} to t_0 that impact our time-series results. Though our CPI adjustments are a useful first pass, a more comprehensive and targeted comparison is warranted.

To this end, we use Truven Analytic's MarketScan Data, a nationally representative individual-level database of medical claims across the spectrum of private insurers.[15] We obtained the Truven data for the two years t_{-1} and t_0. We form a comparison group for our primary sample over these two years in several steps. First, we restrict the Truven sample to consumers receiving care in the state where the firm we study employs most (approximately 75%) of its employees. Second, we restrict the Truven sample to consumers with private health insurance (i.e. not Medicare or Medicaid). With these restrictions, we observe roughly 600,000 consumers' medical spending and claims each year in the Truven data.

To form a more precise comparison group, we weight the Truven sample so that it reflects the exact age and gender profile of our primary sample.[16] With this weighted Truven sample, we then perform a difference-in-differences analysis similar to that done with the early switcher sample.

The right panel in Figure II presents mean spending over time for our primary sample and for the weighted Truven comparison group. First, we note that, even weighted for age, gender, and location, mean spending in the weighted Truven sample is about half of that in our primary sample. This is likely due to a number of factors, including that the Truven group includes consumers in less generous financial plans and less generous plans in terms of provider access (e.g. HMOs) on average. Additionally, the Truven sample is, on average, likely to be lower income than the consumers we study. With this in mind, the figure shows an upward trend in spending over time moving from year t_{-1} to t_0, as compared to the sharp downward break in spending observed in our primary sample. The final column in Table IV quantifies the relative

[15] This dataset has been used in past studies to look at trends in healthcare markets, such as in Baker, Bundorf, and Kessler (2015) and Ellis, Jiang, and Manning (2015). We describe it in more detail in Online Appendix A6.

[16] See Online Appendix A6 for more details on this weighting procedure. The Online Appendix contains an additional exercise that weights the Truven sample by income as well as by age and gender. We include this in the Online Appendix, rather than the main text, because income data are only available for approximately 7% of the overall Truven sample we use. The results with those income weights are similar, though less statistically precise.

spending reduction in our primary sample, which is bounded between -22.6% and -26.6%. The increase in spending over time in the weighted Truven sample is larger than the coarse estimate from the Bureau of Labor Statistics used in our earlier adjustment, leading to a larger percentage reduction in spending. See Online Appendix A6 for more detail.

Truven External Validity Difference-in-Differences. In addition to using the weighted Truven data as a comparison group for our primary sample, we perform an analysis that weights our primary sample to match the Truven data age and gender profile. We weight our primary sample to look like the under-65 private insurance market in our firm's main state, so that the analysis can be thought of as externally valid for this state's age and gender demographic profile.[17]

We perform a difference-in-differences exercise similar to those just described, but instead comparing the spending change for our Truven-weighted primary sample with the spending change for the actual Truven sample. The second column in Table IV presents the main result for this exercise, a relative reduction in spending for our weighted primary sample of between -11.5% and -16.6%. Thus, overall, this exercise returns a spending change result that is quite similar to our primary sample time-series result.

Heterogeneous Treatment Effects. Table A5 in Online Appendix A7 presents treatment effect estimates for different cohorts of consumers categorized by health status, as well as by consumer demographics and broad categories of medical services. Table A5 also presents treatment effects broken down by age and employee income. Table A7 presents the standard errors for these category-specific, all of which are statistically different from zero at the 1% level, except for inpatient spending which is at the 10% level. Section IV dives deeper into spending reductions for specific services, and whether those reductions are achieved via changes in prices paid or quantities consumed.

IV Spending Reduction: Decomposition

In the previous section we provided a range of evidence illustrating the impact of increased cost sharing on medical spending, both overall and for specific types of patients and procedures. In this section we decompose the overall change in spending from the required switch to the HDHP into three main effects

[17] In Online Appendix A6, we replicate this analysis including income.

(i) consumer price shopping (ii) outright quantity reductions and (iii) quantity substitutions to lower-cost procedures. In doing so, we also control for any provider price changes that occur (potentially in response to the large-scale change in insurance).

For this decomposition, we restrict the set of provider-procedure combinations to those that have at least 15 observations over a given two years we study the change in spending for. This ensures that we have accurate price data for the services performed, and are using a consistent set of providers and procedures in the analysis. As they are based on specific procedure (CPT) codes, provider-procedure combinations are a relatively granular measure (e.g. a particular physician performing a diagnostic colonoscopy). Depending on the specialty and the specific procedure the degree of homogeneity can vary but for a substantial portion of our analysis this definition reflects a relatively homogeneous good. We discuss this at more length when we consider specific procedures, particularly the most common by volume and spending, as presented in Online Appendix A8.

The procedure-provider combinations used account for 77% of overall spending. In addition, we focus this analysis on the main region where the company employs people, in order to allow for the possibility that provider price changes could reflect market responses for providers in area where the firm has some monopsony power with respect to providers. The regional restriction reduces the number of employees in our analysis to an average of 16,814 (50,219 covered lives) per year, or about 70% of our primary sample. Online Appendix A8 performs some additional sensitivity analysis with respect to these restrictions.

Framework. We define the factors that we consider so that they are mutually exclusive and exhaustive for explaining the total change in medical spending, which we studied in the previous section. Total medical spending is composed of the prices consumers pay for care multiplied by the quantities they consume:

$$TS_t = \sum_{m,j} P_{m,j,t} C_{m,j,t}$$

Here, P is the price for a service m purchased from provider j at time t, and C is the number of services purchased by employees at the firm. The change in total spending from year t to $t+1$ is:

$$\Delta TS_{t+1,t} = \frac{TS_{t+1} - TS_t}{TS_t} = \frac{\mathbf{P}_{t+1} \cdot \mathbf{C}_{t+1} - \mathbf{P}_t \cdot \mathbf{C}_t}{\mathbf{P}_t \cdot \mathbf{C}_t}$$

Here, $\mathbf{P}_t$ refers to the vector of prices at time t across combinations procedures m performed by a given

provider j offering that procedure at t. $\mathbf{C}_t$ is the equivalent vector of health care consumed at t, giving the total quantities of procedures m performed by provider j at time t. Thus, $\mathbf{C}$ reflects the choices of specific procedure-provider combinations at a given point in time.

We decompose the change in total spending from one year to the next into specific factors that relate to either prices or quantities. We define the *provider price change index* as the average increase in medical prices paid, holding constant the providers visited, as well as the mix and quantity of services consumed. This procedure defines a Laspeyres index for provider price levels:

$$\Delta PPI_{t+1,t} = \frac{\mathbf{P}_{t+1} \cdot \mathbf{C}_t - TS_{t,t}}{TS_{t,t}} \tag{1}$$

Here, $PPI_{t+1,t}$ is the provider price change index resulting from provider price changes from year t to year $t+1$. Thus, e.g., if $t+1 = 2013$ and $t = 2012$, the index measures the increase in spending if the same provider-procedure combinations purchased in 2012 at 2012 prices were purchased instead at 2013 prices. This index takes into account a number of factors that lead to provider price changes including (i) basic medical price inflation and (ii) providers changing their prices in response to the regime shift to the HDHP.[18] We also present $\Delta PPI_{m,t+1,t}$, this provider price index for different specific procedures m.

The second component of our decomposition is the *price shopping effect*, which measures the extent to which consumers substitute to lower price providers conditional on receiving a specific kind of procedure m. To do this, e.g., for $2012 - 2013$, we hold the 2013 distribution of prices for provider-procedure combinations fixed, and examine whether, **for a given procedure**, consumers substituted to differently priced providers in their 2013 choices, relative to their 2012 choices. This decomposition assumes that the ranking of prices across providers within a class of procedures is constant over time, something that we verify is approximately true in Online Appendix A8.

Formally, take $\mathbf{P}_{m,J^m,t}$ to be the vector of prices for procedure m across the set of providers J^m offering that procedure, at year t. Define $\mathbf{C}_{m,J^m,t}$ as the vector of provider choices by consumers for procedure m in year t across the feasible set of providers J^m. Then, we define the price shopping statistic for procedure m as:

[18] Provider prices are typically set through negotiations with the insurer, who typically presents in-network inclusion as a 'take-it-or-leave-it' offer for smaller scale providers. If renegotiations are 'sticky' in the sense that they occur infrequently, our price index may overstate or understate the long-run impact of the HDHP plan on price changes.

$$\Delta PS_{m,t+1,t} = \frac{\mathbf{P}_{m,J^m,t+1} \cdot \mathbf{C}_{m,J^m,t+1} - \mathbf{P}_{m,J^m,t+1} \cdot \mathbf{C}_{m,J^m,t}}{\mathbf{P}_{m,J^m,t+1} \cdot \mathbf{C}_{m,J^m,t}} \tag{2}$$

For procedure m, the price shopping effect tells us, holding prices constant at $t+1$ prices, whether consumers shifted towards cheaper or more expensive providers from t to $t+1$, conditional on doing that procedure. We compute the price shopping effect for overall spending by holding the spending mix of procedures constant across procedures at year t spending, so that substitution across procedures does not impact our price shopping measure. Specifically, define $Y_{m,t}$ as the total spending for procedure m in year t and Y_t as total spending across all procedures in year t. Then, the overall price shopping effect is:

$$\Delta PS_{t+1,t} = \Sigma_{m=1}^{M} \frac{Y_{m,t}}{Y_t} \Delta PS_{m,t+1,t} \tag{3}$$

The overall price shopping effect tells us the extent to which consumers substitute to higher or lower priced providers from one year to the next year, conditional on doing a specific procedure, summed up across procedures. This statistic incorporates any effect related to the mix of providers patients see for a given procedure moving from one year to the next year. This includes, e.g., consumers shopping for providers with lower prices (as a result of the HDHP switch) or trends whereby consumers are moving over time towards seeing more (or fewer) expensive doctors.

The third part of the decomposition reflects *quantity changes* by consumers. We break down the contribution of quantity changes on total medical spending changes into two components: (i) quantity reductions and (ii) quantity substitutions to different medical procedures. We define quantity reductions in a straightforward manner:

$$\Delta Q_{t+1,t} = \frac{\sum_m \mathbf{C}_{m,J^m,t+1} - \sum_m \mathbf{C}_{m,J^m,t}}{\sum_m \mathbf{C}_{m,J^m,t}} \tag{4}$$

$\Delta Q_{t+1,t}$ represents the total increase or decrease in procedures performed from t to $t+1$. We also investigate this measure for specific m, i.e.$\Delta Q_{m,t+1,t}$. For this measure, we assume that provider billing behavior, apart from price changes, does not change over time. This is a weak assumption given that the policy change we study did not affect how providers were paid nor did the insurer studied make meaningful changes to billing over time.

If consumers shifted to lower priced procedures as a result of the HDHP plan shift, this would be ac-

counted for by a change in the average price per medical procedure consumed overall. This quantity substitution is the fourth and final part of our decomposition. We define the impact of quantity substitutions on total medical spending indirectly, as the residual of the change in total spending net of the first three parts of our decomposition, defined above:

$$\Delta QS_{t+1,t} = \Delta TS_{t+1,t} - \Delta PPI_{t+1,t} - \Delta PS_{t+1,t} - \Delta Q_{t+1,t} \tag{5}$$

This measure represents changes in spending unaccounted for by our conditional-on-procedure approach. For example, if a consumer responds to a change from year to year by choosing to treat their cancer with an intense chemotherapy approach rather than watchful medical management, it will be represented by a change in this measure.

We note here that our quantity change measures do not explicitly account for the anticipatory spending documented in the previous section, which reduced our estimate of the reduction in medical spending by between 3-7%. Figure A2 illustrates that anticipatory spending is associated with quantity changes: such spending is unlikely to impact the provider price index and price shopping statistics presented here. We discuss this further in the context of our results.

Price Shopping: Discussion. There are several important details to discuss for our analysis of price shopping, before presenting the results. First, concurrent with the required switch to high-deductible health care, the firm partnered with a leading health data technology firm to offer a tool to help employees search for lower medical prices in advance of getting care. This kind of tool is at the cutting edge of initiatives to increase consumer engagement and information in shopping for health care [see, e.g., Whaley (2015) or Lieber (2015) for related discussion and analysis.] Consequently, our setting is likely closer to the best-case scenario to expect price shopping to occur, rather than the typical environment consumers face. During year t_0, we partnered with the firm to run a survey studying consumer engagement with this more sophisticated shopping tool. Of consumers responding to the survey, 33% had heard of this price shopping tool, 22% had used this price shopping tool, and 4% reported having benefited from use of this tool.[19] These engagement levels are similar to those reported in, e.g., Whaley (2015), suggesting that consumers are still learning about such technology, and how they can use it to beneficially reduce health care spending.

[19]The survey was sent to a random sample of 6,000 employees at the firm and had a 25% response rate, likely selecting consumers more engaged with the health care shopping process.

In addition, we note that our aggregate price shopping statistic is performed **conditional on procedure** and not **conditional on episode of illness**. Thus, our measure incorporates shifting to lower priced providers for a given procedure, but not the impact of shifting to lower priced kinds of procedures for a given episode of illness. We quantify the impact of shifting to lower priced procedures in the quantity substitution measure we estimate. Of course, when we apply this price shopping measure to a specific procedure, this distinction is immaterial.

Results: Overall Spending. We now present the results for this decomposition, first for overall spending patterns, and second for specific types of spending. The top portion of Table V describes the results of this decomposition for the overall change in medical (non-drug) spending for consecutive years in our data. We report the results for all pairs of consecutive years from $t_{-4} - t_1$. Our main focus is on the t_{-1}-t_0 period when the required switch to the HDHP occurred (and subsequent t_0-t_1 trends). We present the results for the prior years to have a baseline for each effect.

The first column presents the year-on-year change in total spending for our modified primary sample, showing similar results to our Section III analysis. The second column presents the results for $\Delta PPI_{t+1,t}$, the provider price inflation index. The table illustrates how this effect is consistent and small across the four pairs of years studied, ranging from 0.2% for t_{-2}-t_{-1} to 3.4% from t_{-4}-t_{-3}. The effects for t_{-1}-t_0 and t_0-t_1 are both 1.7%. Given the similarity of these effects to those in the pre-treatment period, as well as to the overall medical price inflation index, we can rule out a large provider price change as a result of the required HDHP shift.

[TABLE V ABOUT HERE]

The overall price shopping effect $\Delta PS_{t+1,t}$, presented in the third column, is fairly small across the pairs of years studied ranging from -0.6% for t_{-4}-t_{-3} to 3.6% from t_{-1}-t_0. Interestingly, this effect is **largest** for t_{-1}-t_0, implying that after the required switch to the HDHP consumers are actually increasing the expense they are paying for a given procedure, rather than price shopping and moving to lower priced providers when they face a higher marginal price for care. The fact that this estimate goes in the 'wrong direction' (both overall and relative to prior trends) suggests both that (i) consumption trends may have shifted consumers towards more expensive providers conditional on a given procedure and, importantly, that (ii) medical spending was not markedly reduced due to consumers shopping for cheaper providers

for a given procedure.[20] These results are particularly striking insofar as we study an environment where consumers were given a comprehensive online tool to help them shop for prices for different procedures. The t_0-t_1 price shopping statistic is 0.7%, which is not sufficiently different from the prior year values to conclude that consumers learn to price-shop in year two after the required switch.

We note that these results do not imply that there is **no price shopping** or that consumers are **not learning at all**. Instead, they suggest that to extent such price shopping and learning to price shop occur, they do not meaningfully contribute to reduced spending in our environment. It is possible that as price shopping tools improve and consumers learn to use them over time, that price shopping could meaningfully contribute to reduced spending.

To provide some additional context for these price shopping results, Table VI presents a measure of *potential savings from price shopping* to give a sense of how large such savings could be in our environment, in a partial equilibrium sense. We compute a statistic that assesses what percentage of total spending would be saved if consumers who spend above the median price for a given procedure substituted to the median priced provider for that procedure in their region. For our overall spending metric, we then aggregate these statistics over all procedures. For each two year pair presented, the percentage that could be saved is based on potential substitutions in the second year of each pair. Column 1 shows potential price shopping savings for overall spending, which ranges from 18.3% from t_{-4}-t_{-3} to 21.1% in t_{-2}-t_1. t_{-1}-t_0 and t_0-t_1 values are 20.1% and 20.8% respectively. These results give a sense that there are quite a bit of potential savings from price shopping that are not currently being realized, though a complete welfare analysis would have to integrate factors such as travel costs and provider quality.

[TABLE VI ABOUT HERE]

Spending is not decreasing in t_0 and t_1 because of provider price decreases or consumer price shopping. The main reason for the total medical spending reduction after the required switch was quantity reductions by consumers. For the three pairs of years between t_{-4}-t_{-1}, the % change in overall medical service quantities ranges from 6.0-8.4%., indicating increasing quantities over that time frame. For t_{-1}-t_0, the quantity of services consumed dropped by 17.9%, and, thus, was the primary contributor to the drop in total medical spending over those two years as a result of the required HDHP shift. Interestingly, from t_0 to t_1,

[20] For robustness, in Online Appendix A8 we perform this decomposition for new employees, using a cross-sectional approach. The approximately 2,600 New employees (4,300 covered lives) in each year should be less likely to have existing provider relationships, potentially making them more likely to price shop. The results for new employees are almost identical to those for existing employees.

quantities increase by only 0.7%, indicating a lower growth rate than prior to the HDHP switch. The table also reports the impact of substitution across types of procedures on medical spending, and shows that this effect is negligible over time, ranging from -2.2% for t_{-1}-t_0 to 3.5% from t_0-t_1.

Since the nature of shopping is inherently different for prescription drugs than for non-drug medical services and providers, we perform a separate decomposition for prescription drugs. For prescription drugs, because allowed drugs prices are essentially the same across all in-network pharmacies, we combine the provider price index and price shopping index into one average price change index. The bottom panel of Table V shows these price and quantity changes for drugs for year pairs spanning t_{-4}-t_1, with the quantity change broken down into straight quantity reductions and the impact of substitution across drug types on spending.

As for all non-drug medical spending, drug spending increased at a steady rate from t_{-4}-t_{-1}, decreased sharply for t_0, and began to increase again in t_1. For all drugs, the drop in spending for t_0 was almost entirely due to quantity reductions (-17.8%). Price impacts on drug spending changes range from -4.3% to 6.4%, while quantity substitutions have limited impact on overall drug spending changes. Table A14 in Online Appendix A8 studies this decomposition separately for brand drugs and generic drugs. During the treatment period t_{-1}-t_0 the quantity of brand drugs consumed decreases by 30.3% while that of generics only decreases by 11.8%, both meaningful departures from the pre period trend. Quantity substitutions across the mixture of brand drugs reduces spending by 4%, while for generics this increases spending by 1.4%, suggesting together that consumers are substituting away from more expensive brand drugs to their generic counterparts. Additionally, price inflation for brand drugs is quite high over time (13.6% for $t_{-1} - t_0$), while generic drugs prices are decreasing in a meaningful way over time (-12.0% (13.6% for $t_{-1} - t_0$). Our upcoming analysis in this section investigates specific classes of drugs in more detail.

The middle panel of Table V presents the same decomposition for the sickest quartile of consumers in the population. As shown in Section III these consumers substantially reduce spending and it is particularly interesting to understand how and why they do so given that (i) over half of these consumers reach the out-of-pocket maximum in t_0 and (ii) these consumers may be economizing on valuable care. These consumers have similar contributing factors to their spending reductions as the population overall. Total spending decreases 19.5% from t_{-1} to t_0, with total spending increases of 6.1% and 5.9% for the prior two pairs of years. Over all two year pairs, the price inflation index ranges between -0.1% and 1.1%, with similarly small values for the price shopping index. The key component of spending reductions from t_{-1}-t_0 are quantity

reductions, which are responsible for a 20.0% reduction in spending (in prior years, this ranges from 3.5% to 4.1%). Quantity substitutions across procedures account for a 3.3% reduction in spending from t_{-1}-t_0. Unlike the population overall, there is a rebound effect at t_1 for these consumers: quantities rise by 9.0% from t_0-t_1, with a quantity substitution effect of 7.9%, indicting a movement / trend towards higher priced procedures.

We note that due to anticipatory spending, our t_{-1}-t_0 effects presented in Table V may overstate the total spending reduction and total quantity reduction. Section III showed that such spending accounts for between 3-7% of the t_{-1}-t_0 spending reduction: if this all comes from quantity substitution, for a representative set of quantities, then the total medical (non-drug) spending change for t_{-1}-t_0 will be roughly between 8.3-12.3% in this section, and the total quantity reduction between 10.9-14.9%. It is clear that, regardless of the anticipatory spending adjustment made, quantity reductions are the primary reason for the documented drop in medical spending due to the HDHP. Supporting analysis finds that this decomposition produces steady results throughout the year when comparing spending in a given month to spending in that same month a year earlier.

Results: Specific Procedures. Given that quantity reductions are responsible for almost all of the significant spending drop at the firm moving from t_{-1} to t_0, it is natural to ask what types of care consumers are reducing. In a typical model of moral hazard resulting from insurance, consumers would only reduce wasteful care that provides them with a benefit that is less than their out-of-pocket spending. However, as summarized nicely in Baicker, Mullainathan, and Schwartzstein (2015)), there is now ample evidence that consumers also reduce care that is likely valuable when faced with higher cost-sharing, a phenomenon which they term "behavioral hazard." In our context, where sicker consumers reduce quantities of care by meaningful amounts, it is important to understand exactly which types of care they economize on.

We begin with a broad analysis that documents and decomposes the spending changes over time for the 30 procedures on which consumers spend the most at the firm over our sample period. Table A12 in the Online Appendix presents summary statistics for this analysis aggregated across all 30 procedures, for each year pair we analyze. Overall, for these top 30 procedures, 73% had increases in quantity consumed from t_{-3} to t_{-2}, 80% had increases in quantity consumed from t_{-2} to t_{-1}, but only 17% had increases in quantity consumed over the treatment period t_{-1}-t_0. This number rebounded back to 80% for t_0-t_1. Price shopping and price index statistics are much more even over time: over all year pairs studied between

43-63% (37-70%) of these procedures had positive spending increases due to price shopping (rising price index). This suggests that cost-sharing might be an effective but blunt instrument to control health spending: higher cost-sharing reduces medical spending, but does so across the spectrum of medical procedures, some of which are likely valuable and others which are likely not. Table A13 in Online Appendix A8 shows a disaggregated view of this analysis for $t_{-1} - t_0$, presenting the decomposition separately for each of the 30 procedures. Of special note in that list are pregnancy related procedures, which have close to zero quantity changes over time and form a nice placebo test.

It is also important to specifically assess how consumers change spending and consumption both for procedures that are typically thought to be high-value and those typically thought to be low value. Though it is difficult to comprehensively classify the thousands of procedures we observe into high versus low value, it is possible to highlight and study specific procedures that are easier to classify.[21]

Table VII presents our spending change decomposition results for a collection of services that are generally considered to be high value. The results are presented for the treatment change period $t_{-1} - t_0$, as well as for an earlier year pair $t_{-3} - t_{-2}$ as an indicator of spending trends in the pre-treatment period.[22] The first high value services we consider are preventive health services, a large collection of medical services intended to improve population health in the long run by preventing the onset of costly and debilitating medical conditions (see, e.g., Chernew, Schwartz, and Fendrick (2015)). For an in depth discussion of the value proposition of preventive health services, see, e.g., Stange and Woolf (2008), and for an in depth discussion of other evidence of consumer take-up (or lack thereof) of preventive services see, e.g., Baicker, Mullainathan, and Schwartzstein (2015). The Affordable Care Act specifically seeks to encourage the use of preventive care by requiring it to be free of charge to consumers in all health plans (Kaiser Family Foundation, 2015b), a strong signal that such care is considered to be of high value. As a consequence of this regulation, preventive care in our context is free both before and after the required switch to high-deductible health care.

Despite the fact that preventive services are free both before and after the switch to high-deductible care, we find that consumers meaningfully reduce consumption of these services. For general preventive services consumers reduce quantity consumed by 7.5% from t_{-1} to t_0, with a further 5.2% reduction from $t_0 - t_1$.

[21] We also individually study each of the 30 medical procedures for which the employer and employees spent the most money. Table A12 in the Online Appendix presents summary statistics for this analysis across all 30 procedures. Online Appendix A8 presents the spending decomposition for $t_{-1} - t_0$ separately for each of these 30 procedures. These results add context to the aggregate results: Consumers reduce quantities across almost all of the medical procedures in this group.

[22] See Tables A15 and A16 in Online Appendix A8 for a complete time series of these decompositions.

Similar results hold for preventive services where a prior diagnosis is required (which may encompass more essential care): quantity reductions are 12.2% from t_{-1} to t_0 with only a small rebound effect (3.8%) from t_0 to t_1.[23] Both categories of preventive care have flat or upward quantity trends in years prior to the shift to high-deductible care. These two categories together comprise a meaningful portion (approximately 20%) of total medical spending studied in our modified sample. For prices, general preventive care has a 6.4% price index increase from $t_{-1} - t_0$, while preventive care with a prior diagnosis has 2.0% price increase. Neither type of preventive care has meaningful impacts of price shopping on overall spending.

At first glance, it is puzzling that consumers reduce free preventive care when consumers are required to switch to high-deductible health care at t_0. There are several possible hypotheses for why this occurs. First, consumers could have limited information on what services are considered preventive, as well as limited information about the fact that all preventive services are free under the HDHP. A second explanation is that consumption of preventive services are typically bundled together with more expensive services during visits to providers. If consumers reduce visits to providers because non-preventive health care is now costly, and preventive care is often an "add-on" to such visits, this could cause a reduction in preventive care consumption.

Online Appendix A9 presents analysis intended to help distinguish between these hypotheses. We decompose the reduction in preventive care consumption into extensive margin (fewer primary care visits) and intensive margin (fewer preventive services per primary care visit). If consumers consume the same amount of preventive care conditional on making an office visit, this suggests that they are not reacting heavily to a perceived price increase in preventive care, and instead going to their providers less because of the costs of other bundled services. If consumers reduce preventive care on the intensive margin, conditional on visiting their provider, this suggests that they are responding to a perceived price increase. We present a range of approaches to distinguish extensive from intensive margin changes, all delivering similar results and discussed in detail in the Online Appendix. These approaches clearly show that preventive care reductions are entirely on the extensive margin (-12.1% w/ main approach) rather than the intensive margin, where preventive care per visit actually increases (+3.5%). This supports the hypothesis that consumers reduce preventive care because it is bundled with other, costly, care consumed during primary care visits, as opposed to the hypothesis that consumers reduce such care because they think it is costly itself.

[23] Preventive care in our setting does not require a referral from a physician, even for care classified as preventive with a prior diagnosis. Preventive care is defined based on specific diagnosis and procedure codes, defined by the employer and insurance carrier. Care that is preventive with a prior diagnosis relates to specific medical conditions or demographics of an individual that automatically causes certain procedures to be classified as preventive, with zero cost-sharing.

[TABLE VII ABOUT HERE]

Table VII illustrates that consumers also reduce other kinds of care generally considered to be high value. Consumers reduce quantities of mental health care services by 5.4% from $t_{-1} - t_0$ and, notably, reduce quantities of physical therapy services by 29.7%. Consumers reduce quantities of diabetes drugs by 48%, statins for cholesterol management by 19.6%, antidepressants by 18.0%, and hypertension drugs by 24.2%. These quantity reductions are all strong departures from pre-period trends, and are not due to intertemporal substitution (increased purchases as t_{-1}). Online Appendix A8 provides more detail on this analysis, including the full time series of results for each service.

[TABLE VIII ABOUT HERE]

Table VIII presents our spending change decomposition for a collection of low-value services. Schwartz, Landon, Elshaug, Chernew, and McWilliams (2014) defines a collection of 26 low-value services with claims data, using clinically-based classifications from the American Board of Internal Medicine, the US Preventive Services Task Force, the National Institute for Health and Care Excellence, and the Canadian Agency for Drugs and Technologies in Health. They study Medicare beneficiaries, and show that that the use of these low-value services is widespread. We adopt the subset of services they classify that are directly relevant to an under-65 population, and investigate the impact of the shift to high-deductible care on their consumption. See Online Appendix A8 for more detail on this classification.

We find that the shift to high-deductible care causes large reductions from $t_{-1} - t_0$ across all of the low-value services we study, marking large departures from trend. Consumers reduce CT scans for sinuses with acute sinusitis by 26.0%, back imaging for non-specific low back pain by 21.3%, head imaging for uncomplicated headaches by 30.7%, and colorectal cancer screenings for patients under 50 by 26.2%. For drugs, antibiotics for acute respiratory infections are reduced by 44.4%. The low value services we study defined in Schwartz, Landon, Elshaug, Chernew, and McWilliams (2014) comprise approximately 1% of medical spending in total, they are potentially indicative of broader reductions in such low value care that we are unable to classify.

To get a sense of the broader implications for low value services, we also present the spending decomposition for the entire set of imaging services, which comprise 10% of medical costs in the modified primary sample. Many of the set of low value services are imaging services, and imaging services are often cited as

one area where wasteful care exists ((White House Report, 2014)). We find a substantial reduction in imaging spending from $t_{-1} - t_0$ (19.5%), while for prior year pairs spending increased between 5.5% and 12.4%. As with other kinds of care, spending reductions are almost entirely linked to quantity reductions (17.7%) as opposed to consumer price shopping or price index changes. The non-impact of price shopping is especially interesting for imaging, for which Table VI shows potential for 34.2% savings from price shopping if above median costs are reduced to median costs. See Online Appendix A8 for more detail on these results.

V Consumer Responses to Non-Linear Contract

As a result of the required shift to high-deductible health care from free health care, the consumers we study reduced health care spending between 11.79% and 13.80%. These spending reductions came in large part from well-off and predictably sick consumers facing reasonably low yearly out-of-pocket maximums. Moreover, consumers reduced spending almost exclusively by buying lower quantities of health care services, rather than through price shopping for cheaper services, or, indirectly, by having access to lower priced providers over time.

These facts clearly establish who reduced spending, and how they did so but they do not explain why. In this section, we investigate how consumers respond to the complex yearly price structure of the HDHP in order to explain why predictably sick and well-off consumers with low out-of-pocket maximums reduce medical spending. Our analysis is motivated by research across a range of industries suggesting that consumers may respond to 'spot' prices, i.e. the prices they face on any given day, rather than the price a fully rational consumer would respond to, which is the actual shadow price of current spending given the contract and expected future spending (we also refer to this as the expected marginal price). In the context of Medicare Part D prescription drug coverage, Einav, Finkelstein, and Schrimpf (2015), Dalton, Gowrisankaran, and Town (2015), and Abaluck, Gruber, and Swanson (2015) use different approaches to show that consumers markedly reduce consumption after they hit the 'donut hole' (a region where they pay 100% of cost), even when they should have clearly expected to end their year in that coverage region and face the full cost of marginal drug purchases. Aron-Dine, Einav, Finkelstein, and Cullen (2015) study consumer responses to non-linear insurance contracts in a large-employer health insurance setting, and conclude that consumers respond to both spot and true shadow prices for care during the year. Grubb and Osborne (2015), Nevo, Turner, and Williams (2016), and Ito (2014) study similar consumer responses to non-linear tariffs in

the contexts of cellular phone, broadband, and electricity markets, respectively. Liebman and Zeckhauser (2004) refer to this phenomenon as "schmeduling," and discuss behavioral foundations for why consumers may not respond to expected marginal prices in complex non-linear contracts.

In our environment, if consumers respond to simpler spot prices, rather than the true marginal (i.e. shadow) price of care, then they will under-consume care relative to what a fully rational dynamically optimizing consumer would do. This is true because the spot price in the HDHP is weakly decreasing during the year, and will thus always be weakly higher than the true shadow price of care. In some cases it will be much higher: for example, a predictably sick consumer will be under the deductible early in the year (spot price of 100% of cost) but will have a true shadow price close to 0%, since they can expect to get close to, or surpass, the plan out-of-pocket maximum. This could be one potential explanation for why predictably sick and relatively well-off consumers still reduce spending under the HDHP.

Our empirical environment is uniquely well suited to study consumer dynamic responses to spot and shadow prices in non-linear contracts. In the pre-period consumers are enrolled in free care and there are no within-year price dynamics. With the required switch to high-deductible care, the entire population is shifted to an environment where spot and shadow prices differ, and price dynamics matter. Given this setting, we use simple cross-sectional assumptions on population health together with detailed micro-level data on health status and incremental spending throughout the calendar year, pre and post switch, to trace out consumer responses to spot prices vs. shadow prices and the consequent implications for spending reductions.

Model. Denote consumer health status at the beginning of a calendar year by H_t and consumer demographics as X_t. Our key assumption maintains that, conditional on H_t and X_t, the cross-sectional distribution of population health needs at any month m during treatment year t is the same as that cross-sectional distribution at the same point in month m in control year t'. Formally, using t_0 as an example treatment year and t_{-2} as an example control year, we assume:

$$F_{t_0}[s_m, |H_{t_0}, X_{t_0}] = F_{t_{-2}}[s_m|H_{t_{-2}}, X_{t_{-2}}] \; \forall m = 1, ...12$$

Here, s_m describes the health state of consumers at the beginning of month m and F denotes the distribution of that health state. This assumptions implies that, conditional on ex ante health status and demographics,

the dynamic evolution of population health needs throughout the year is identical in the treatment year and the control year.[24]

We define the mapping from the health state and insurance contract to incremental consumer spending as:

$$G[S_{m+x} - S_m | s_m, H, X, Ins_m]$$

Here, S_m is year-to-date spending at the beginning of month m and S_{m+x} is the year-to-date spending at the beginning of month $m + x$. So, if $x = 1$, G reflects the distribution of incremental monthly spending in the population for month m, given the health state, insurance contract Ins_m, ex ante health status, and ex ante demographics. For a given month m, if $x = 12 - m$ then G reflects the distribution of rest of year spending from the beginning of m.

To implement our analysis, we assume that there is a one-to-one monotonic mapping between s_t, which is unobserved, and year-to-date spending S_m, conditional on H and X. This means, e.g., that if 35% of consumers have S_m that places them in the coinsurance region for the high-deductible plan at the beginning of June, t_0, those consumers can be directly compared to the 35% of consumers in t_{-2} in the same quantile range for S_m in that year.[25] This permits direct comparison between spending patterns within the calendar year for consumers under the HDHP in t_0, as a function of insurance contract prices, and those patterns for equivalent consumers in t_{-2} under free health care.

The final part of the model is the definition of different potential prices consumers might respond to in the HDHP as the calendar year evolves (the components of Ins_m). These prices are:

- **Spot Price,** P_m^s: This is the marginal price a consumer faces at the time they make the decision to consume health care. This corresponds directly to the three arms of the non-linear high-deductible contract, and equals 100% of the cost of care if consumers have not yet reached the deductible, 10% if in the coinsurance region, and 0% after reaching the out-of-pocket maximum.

[24] This assumes that, in the treatment years of t_0-t_1, consumers do not become, on average, sicker throughout the year due to dynamic effects from reducing the care consumed earlier in the year.To the extent that this assumption is violated, this will work against our main results as we will predict **lower** differences in spending for t_0 and t_1 relative to t_{-2} because consumers will be conditionally sicker in those years. Our upcoming analysis of consumers who have already passed the out-of-pocket maximum in the treatment years also supports the notion that such within-sample health effects on spending are minimal, since their incremental spending is identical to equivalent pre-period consumers.

[25] This concept manifests slightly differently for families, as opposed to individual consumers. For families, in the descriptive analysis we assume that families have one health state measure s_t. For our regression analysis, we pursue a more sophisticated approach that studies individual behavior within the family structure.

- **Shadow Price / Expected Marginal EOY Price,** $P_m^e = E_m[P_{EOY}^s | S_m, H, X, Ins_m]$: The shadow price is the expected marginal end-of-year price for a given consumer, given their health status and year-to-date spending at t. This price evolves dynamically throughout the year as risks are realized, and is the only price that a fully rational and informed consumer without liquidity constraints would use when making health care decisions.

- **Prior Year End Marginal Price,** P_m^L: This is the actual end of year price a consumer would have faced if their total medical spending during the prior year occurred in the HDHP. For consumers in t_1, this is their actual end-of-year price from t_0. For consumers in t_0, this is what their end-of-year price in t_{-1} would have been if they had been in the HDHP in that year. Consumers' behavior may respond to this price if they use their most recent risk realizations to project their shadow price of care.

Computing P_m^s is straightforward for each consumer and each month by mapping S_m to the corresponding non-linear contract spot price. Computing P_t^L is also straightforward, taking the spot price implied by the previous year's total spending applied to the HDHP. Computing the shadow price is more complex as it involves forming expectations about total end-of-year spending for each consumer at the beginning of each month. To construct P_m^e we use the following process:

1. For each month m define cells of equivalent consumers using the triple (H, X, S_m). We define these cells to be as precise as possible while maintaining sufficient sample sizes to determine a distribution of end-of-year spending realizations for each cell. In practice, we divide individuals by sextiles based on H_t. We use age as our only X variable, and split consumers into five age bins (0-15, 16-25, 26-35, 36-45, 46+). Then, for each cell combination of age and health, we divide consumers into deciles based on year-to-date spending S_m.

2. Assign individual i to one of these cells for each month m.

3. Form non-parametric end-of-year spending distribution for individuals i in each cell using all observations for actual end-of-year spending in that cell. Denote this $f_{i,m}(S_{i,EOY} | H, X, S_{i,m})$.

4. Combine individual end-of-year spending distributions into family distributions, assuming no correlation in spending for individuals with a family. The family distribution of end-of-year total spending is just the distribution of the sum of individual end-of-year spending across individuals in that family:

$$f_{j(i),m}(S_{EOY}) = \Sigma_{\Sigma_{S_{i,EOY}=S_{EOY}}} \Pi_i^{j(i)} f_{i,m}(S_{i,EOY})$$

5. The distribution of family end-of-year prices $P^s_{j,M}$ is the distribution that results from mapping the S_M coming out of $f_{j(i),m}(S_M)$ to the corresponding spot prices for each S_M:

$$P^e_{j,m} = \Sigma_{S_{EOY} \in \mathbf{S}_{EOY}} P^s_{j,EOY}(S_{EOY}) f_{j,m}(S_{EOY})$$

$P^e_{j,m}$ in our model is intended to serve as the price a rational and fully informed consumer should perceive as their true price of incremental care at m. We note that this framework is not intended to be a model of how consumers **actually** behave but rather a model of how a rational consumer in their situation would behave. Our upcoming analysis investigates whether consumers respond to alternative prices (e.g. spot prices or last year's end marginal price): if they do so, this suggests a departure from what a fully informed and rational consumer would do.[26]

Finally, we note that, when forming the expected end-of-year price, we deal with the issue of reverse causality (where cohort spending reductions imply changes to the expected end-of-year prices) by instrumenting for expected end-of-year prices in treatment years with the projected end-of-year prices for similar consumers prior to the required HDHP switch. These prices are correlated with those from equivalent consumers post-switch, but not correlated with changes to incremental spending that result post-switch. We use these instrumented versions of P^e_m throughout the descriptive and regression analysis.

Descriptive Analysis. We first use this framework as the basis for a series of descriptive analyses that investigate incremental consumer spending as a function of S_m and Ins_m across the calendar year. Then, we turn to regression analyses that formally quantify how consumers respond to the different possible prices they respond to. For parsimony, we present the descriptive analysis in this section for families (covering 3+ individuals total) since the majority of employees are in this coverage tier and the vast majority of spending comes from employees and dependents in this tier. Similar analysis for individuals and those with just one dependent are presented in the Online Appendix.

[26] It is important to note that, to the extent that our expected end-of-year price has statistical error, or is biased, this will suggest that consumers place some weight on other prices in our regression analysis. Additionally, it is possible that some measurement error in spot prices occurs if consumers undertake inflexible care plans that span multiple months where they pass through different regions of the non-linear contract. Given the precision of our model for expected EOY prices, and the large emphasis on spot prices we find in our results, these issues seem like secondary concerns.

Our first set of descriptive analyses examines incremental spending (CPI adjusted) by month for consumers in t_0 (or t_1) relative to that spending by equivalent consumers under free insurance in t_{-2}.[27] We examine the distribution of consumers' incremental spending for (i) the next month and (ii) the rest of the year, starting at any given month m. We begin by examining incremental spending as a function of the spot price consumers face at the beginning of month m in t_0, and compare that to the incremental spending of the equivalent quantiles of consumers for S_m in t_{-2}.

Figure III shows the mean and median incremental spending *for the next month* (left panel) for families who have passed the out-of-pocket maximum by month m in t_0 and the comparison group in t_{-2}. The figure presents the results for July-December of the calendar year, since few families pass the out-of-pocket maximum prior to those months in t_0.[28] The figure illustrates that incremental spending for the next month is essentially the same for families in t_0 who have passed the out-of-pocket maximum at t and their comparison quantiles of families in t_{-2}.

These results suggest that once consumers have passed the out-of-pocket maximum under the HDHP in t_0, they spend exactly as much as they would have spent incrementally as in t_{-2}. Since consumers who pass the out-of-pocket maximum always have $P_m^s = P_m^e = 0$, the same spot and shadow prices as the pre-period, the fact that these consumers spend the same in t_0 as their comparison groups do in t_{-2} provides a check showing that consumers respond equivalently to a price of zero in both periods. It also provides a simple test for our empirical strategy, akin to a placebo test. Were our assumptions about disease dynamics driving biased results we would expect to find differences even when prices are the same in both t_0 and t_{-2}. Additionally, it implies that all of the spending and quantity reductions that we document earlier in this paper, including those for the sickest ex ante quartile of consumers, must come from consumers when they are either in the deductible arm or the coinsurance arm of the HDHP.

[FIGURE III ABOUT HERE]

The right panel in Figure III presents the analogous results for consumers who begin a month in the coinsurance arm of the high-deductible plan in t_0. It is evident that both incremental monthly spending and incremental rest of year spending (Figure A12) are essentially the same for the treatment cohorts in t_0

[27] We use t_2 as our main control year to remove pre-period anticipatory spending that occurs in t_{-1}. In t_0, spending in January and February may be depressed because of anticipatory t_{-1} spending, as discussed in Section III. This becomes a smaller concern as we move through the year t_0 and is not of high enough magnitude to markedly impact our results. Our results for t_1 are consistent with those from t_0.

[28] Table A25 shows the share of families who are in each non-linear contract arm at the beginning of a given month. In t_0, 673 are in the out-of-pocket maximum region in July, increasing up to 1,655 by December.

and their relevant comparison groups in t_{-2}. This is true uniformly throughout the calendar year. Once consumers reach the coinsurance region, their spending does not drop relative to the pre-period in free health care. Taken together with the out-of-pocket maximum results, this suggests that **essentially all** the reductions we have documented for reduced post-period spending come from consumers when they are actually under the deductible in the calendar year.

This is borne out when we examine the analogous figures for families who begin a given month under the deductible. The figure shows substantial decreases in incremental monthly spending for consumers under the deductible in t_0, relative to their t_{-2} comparison groups. This decrease is approximately 25-30% throughout the calendar year for mean monthly spending, and 50% throughout the year for median spending. As expected, rest of year spending also drops for consumers in the treatment cohorts relative to the comparison cohorts.

[FIGURE IV ABOUT HERE]

When combined with our earlier descriptive evidence on predictably sick consumers reducing spending, these analyses suggest that these consumers only reduce spending when under the deductible, even though they should predictably go well past the deductible during the calendar year. We explore this more precisely in Figure V.

[FIGURE V ABOUT HERE]

Figure V presents incremental monthly spending (left panel) and rest-of-year spending (right panel) for families who (i) start a month under the deductible in t_0 and (ii) are in the lowest quartile of expected end-of-year price (sickest quartile).[29] This panel shows that these consumers substantially reduce incremental monthly spending early in the year: for example, in March, the sickest quartile of consumers under the deductible reduce mean spending by about 25% relative to their t_{-2} comparison group, despite the fact that these consumers average about $15,000 in spending for the rest of the year, suggesting that they will easily pass the deductible on average.[30] Rest-of-year spending declines by a meaningful amount for these

[29] It is important to note that the mixture of consumers under the deductible becomes notably healthier as the year goes on (since sick consumers spend money and move to the coinsurance region). Consequently, though we present the analysis for February - December for completeness, the months early in the year are most relevant since this is when truly predictably sick consumers are still under the deductible.

[30] As shown in Table A26, these consumers have expected end-of-year prices of 0.08, and almost certainly end the year in either the coinsurance or out-of-pocket maximum region (where they no longer reduce incremental spending).

predictably sick consumers, suggesting that reduced spending early in the year when under the deductible is not compensated for by larger spending later in the year once the deductible has been passed.

Applying a more stringent criterion — the sickest 10% of the population — we find patterns that mimic those for the sickest quartile, and show that these consumers reduce spending early in the year, despite having mean true shadow prices of 0.06. See Figure A17 in the Online Appendix for these additional results.

[TABLE IX ABOUT HERE]

Table IX brings together these descriptive analyses to illustrate the proportion of total yearly savings due to incremental monthly spending changes for consumers who start a given month in a given plan arm. 91% of the total yearly spending reductions from t_{-2} to t_0 comes from consumers who started a given month under the deductible. The Table shows that 25% of all spending reductions during the year come from consumers who are (i) under the deductible and (ii) predictably sick in the sense that they have low expected shadow prices of care. Interestingly, 24%, 19%, and 23% of total spending reductions come from families in quartiles 2, 3, and 4 of shadow prices: this suggests that healthier consumers ex ante are also responsible for large portions of overall spending reductions, and that those occur when they are under the deductible during the year.

Figure A16 in Online Appendix A10 replicates this analysis for t_1 spending. The figure highlights that the patterns we discussed in depth for t_0 spending continue to hold in t_1. This suggests that consumers do not rapidly learn to respond to their true shadow prices, as opposed to the spot prices throughout the year, after one year of experience in the HDHP.

Regression Analysis. Now, we perform a series of regression analyses to deal with underlying correlations in the data and more precisely quantify the impacts of different non-linear contract prices on total medical spending. Our primary regression studies incremental monthly spending for families in the t_0 and t_1 treatment years relative to their t_{-2} comparison quantile groups. Our main specification is:

$$\begin{aligned}\log(Y_{i,m}+1) &= \alpha + [\beta_e P^e_{i,m} + \beta_s P^s_{i,m} + \beta_L P^L_i] + [\theta_e P^e_{i,m} + \theta_s P^s_{i,m} + \theta_L P^L_i] I_{t_0-t_1} \\ &\quad + [\kappa_e P^e_{i,m} + \kappa_s P^s_{i,m} + \kappa_L P^L_i] I_{t_1} + \gamma_H H_i + \gamma_X X_i + \gamma_{Y^l} \Sigma^2_{l=1} \log(Y_{i,t-l}+1) \\ &\quad + \Sigma_{m\in M}\gamma_m I_m + \Sigma_{t\in T}\gamma_t I_t + \epsilon_{i,m}\end{aligned}$$

Here, $Y_{i,m}$ is total monthly incremental spending (insurer + out-of-pocket) in month m for a given family. P^k are the three prices defined at the family-level for each month m. The regression includes observations from one control year, t_{-2}, and both treatment years, t_0 and t_1. Importantly, we define counterfactual HDHP non-linear contract prices for the t_{-2} control population using the same quantile comparison method discussed earlier in this section: this means that conditional on (H, X) we match deciles of S_m in t_{-2} to comparable deciles in t_0 and t_1, and assign the t_{-2} consumers the same prices as those treatment year consumers. This mimics the approach used in the descriptive analysis comparing treatment consumers to comparable control consumers, leveraging the cross-sectional assumptions described earlier. The regressions control for ex ante family health status (adding up individual family spending predictions), demographics (ages, family size, gender mixture), and calendar month and year fixed effects. Additionally, the regressions control for lagged spending from each of the prior two months, to deal with spending autocorrelation.

Our primary parameters of interest are the interaction of price measures and treatment years. The θ_k coefficients gives an estimate for the % reduction in incremental monthly spending as a function of each kind of non-linear contract price in the treatment years. For example, $\theta_k = 0$ would imply that, conditional on health status, demographics, and other prices, families do not change spending in response to changes in P^k. Negative values imply that consumers reduce spending by $\theta_k\%$ in response to a price change of 1 (i.e. 100%). The κ_k parameters are also of interest, and measure whether consumers' responses to the different non-linear contract prices change in t_1, after they have already been enrolled in the HDHP for a full year. By including prices directly in the regression in the period prior to the introduction of the HDHP we can flexibly capture any mechanical correlations between estimated prices and spending.[31]

When we implement these regressions, we use indicator variables to represent various values of each P^k. For spot prices and prior year-end marginal prices this is natural, since 0,.1, and 1 are the only possible values for these prices. We omit the value of 0 (consumers passed the out-of-pocket maximum) and include two dummies for starting a month (ending the year) in the deductible arm or coinsurance arm. For the shadow price in the current year (expected end-of-year marginal price) our main specification considers quintiles of this price, described in our results table, though we also examine a specification with ventiles. We note that, as discussed earlier, we use instrumented versions of expected end-of-year prices in the treatment years to deal with the issue of reverse causality (where cohort spending reductions imply changes to the expected end-of-year prices).[32] Finally, it is important to note that if our measures of expected future prices are noisy

[31] Table A27 presents the correlations in these three prices at different months during the calendar years in t_0 and t_1.

[32] To do this we use projected end-of-year prices for comparable quantiles of consumers in t_{-3}, prior to the required HDHP

projections of true shadow prices, this will reduce the magnitude of our expected price coefficients (biased towards 0) which works against the results we eventually find.

[TABLE X ABOUT HERE]

Table X presents the results from our primary specification, along with five robustness analyses. Our primary specification shows that on average, in t_0, consumers under the deductible reduce incremental monthly spending by 42.2%, significant at the 1% level, **controlling for their shadow prices and prior year-end marginal price**. This treatment effect for t_1 is not statistically different from that for t_0, with a small standard error of 0.0374 for this difference. Consumers in the coinsurance region at the start of a month in t_0 reduce incremental spending by 14.4% on average, with this t_0 effect statistically the same as the t_1 effect.

Consumers' responses to their true shadow prices are much lower in magnitude: for example, consumers in the 4th highest shadow price quintile (0.275, 0.730) only reduce incremental spending by 6.66%, statistically significant at 1%, relative the control group consumers (and omitted t_0 OOP-max consumers) who have shadow prices of 0. These results are similar across the quintiles, except for quintile 5 (highest shadow prices) which shows **higher** relative spending, likely due to the presence of many consumers spending 0 in this group regardless of the price regime. The coefficients which examine the t_1 differential for these treatment effects are positive and small, suggesting that consumers are not learning that the shadow prices are the true prices they should consider.

The coefficient on prior year-end marginal price is small and positive for t_0 when t_{-1} end of year spending would have placed the consumer under the HDHP deductible. This suggests that this is not a meaningful driver of spending reductions in t_0. However, the coefficient examining the t_1 differential is -9.6%, statistically significant at 1%, suggesting that consumers in t_1 who ended t_0 under the deductible reduce incremental monthly spending by 10% in t_1. This suggests that, to the extent that consumers learned about the HDHP from t_0 to t_1, they learned based on their prior-year end-of-year price realization, rather than through an understanding of the more complex shadow price. Ending the prior year in the coinsurance arm does not have a meaningful impact on next year spending next year, either in t_0 or t_1.

switch (and prior to the observations included in the regression). These prices are correlated with those from equivalent consumers post-switch, but not correlated with changes to incremental spending that result post-switch. It is important to note that these prices will be biased slightly lower than actual t_0 and t_1 shadow prices (because spending in the pre-period is higher). However, because the change in total spending implies only small changes in these shadow prices, this should not have a meaningful impact on our results.

Table X also presents five regressions to assess the robustness of our primary specification. The results in these alternative specifications, described in more detail in Online Appendix A10, are similar to those from the primary specification just presented. Additionally, Online Appendix A11 presents results from a LASSO penalized regression model that supports the key findings presented here.

Non-Linear Contract Discussion. Taken in sum, these regression results illustrate that relative to shadow prices and last year's ending marginal price, spot prices are the primary driver of the spending reductions we document. Shadow prices have a limited impact on spending reductions. Consumers also have limited responses to the prior year's end-of-year marginal price in the first HDHP plan year, t_0, but increasingly respond to that price in t_1, the second year of HDHP enrollment. Though our analysis cannot assess whether consumers will learn to respond to the shadow price of care over a longer time horizon than two years, the prevalence of related results in different contexts suggests that consumers' emphasis on spot prices persists to a meaningful extent.

There are several possible micro-foundations for why consumers respond heavily to spot prices, rather than their true shadow prices. As modeled in Dalton, Gowrisankaran, and Town (2015) consumers could be myopic, or, more generally, consumers could have high discount rates. Liebman and Zeckhauser (2004) discusses how in the context of complex prices, consumers may engage in "schmeduling," constructing a heuristic price that they feel reflects their choice environment. Another potential explanation is limited information: consumers could either have limited information about (i) their own health risks or (ii) key non-linear contract features (Handel and Kolstad (2015)). Though we do not differentiate between these foundations, the facts we establish have important implications for cost control and consumer health behaviors regardless of the underlying explanation. Studying these mechanisms, and their implications for policies to reduce these biases, is an important path for future research.

An additional potential explanation for spot price responses is liquidity constraints, whereby consumers are more likely to reduce spending when under the deductible because they don't have the discrete amount of money they need to make deductible payments. In our setting, liquidity constraints are highly unlikely to be material. First, our employee base is quite high income, implying that their flows of flexible income are substantial relative to the deductible being paid. Second, these consumers generally have easy access to credit. Third, perhaps most importantly, consumers directly receive the amount of their deductible into their HSAs at the beginning of the year, money that is specifically earmarked for health spending. Thus,

while liquidity constraints may be potentially quite important in other settings for explaining responses to spot prices, they are unlikely to be so in our context.

VI Conclusion

We studied the health care decisions and spending behavior for a large population of consumers who were required to switch into high-deductible insurance after years of having access to completely free health care. The change caused a spending drop between 11.79% and 13.80%, occurring across the spectrum of health care service categories. We investigated whether spending reductions came from (i) consumer price shopping for cheaper providers (ii) quantity reductions or (iii) substitution across procedures by consumers. We clearly documented that spending reductions were due almost entirely to consumer quantity reductions across a broad range of services, including some that were likely of high value in terms of health and potential to avoid future costs. Consumers did not shift to cheaper providers, in either of the two years we observe post-switch.

A meaningful portion of all spending reductions came from well-off consumers who were predictably sick, implying that the true marginal prices they faced under high-deductible care were actually quite low. We investigated consumers' responses to the different potential prices they might perceive in the non-linear high-deductible insurance contract to help explain the puzzle of why these consumers reduce spending. We found that almost all spending reductions during the year occurred while consumers were still under the deductible, despite the fact that the majority of incremental spending occurs for consumers that have already passed the deductible. Moreover, about 30% of **all** spending reductions come from consumers in months when they (i) began that month under the deductible but (ii) were predictably sick, in the sense that they had very low shadow prices for health care. Once these consumers (predictably) reached the coinsurance arm and out-of-pocket maximum arms of the non-linear contract, they did not reduce spending further. These spending patterns are almost identical for t_1, implying that consumers did not learn to respond to the true shadow prices of care by the second-year of enrollment in high-deductible health care. Our regression analysis shows that consumers reduce spending by 42.2% when under the deductible, controlling for both their shadow prices and last year's end-of-year marginal price. Additionally, consumers reduce relative spending by 10% in t_1 when they ended t_0 under the deductible.This suggests that while consumers may not respond to their true shadow price of care in the second-year, they do respond somewhat to their price

experience in the prior year.

We assess not only **whether** consumers reduced spending but **how**, leading to insights with potentially important normative implications. Despite studying an environment with educated, technologically-savvy, and high-income consumers who have access to a near state-of-the-art price shopping tool, we find that price shopping is not an important component of the spending reductions resulting from the switch to high-deductible care. Instead, we find that outright health care quantity reductions across the spectrum of services drive lower spending. This suggests that the nature of those quantity reductions is crucial, in the current climate, for assessing the welfare impact of increased cost-sharing (Baicker, Mullainathan, and Schwartzstein (2015)). We document meaningful reductions in care that is likely valuable and care that is potentially wasteful. We believe that a comprehensive assessment of whether such quantity reductions are welfare increasing on net is an important path for future research. Additionally, we believe that further research on the positive and normative implications of different "value-based" contract designs (see, e.g., Chernew, Rosen, and Fendrick (2007)) is crucial to assess the degree to which tailoring out-of-pocket payments to specific health behaviors can drive purchasing value. It is clear that such contracts can improve on designs that lump all services together. It is less clear, however, how specific such contracts can be before they become too complex for consumers to effectively navigate. If the effectiveness of such contracts is limited by their complexity, the best supply-side policies may be more effective for efficiently cutting back on high cost, low value care.

Our results also suggest the typical structure of non-linear health insurance contracts, with decreasing marginal prices throughout the year, reduces medical consumption and may yield dramatically different behavior relative to plans that have flatter structures throughout the year. This creates a challenge for employers and regulators: highly non-linear contracts, such as a catastrophic contract with a large deductible that transitions directly to a stop-loss, will help control spending and protect consumers from large financial risks, relative to flatter contracts, but may also discourage the use of valuable services (as well as wasteful ones). For example, a transition to decreasing non-linear tariffs in Medicare Part D may reduce overall spending and better protect consumers from financial risk, but also discourage adherence to important medications (Einav, Finkelstein, and Schrimpf (2015)). Further, when consumers can choose between different kinds of non-linear contracts, it will be important to consider whether their bias towards spot prices also biases them away from choosing high-deductible plans (Bhargava, Loewenstein, and Sydnor (2015)). We believe that a careful empirical investigation of optimal non-linear contract design in the context of these responses

to different price signals, building on work such as Vera-Hernandez (2003), is a valuable avenue for future research.

Author Affiliations

a: University of California Berkeley, Department of Economics

b: Harvard University, John F. Kennedy School of Government, and NBER

c: University of California Berkeley, Department of Economics, and NBER

d: University of California Berkeley, Haas School of Business, and NBER

References

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

Anastov, Pavel and Tom Baker. "Putting Health Back Into Health Insurance Choice." *Medical Care Research Review*, 71(2014), 337–355.

Aron-Dine, Aviva, Liran Einav, and Amy Finkelstein. "The RAND Health Insurance Experiment, Three Decades Later." *Journal of Economic Perspectives*, 27(2013), 197–222.

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

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein. "Behavioral Hazard in Health Insurance." *Quarterly Journal of Economics*, 130(2015), 1623–1667.

Baker, Laurence C., M. Kate Bundorf, and Daniel P. Kessler. "Does Health Plan Generosity Enhance Hospital Market Power?" *Journal of Health Economics*, 44(2015), 54–62.

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor. "Do Individuals Make Sensible Health Insurance Decisions? Evidence from a Menu with Dominated Options." (2015). NBER Working Paper No. 21160.

Blinder, Alan S. "Wage Discrimination: Reduced Form and Structural Estimates." *Journal of Human Resources*, 8(1973), 436–455.

Bundorf, M. Kate. "Consumer-Directed Health Plans: Do They Deliver?" (2012). Robert Wood Johnson Foundation.

Buntin, Melinda B., Amelia M. Haviland, Roland McDevitt, and Neeraj Sood. "Healthcare Spending and Preventive Care in High-Deductible and Consumer-Directed Health Plans." *American Journal of Managed Care*, 17(2011), 222–230.

Cabral, Marika. "Claim Timing and Ex Post Adverse Selection." (2013). University of Texas Working Paper.

Carlin, Caroline and Robert Town. "Adverse Selection, Welfare, and Optimal Pricing of Employer Sponsored Health Plans." (2009). University of Minnesota Working Paper.

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

Chernew, Michael, Allison B. Rosen, and Mark Fendrick. "Value-Based Insurance Design." *Health Affairs*, 26(2007), 195–203.

Chernew, Michael, Sanford Schwartz, and Mark Fendrick. "Reconciling Prevention and Value in the Health Care System." (2015). `http://healthaffairs.org/blog/2015/03/11/reconciling-prevention-and-value-in-the-health-care-system/`.

CMS. "National Health Expenditure Data." (2015). `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html`.

Cutler, David. "From the Affordable Care Act to Affordable Care." *Journal of the American Medical Association*, 314(2015), 337–338.

Dalton, Christina M., Gautam Gowrisankaran, and Robert Town. "Myopia and Complex Dynamic Incentives: Evidence from Medicare Part D." (2015). NBER Working Paper No. 21104.

Einav, Liran, Amy Finkelstein, and Mark Cullen. "Estimating Welfare in Insurance Markets Using Variation in Prices." *Quarterly Journal of Economics*, 125(2010), 877–921.

Einav, Liran, Amy Finkelstein, Stephen P. Ryan, Paul Schrimpf, and Mark R. Cullen. "Selection on Moral Hazard in Health Insurance." *American Economic Review*, 103(2013), 178–219.

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*, 130(2015), 841–899.

Ellis, Randall P., Shenyi Jiang, and Willard G. Manning. "Optimal Health Insurance for Multiple Goods and Time Periods." *Journal of Health Economics*, 41(2015), 89–106.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and The Oregon Health Study Group. "The Oregon Health Insurance Experiment: Evidence From the First Year." *Quarterly Journal of Economics*, 127(2012), 1057–1106.

Grubb, Michael D. and Matthew Osborne. "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock." *American Economic Review*, 105(2015), 234–271.

Handel, Benjamin R. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *American Economic Review*, 103(2013), 2643–2682.

Handel, Benjamin R. and Jonathan T. Kolstad. "Health Insurance for "Humans:" Information Frictions, Plan Choice, and Consumer Welfare." *American Economic Review*, 105(2015), 2449–2500.

Haviland, Amelia M., Matthew Eisenberg, Ateev Mehrora, Peter Huckfeldt, and Neeraj Sood. "Do Consumer Directed Health Plans Bend the Cost Curve Over Time?" *Forum for Health Economics & Policy*, 46(2016), 33–51.

Haviland, Amelia M., Susan Marquis, Roland McDevitt, and Neeraj Sood. "Growth of Consumer Directed Health Plans To One-Half of All Employer-Sponsored Insurance Could Save 57 Billion Annually." *Health Affairs*, 31(2012), 1009–1015.

Ito, Koichiro. "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing." *American Economic Review*, 104(2014), 537–563.

Kaiser Family Foundation. "Employee Health Benefits: 2014 Summary of Findings." (2015a).

———. "Preventive Services Covered by Private Health Plans Under the Affordable Care Act." `http://kff.org/health-reform/fact-sheet/preventive-services-covered-by-private-he` (2015b).

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

Lieber, Ethan M.J. "Does it Pay to Know the Prices in Health Care?" (2015). Notre Dame Working Paper.

Liebman, Jeffrey B. and Richard J. Zeckhauser. "Schmeduling." (2004). Harvard Working Paper.

Lo Sasso, Anthony T., Lorens A. Helmchen, and Robert Kaestner. "The Effects of Consumer-Directed Health Plans on Health Care Spending." *The Journal of Risk and Insurance*, 77(2010).

Nevo, Aviv, John L. Turner, and Jonathan W. Williams. "Usage-Based Pricing and Demand for Residential Broadband." *Econometrica*, 84(2016), 411–443.

Newhouse, Joseph P. and the Insurance Experiment Group. *Free For All?: Lessons from the RAND Health Insurance Experiment*. Harvard Univ. Press (1993).

Oaxaca, Ronald. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review*, 14(1973), 693–709.

Schwartz, Aaron, Bruce Landon, Adam Elshaug, Michael Chernew, and Michael McWilliams. "Measuring Low-Value Care in Medicare." *JAMA Internal Medicine*, 174(2014), 1067–1076.

Stange, Kurt and Steven Woolf. "Policy Options in Support of High-Value Preventive Care." *Partnership for Prevention*, (2008).

Towers Watson. "High-Performance Insights: Best Practices in Health Care." *20th Annual Towers Watson / National Business Group on Health Employer Survey on Purchasing Value in Health Care*, (2015).

Vera-Hernandez, Marcos. "Structural estimation of a principal-agent model: moral hazard in medical insurance." *RAND Journal of Economics*, 34(2003), 670–693.

Whaley, Christopher. "Searching for Health: The Effects of Online Price Transparency." (2015). Berkeley Working Paper.

White House Report. "The Affordable Care Act and Trends in Health Care Spending." `https://www.whitehouse.gov/sites/default/files/docs/fact_sheet_implementing_the_affordable_care_act_from_the_erp_2013_final1.pdf` (2014).

Table I: Sample Demographics

	PPO or HDHP in t_{-1}	PPO in t_{-1}	Primary Sample
N - Employees	[35,000-60,000]*	[35,000-60,000]*	22,719
N - Emp. & Dep.	[105,000-200,000]*	[105,000-200,000]*	76,759
Enrollment in PPO in t_{-1}	85.21%	100%	100%
Gender - Emp. & Dep. **% Male**	51.9%	51.5%	51.4%
Age, t_{-1} - Employees			
18-29	12.0%	10.3%	4.3%
30-54	83.2%	84.8%	91.4%
≥ 55	4.8%	4.9%	4.3%
Age, t_{-1} - Emp.& Dep.			
< 18	34.5%	35.3%	36.1%
18-29	12.3%	11.5%	8.8%
30-54	50.1%	50.1%	52.0%
≥ 55	3.1%	3.1%	2.8%
Income, t_{-1}			
Tier 1 ($<$ \$100K)	8.4%	8.2%	7.3%
Tier 2 (\$100K-\$150K)	65.0%	64.9%	64.7%
Tier 3 (\$150K-\$200K)	21.8%	22.0%	22.6%
Tier 4 ($>$ \$200K)	4.9%	4.9%	4.7%
Family Size, t_{-1}			
1	23.7%	21.4%	16.1%
2	19.6%	19.1%	17.9%
3+	56.7%	59.5%	65.9%
Individual Spending, t_{-1}			
Mean	\$5,020	\$5,401	\$5,223
25th Percentile	\$609	\$687	\$631
Median	\$1,678	\$1,869	\$1,795
75th Percentile	\$4,601	\$5,036	\$4,827
95th Percentile	\$18,256	\$19,367	\$18,810
99th Percentile	\$49,803	\$52,872	\$52,360

*Exact numbers concealed to preserve firm anonymity.

This table presents summary demographic statistics for (i) employees enrolled in the PPO or HDHP plan options at the firm in t_{-1}; (ii) employees enrolled in the PPO plan option at the firm in t_{-1}; and (iii) our final sample, which is restricted to employees present in all six years of our data, and their dependents. This sample is described in depth in the text. When relevant, statistics for the primary sample are presented for the year t_{-1}. Table A1 in Online Appendix A1 replicates our key statistics for an alternative primary sample.

Table II: Health Plan Characteristics

	PPO	HDHP*
Premium	\$0	\$0
Health Savings Account (HSA)	No	Yes
HSA Subsidy	-	[\$3,000-\$4,000]**
Max. HSA Contribution	-	\$6,250***
Deductible	\$0****	[\$3,000-\$4,000]**
Coinsurance (IN)	0%	10%
Coinsurance (OUT)	20%	30%
Out-of-Pocket Max.	\$0****	[\$6,000-\$7,000]**

* We do not provide exact HDHP characteristics in order to help preserve firm anonymity.

**Single employees (or employees with one dependent) have .4× (.8×) the values given here.

***Single employee legal maximum contribution is \$3,100. Employees over 55 can contribute an extra \$1,000 as a 'catch-up' contribution.

****For out-of-network spending, PPO has a very low deductible and out-of-pocket max. both less than \$400 per person.

This table presents key characteristics of the two primary plans offered over time at the firm we study. We present characteristics for the family tier (the majority of employees), with levels for single employees and couples noted below. Both plan options were present at the firm from $t_{-4} - t_{-1}$, but the PPO option was removed in t_0: plan characteristics remained the same throughout the study period.

Table III: Spending Impact of HDHP Switch

	Model		
	(1) Raw Spending	(2) CPI & Age Adj.	(3) Intertemp. Substitution
Year, Mean Spend			
t_{-4}	4,031.49	3,910.87	3,910.87
t_{-3}	4,256.21	3,858.78	3,858.78
t_{-2}	4,722.03	4,055.01	4,051.01
t_{-1}	5,222.60	4,277.84	4,112.61
t_0	4,446.08	3,490.97	[3,490.97 , 3,656.20]
t_1	4,799.14	3,599.25	3,599.25
% Decrease			
t_{-1}-t_0	-14.9%***	-18.4%***	[-11.1%, -15.1%]***
	(1.4%)	(1.4%)	[(1.5%),(1.4%)]
t_{-1}-t_1	-8.0%***	-15.9%***	-12.5%***
	(1.5%)	(1.4%)	(1.4%)
Semi-Arc Elasticity*	-0.57	-0.85	[-0.59,-0.69]

*Elasticities average t_{-1}-t_0 and t_{-1}-t_1 estimated effects

*** Statistically significant from no change at 1% level.

This table details the treatment effect of the required HDHP switch under different frameworks: (i) nominal spending (ii) age and CPI adjusted spending and (iii) estimates with anticipatory spending (age and CPI adjusted).

Table IV: Difference-in-Differences Analysis of HDHP Switch

	Model		
	(4)	(5)	(6)
	Early Switcher DID	Ext. Validity Truven Weighted DID	Truven-Control DID
% Decrease			
t_{-1}-t_0	[-11.31%, -15.20%*]	[-11.5% , -16.6%]***	[-22.6% , -26.6%]***
Semi-Arc Elasticity	[-0.56,-0.76]	[-0.57,-0.82]	[-1.12,-1.32]

* Statistically significant from no change at 10% level.

*** Statistically significant from no change at 1% level.

This table details the treatment effect of the required HDHP switch under three specifications described in the text: (i) early switcher difference-in-differences (ii) external validity difference-in-differences using weights derived from Truven MarketScan data and (iii) Truven control group difference-in-differences.

Table V: Spending Impact Decomposition

Medical Care	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$	$\Delta PS_{t+1,t}$	$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Full Sample					
t_{-4}-t_{-3}	9.3%	3.4%	-0.6%	6.0%	0.5%
t_{-3}-t_{-2}	11.1%	2.0%	2.4%	6.8%	-0.1%
t_{-2}-t_{-1}	10.4%	0.2%	0.3%	8.4%	1.5%
t_{-1}-t_0	**-15.3%**	**1.2%**	**3.6%**	**-17.9%**	**-2.2%**
t_0-t_1	6.6%	1.7%	0.7%	0.7%	3.5%
Sickest Quartile*					
t_{-3}-t_{-2}	6.1%	1.1%	-0.4%	4.1%	1.3%
t_{-2}-t_{-1}	5.9%	-0.1%	-0.5%	3.5%	3.0%
t_{-1}-t_0	**-19.5%**	**0.4%**	**3.4%**	**-20.0%**	**-3.3%**
t_0-t_1	19.2%	0.0%	2.3%	9.0%	7.9%
Drugs	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$**		$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Full Sample					
t_{-4}-t_{-3}	10.1%	6.4%		3.6%	0.1%
t_{-3}-t_{-2}	6.6%	5.3%		1.2%	0.1%
t_{-2}-t_{-1}	4.2%	-0.2%		4.5%	-0.1%
t_{-1}-t_0	**-21.3%**	**-4.3%**		**-17.8%**	**0.8%**
t_0-t_1	13.9	5.3%		8.1%	0.5%

*Sickest quartile makes up, on average, 48.9% of total spending $t_{-3} - t_1$.

** For drugs, the price shopping and price index effects are combined into one price effect.

This table presents the results for our decomposition of the total reduction in medical spending from one year to the next into three effects: (i) provider price inflation index (ii) price shopping effect and (iii) quantity change effect, broken down into straight quantity reductions and the impact of substitution across types of procedures on medical spending. The second section of the table presents this decomposition for the sickest quartile of consumers. The third section presents this decomposition for drug spending.

Table VI: Potential Savings from Price Shopping

	Overall	Imaging	Preventive	Preventive w/ Diag.	Sickest 25%
t_{-4}-t_{-3}	18.3%	24.9%	11.8%	8.8%	18.1%
t_{-3}-t_{-2}	18.7%	28.1%	12.2%	10.5%	19.0%
t_{-2}-t_{-1}	21.1%	37.1%	12.4%	10.4%	21.5%
t_{-1}-t_0	**20.1%**	**34.2%**	**12.5%**	**12.0%**	**21.3%**
t_0-t_1	20.8%	37.0%	11.4%	12.5%	21.3%

This table presents the potential savings from price shopping, defined as the savings that would occur if consumers spending above the median for a given procedure reduced their spending to the median value for that procedure being offered by a different provider. Potential savings are calculated for the second-year of each two year pair.

Table VII: Spending Impact Decomposition: Potentially High Value Care

Medical Care	% Tot. Spend	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$	$\Delta PS_{t+1,t}$	$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Preventive Care,	8.2%*	-0.3%	6.4%	2.1%	-7.5%	-1.3%
General		4.1%	-1.6%	9.2%	-0.4%	-3.1%
Preventive Care,	14.5%*	-10.6%	2.0%	1.0%	-12.2%	-1.4%
w/ Prior Diag.		3.0%	2.4%	-0.7%	0.1%	1.2%
Preventive Care,	0.04%*	-1.4%	-2.0%	-0.5%	-1.6%	2.7%
Diabetics		15.9%	-1.9%	2.9%	12.5%	2.4%
Mental Health	14.11%*	-2.9%	-1.0%	0.0%	-5.4%	3.5%
		16.2%	-1.3%	0.0%	14.8%	2.7%
Physical Therapy	12.68%*	-23.8%	0.3%	7.1%	-29.7%	-1.5%
		13.5%	0.8%	3.1%	8.5%	0.9%
Drugs	% Tot. Spend	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$		$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Diabetes Drugs	3.0%**	-44.5%	6.7%		-48.0%	-3.2%
		29.1%	14.8%		12.6%	1.7%
Statins	1.7%**	-47.2%	-34.3%		-19.6%	6.7%
(for cholesterol)		14.6%	16.8%		-1.8%	-0.4%
Antidepressants	5.5%**	-48.7%	-37.4%		-18.0%	6.7%
		12.0%	0.4%		11.6%	0.0%
Hypertension Drugs	1.3%**	-27.9%	-4.9%		-24.2%	1.2%
		16.3%	3.2%		12.7%	0.4%

* % of medical spending, ** % of drug spending

This table presents our spending change decomposition for types of health care that are likely to be of high value to consumers. For each type of care, the top row presents results from the spending change decomposition moving from $t_{-1} - t_0$ while the bottom row presents these results from $t_{-3} - t_{-2}$.

Table VIII: Spending Impact Decomposition: Potentially Low Value Care

Medical Care	% Tot. Spend	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$	$\Delta PS_{t+1,t}$	$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Imaging	10.0%*	-19.5%	-0.4%	0.6%	-17.7%	-2.0%
		5.5%	2.7%	-1.9%	6.3%	-1.6%
CT Scan for Sinuses	0.1%*	-24.8%	0.5%	1.1%	-26.0%	-0.4%
w/ Acute Sinusitis		11.3%	0.4%	3.9%	5.2%	1.8%
Back Imaging for	0.3%*	-26.1%	6.9%	-6.8%	-21.3%	-4.9%
Non-Specific Low Back Pain		22.2%	4.2%	-7.6%	14.5%	11.3%
Head Imaging for	0.2%*	-23.9%	-1.0%	6.6%	-30.7%	1.2%
Uncomplicated Headache		18.0%	0.4%	-1.8%	17.9%	1.5%
Colorectal Cancer Scrng.	0.5%*	-32.2%	0.7%	-0.8%	-26.2%	-5.9%
for Patients Under 50		7.6%	1.3%	5.2%	-3.4%	4.5%
Drugs	% Tot. Spend	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$		$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Antibiotics for Acute	0.9%**	-47.8%	-6.2%		-44.4%	2.8%
Respiratory Infection		-4.8%	-5.3%		0.4%	0.1%

* % of medical spending, ** % of drug spending

This table presents our spending change decomposition for types of health care that are potentially of low value to consumers. For each type of care, the top row presents results from the spending change decomposition moving from $t_{-1} - t_0$ while the bottom row presents results from the spending change decomposition from $t_{-3} - t_{-2}$, in the pre-treatment period.

Table IX: Percentage of Savings Coming From Start-of-Month Plan Arm

	% t_0 Savings	% t_1 Savings	% Member-Months In Plan Arm, t_0
Start of Month Plan Arm			
Deductible	91%	120%	63%
– EOY Q1 (Sick)	25%	33%	
– EOY Q2	24%	30%	
– EOY Q3	19%	24%	
– EOY Q4 (Healthy)	23%	32%	
Coinsurance	-5%	-10%	32%
OOP Max	14%	-10%	5%

This table shows the % of total reduced t_0 and t_1 spending coming from consumers who start a given month in a given plan arm of the non-linear contract. The table integrates spending at the monthly level: e.g., a consumer starting February under the deductible has February spending count towards under deductible, while if that consumer starts March in the coinsurance range, March spending counts in the coinsurance category. t_0 and t_1 consumers' spending are compared to comparable quantiles of consumers' spending from t_{-2} as discussed in the text.

Table X: Non-Linear Contract Analysis: Incremental Spending Regressions

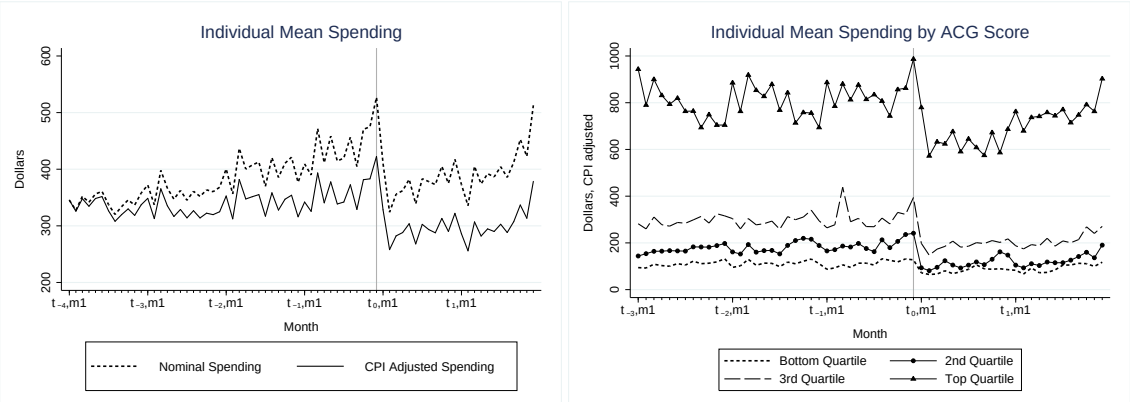
	Specification					
Variable	Primary	Shadow P Ventiles	No Prior Year MP	No Shadow Price	Fewer Controls	t_0 Only
Spot Price X Treatment Year						
1 (Deductible)	-0.422***	-0.414***	-0.434***	-0.347***	-0.525***	-0.411***
	(0.0385)	(0.0458)	(0.0384)	(0.0328)	(0.0395)	(0.0386)
1 (Deductible X t_1)	-0.0547	-0.0727	-0.0671*	0.0323	-0.0860**	–
	(0.0374)	(0.0443)	(0.0372)	(0.0318)	(0.0860)	–
0.1 (Coinsurance)	-0.144***	-0.0938**	-0.143***	-0.117***	-0.181***	-0.139***
	(0.0377)	(0.0401)	(0.0335)	(0.0325)	(0.0346)	(0.0337)
0.1 (Coinsurance X t_1)	-0.0197	-0.0416	-0.0331	-0.001	-0.0314	–
	(0.0328)	(0.0390)	(0.0326)	(0.0307)	(0.0336)	–
Shadow Price X Treatment Yr.						
Quintile 2 – [0.089,0.100]	-0.0570***	$--^a$	-0.0655***	–	-0.0773***	-0.0597***
	(0.0217)	$--^a$	(0.0214)	–	(0.0222)	(0.0219)
Quintile 2 X t_1	0.0424*	$--^a$	0.0211	–	0.0456	–
	(0.0217)	$--^a$	(0.0214)	–	(0.0223)	–
Quintile 3 – [0.100,0.2755]	-0.0424*	$--^a$	-0.0443	–	-0.0479*	-0.0564***
	(0.0255)	$--^a$	(0.0249)	–	(0.0261)	(0.0262)
Quintile 3 X t_1	0.0549**	$--^a$	0.0253	–	0.0615*	–
	(0.0260)	$--^a$	(0.0256)	–	(0.0267)	–
Quintile 4 – [0.2756,0.7303]	-0.0666***	$--^a$	-0.0381	–	-0.0715**	-0.0513*
	(0.0294)	$--^a$	(0.0285)	–	(0.0301)	(0.0311)
Quintile 4 X t_1	0.106***	$--^a$	0.0196	–	0.115***	–
	(0.0292)	$--^a$	(0.0283)	–	(0.0300)	–
Quintile 5 – [0.7304,1]	0.135***	$--^a$	0.205***	–	0.167***	0.160***
	(0.0312)	$--^a$	(0.0288)	–	(0.0320)	(0.0355)
Quintile 5 X t_1	0.0967***	$--^a$	-0.0114	–	0.109***	–
	(0.0307)	$--^a$	(0.0284)	–	(0.0315)	–
Prior Yr. End MP X Treatment Yr.						
1 (Deductible)	0.0657***	0.0509*	–	0.0948***	0.0516*	0.0607
	(0.0262)	(0.0269)	–	(0.0244)	(0.0268)	(0.0384)
1 (Deductible X t_1)	-0.0962***	-0.0822***	–	-0.0569**	-0.0786***	–
	(0.0254)	(0.0260)	–	(0.0236)	(0.0260)	–
0.1 (Coinsurance)	-0.0333	-0.0308	–	-0.0497**	-0.0471**	-0.0384
	(0.0210)	(0.0216)	–	(0.0205)	(0.0215)	(0.0310)
0.1 (Coinsurance X t_1)	-0.0159	-0.0102	–	0.0283	-0.0181	–
	(0.0205)	(0.0216)	–	(0.0200)	(0.0210)	–
Demographics & Seasonality	YES	YES	YES	YES	YES	YES
Prior Month Spend Controls	YES	YES	YES	YES	NO	YES
Health Controls	YES	YES	YES	YES	NO	YES
Observations	749,705	749,705	749,705	749,705	749,705	499,796
R^2	0.381	0.383	0.374	0.371	0.349	0.382

*** p $<$ 0.01, ** p $<$ 0.05, * p $<$ 0.10

a Shadow price ventile coefficients displayed in Table A23 in Online Appendix A10

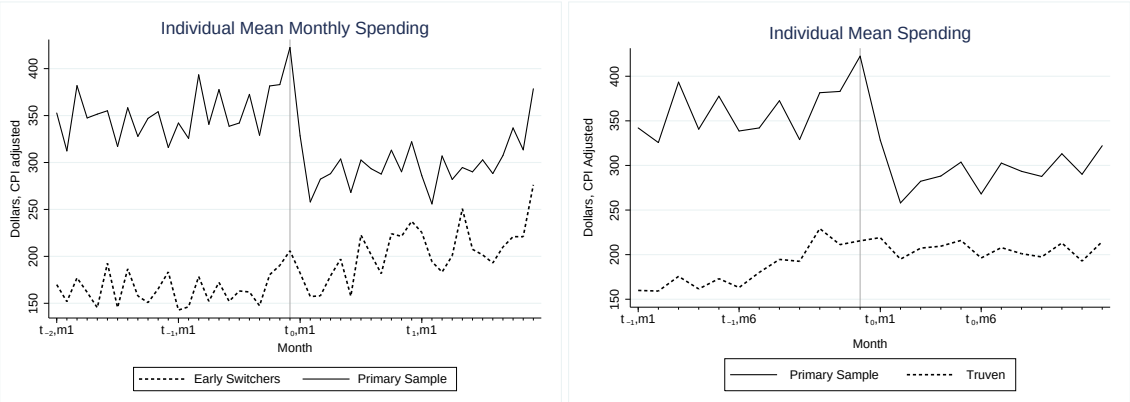
Results for regressions examining consumer responses to non-linear contract prices in the HDHP.

Figure I: Incremental Spending Time Series



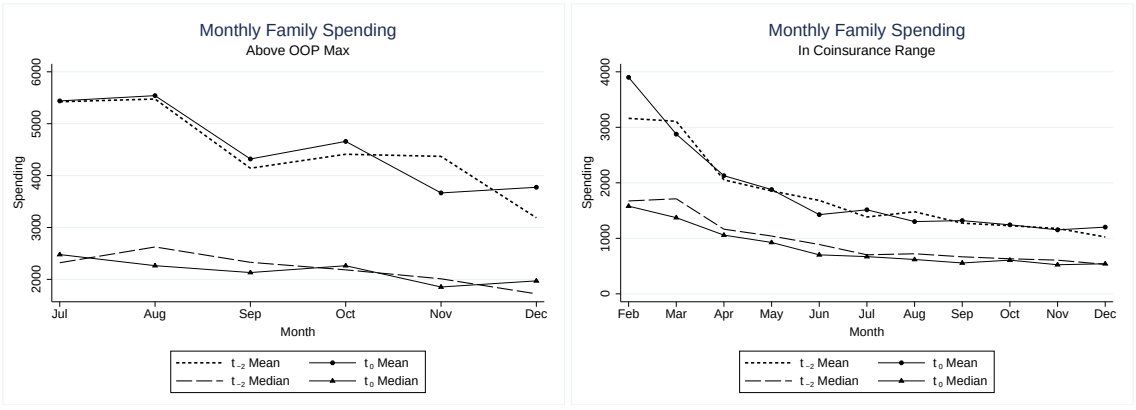
This left panel in this figure plots mean monthly spending by individuals in our primary sample over the six years in our data, both adjusted and unadjusted for age and price trends. The right panel plots adjusted spending for individuals in a given month, by ACG predictive health index quartile (the index is calculated at the beginning of each calendar year).

Figure II: Difference-in-Differences Time Series Analysis: Early Switchers and Truven Control Group



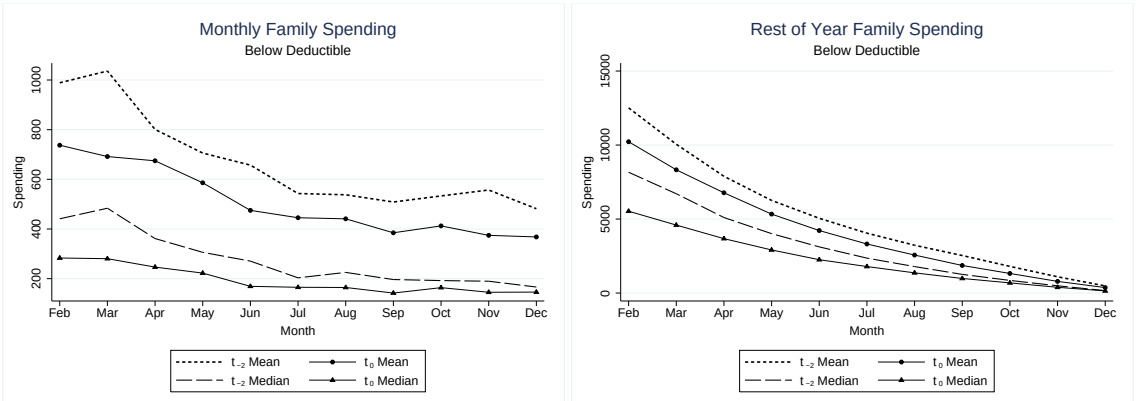
The left panel presents unweighted spending over time for early switchers to the HDHP alongside our primary sample. The right panel presents spending for our primary sample alongside spending for the weighted control group formed from Truven MarketScan data.

Figure III: Incremental Spending for Employees Over Out-of-Pocket Maximum and in Coinsurance Arm



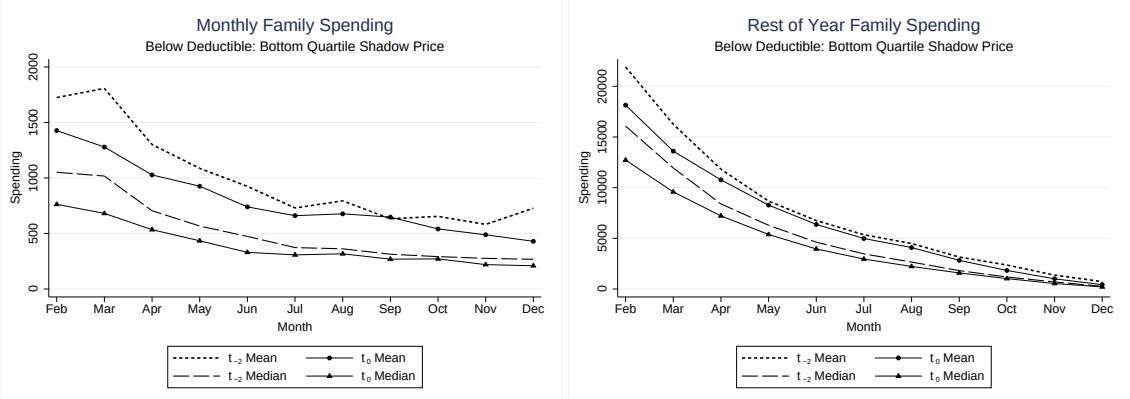
The left panel in this figure shows incremental spending for the next month, for families who have passed the out-of-pocket maximum by the start of a given month in t_0, compared to t_{-2} incremental spending for equivalent quantiles of pre-period consumers. The right panel presents the analogous figure for families who start a given month in the coinsurance arm of the HDHP (and matched t_{-2} consumers).

Figure IV: Incremental and Rest-of-Year Spending for Employees Under Deductible



This figure shows incremental spending for employees who are under the HDHP deductible by the start of a given month in t_0. The left side of the figure studies incremental spending for the next month, while the right side studies incremental spending for the rest of the year. This t_0 incremental spending is compared to t_{-2} incremental spending for the equivalent quantiles of pre-period consumers.

Figure V: Incremental and Rest-of-Year Spending for Very Sick Employees Under Deductible



This figure shows incremental spending for predictably sick (25% of ex ante sickest consumers under the deductible at the start of each month) employees who are under the HDHP deductible by the start of a given month in t_0. The left side of the figure studies incremental spending for the next month, while the right side studies incremental spending for the rest of the year. This t_0 incremental spending is compared to t_{-2} incremental spending for the equivalent quantiles of pre-period consumers.