Title
Agency and Market Efficiency in the U.S. Health Care Industry

Author
Brot-Goldberg, Zarek Chase

Publication Date
2019

Peer reviewed|Thesis/dissertation

Agency and Market Efficiency in the U.S. Health Care Industry

by

Zarek Brot-Goldberg

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Benjamin Handel, Chair
Associate Professor Jonathan Kolstad
Assistant Professor Kei Kawai

Spring 2019

Agency and Market Efficiency in the U.S. Health Care Industry

Abstract

Agency and Market Efficiency in the U.S. Health Care Industry

by

Zarek Brot-Goldberg

Doctor of Philosophy in Economics

University of California, Berkeley

Associate Professor Benjamin Handel, Chair

This dissertation studies the role of principal-agent problems as a barrier to market efficiency in the U.S. health care industry. The rise of health care as a percent of U.S. gross domestic product, as well as the documented dispersion in the productivity and price of health care, demand a policy response. In this dissertation I ask how demand-side (Chapter 1) and supply-side (Chapter 2) incentives can work to increase or decrease competitive forces and affect the efficient functioning of the market.

The first chapter, coauthored with Amitabh Chandra, Benjamin Handel, and Jonathan Kolstad, studies the role of principal-agent problems on the demand side. We study a large U.S. employer that changed their health insurance benefits from a comprehensive plan with no employee cost-sharing to a high-deductible health plan with significant cost-sharing, effectively increasing the price of health care to those employees. We find that, although employees do reduce their health care spending in response to this change, they do so in suboptimal ways. We find that (i) they reduce high-value care in similar proportion to low-value care; (ii) that reductions in spending come from reductions in utilization rather than substitution of care to lower-priced providers; and (iii) that employees do not understand the dynamic incentives embedded in high-deductible plans, so that even those employees whose effective marginal price of care has not changed cut back spending in response to deductibles. Our results suggest that although demand-side policies may reduce spending, they do so in highly inefficient ways, and may not solve the underlying causes of high U.S. health care spending growth.

The second chapter studies the role of principal-agent issues on the supply-side in determining the productivity of the U.S. health care industry. I document extensive vertical integration between primary care physicians (PCPs) and orthopedic joint surgeons. Using a stylized model, I show that this integration can increase productivity through technical production efficiencies, or lower productivity by distorting where PCPs refer their patients. I estimate that both of these effects are present, but that which effect dominates depends critically on the identities of the integrating parties. Echoing my first chapter, I find that demand is insensitive to price, lowering the potential distortionary effects of vertical inte-

gration on demand. I do find that the use of 'global budget' capitation contracts, which make PCPs share the cost of patient care, do induce the reallocation of patients towards lower-cost orthopedists. This suggests that supply-side remedies may be more effective for future policy than demand-side remedies.

To my family, because that's typically to whom one dedicates their dissertation. They did help a little bit.

Contents

Acknowledgments

It is hard to understate how important my doctoral advisors, Ben Handel and Jon Kolstad, were to the completion of this work. They have been present from the start of my research career, as their research assistant, to my move to Berkeley, to the bitter end of this degree. Beyond the obvious (their coauthorship of Chapter 1), it is hard to divorce my identity as a researcher from their influence and guidance. Ben, more than anyone, taught me how to think about research–how to be critical, and how to be interesting. Our weekly lunch meetings in my first years at Berkeley are the highlight of my graduate school experience. Jon is the one who has always kept me grounded, always pushing me to be practical, focused, and policy-relevant. Chapter 2, my job market paper, would be even more disorganized and messy than it already is without his guidance. I am eternally grateful for their support and encouragement throughout the last eight years. Kei Kawai, my third dissertation committee member, provided a much-needed skeptical eye through this work.

I thank Amitabh Chandra for his coauthorship of Chapter 1, and for getting me excited about the economics of health care so many years ago. The inspiration for my job market paper came from things he taught me back then. Mathijs de Vaan's collaboration and data support made Chapter 2 possible.

I could not have gotten through writing this without the support of friends. I am especially indebted to Jon Schellenberg, the best friend I could have hoped to accompany me through this process. I thank Kati Springel, who never stopped reminding me how much worse things could be when I struggled the hardest. I thank Jonathan Holmes, who tolerated more of my flights of fancy than anyone else. And I thank Boris Vabson, whose unbounded optimism kept me afloat.

It takes a village to write a dissertation, and there are many more to thank for their help and/or friendship, in less excruciating detail. With some omissions, I thank Ned Augenblick, Natalie Bachas, Giovanni Compiani, Joe Farrell, Richard Gilbert, Jen Gong, Sean Higgins, Abbie Jacobs, Erik Johnson, Rupal Kamdar, Jen Kwok, Juliana Londono Velez, Jordan Ou, Deepak Premkumar, Walker Ray, Yotam Shem-Tov, Avner Shlain, and Tiffany Tsai for their support.

There are a number of people whose helping hands brought me here in some way: My first economics teacher, Martha Curtis; my undergraduate advisor, Mary Hansen; and Susan Athey, whose decision to hire me at Microsoft Research opened the door into this profession.

Finally, I thank my parents, Alisa Brot and Carl Goldberg, who believed in my abilities earlier and moreso than anybody else; my grandparents, George and Sandra Brot, for their unyielding advice and unconditional love (respectively); and Jessica Vetterli, whose smile brightened the darkest moments.

And thanks to you, reader. Although, if we're being honest, you probably just read this far to see if you were mentioned in this section, right?

Chapter 1

What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Spending

1.1 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 78% 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

[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).

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 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

[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).

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 (2017) 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 et al. (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 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

[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 (2017) 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.

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.

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-

[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.

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 1.2 describes our empirical setting and data. Section 1.3 presents our treatment effect analysis of the overall medical spending response to the required HDHP switch. Section 1.4 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 1.5 presents our analysis of consumers responding to non-linear contract prices, and Section 2.8 concludes.

1.2 Setting & Data

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]

[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.

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.

The first column of Table 3.1 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 3.2 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

[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.

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 3.1, in t_{-1}, 85.2% 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.

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 1.5.

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 3.1 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 Appendix 5 presents summary 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.

[7] Table A22 in Appendix 5 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 3.2, 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.

Figure A1 in Appendix 5 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 Appendix 5 for additional detail.

1.3 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 4.1 plots mean monthly spending at the individual level for our primary sample over the six years in our data (Figure A19 in Appendix 5 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 3.3 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.

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

[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`.

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 1.4.

The left panel of Figure 4.1 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 3.3 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 Appendix 5, to compute the standard errors for all of the estimates presented in this section.

The right panel in Figure 4.1 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 1.2. 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]

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 1.5, 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 4.1 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 1.2, the t_0 HDHP switch was first announced in October t_{-3} with many regular subsequent

[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.

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 et al. (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 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. Appendix 5 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 4.1 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

[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 in Appendix 5 clearly illustrate that claim counts and median monthly spending spike in October-December t_{-1}, but not earlier in t_{-1}.

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 3.3 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 3.3 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 price responsiveness.[13] The three semi-arc elasticity estimates in Table 3.3 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 Appendix 5 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 4.2, 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

[13] As discussed in Aron-Dine, Einav and Finkelstein (2013) and shown in this paper in Section 1.5, 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.

point in time. We implement this matching at a granular level, based on ACG score ventiles: see Appendix 5 for more details.

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 3.4 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 3.3. The lower end of this range is statistically different from a 0% change at the 10% level: the standard errors for this specification is higher than for the other presented in this section because of the relatively small size of the early switcher sample (approx. 12,000). See Appendix 5 for additional figures and details on this early switcher difference-in-differences specification.

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 4.2 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

[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 Appendix 5.

[16] See Appendix 5 for more details on this weighting procedure. The Appendix contains an additional exercise that weights the Truven sample by income as well as by age and gender. We include this in the 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.

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 3.4 quantifies the relative 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 Appendix 5 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 3.4 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 Appendix 5 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 1.4 dives deeper into spending reductions for specific services, and whether those reductions are achieved via changes in prices paid or quantities consumed.

1.4 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 (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).

[17] In Appendix 5, we replicate this analysis including income.

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 Appendix 5.

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. Appendix 5 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:

$$\Delta PPI_{t+1,t} = \frac{\mathbf{P}_{t+1} \cdot \mathbf{C}_t - TS_{t,t}}{TS_{t,t}} \tag{1.1}$$

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 Appendix 5.

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:

$$\Delta PS_{m,t+1,t} = \frac{\mathbf{P}_{m,J^m,t+1} \cdot \mathbf{C}_{m,J^m,t+1} - \mathbf{P}_{m,J^m,t+1} \cdot \mathbf{C}_{m,J^m,t}}{\mathbf{P}_{m,J^m,t+1} \cdot \mathbf{C}_{m,J^m,t}} \tag{1.2}$$

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:

$$\Delta PS_{t+1,t} = \Sigma_{m=1}^{M} \frac{Y_{m,t}}{Y_t} \Delta PS_{m,t+1,t} \tag{1.3}$$

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

[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.

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:

$$\Delta Q_{t+1,t} = \frac{\sum_m \mathbf{C}_{m,J^m,t+1} - \sum_m \mathbf{C}_{m,J^m,t}}{\sum_m \mathbf{C}_{m,J^m,t}} \tag{1.4}$$

$\Delta Q_{t+1,t}$ 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 accounted 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:

$$\Delta QS_{t+1,t} = \Delta TS_{t+1,t} - \Delta PPI_{t+1,t} - \Delta PS_{t+1,t} - \Delta Q_{t+1,t} \tag{1.5}$$

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 (2017) 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.

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 3.5 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 1.3 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.

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,

[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.

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 3.6 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.

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, 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

[20]For robustness, in Appendix 5 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.

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 3.5 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 Appendix 5 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 3.5 presents the same decomposition for the sickest quartile of consumers in the population. As shown in Section 1.3 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 3.5 may overstate the total spending reduction and total quantity reduction. Section 1.3 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 et al. (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 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 Appendix 5 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 3.7 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

[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 Appendix presents summary statistics for this analysis across all 30 procedures. Appendix 5 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.

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 et al. (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 $t_0 - t_1$. 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.

Appendix 5 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,

[22]See Tables A15 and A16 in Appendix 5 for a complete time series of these decompositions.

[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.

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 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.

Table 3.7 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}). Appendix 5 provides more detail on this analysis, including the full time series of results for each service.

Table 3.8 presents our spending change decomposition for a collection of low-value services. Schwartz et al. (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 Appendix 5 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 et al. (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 3.6 shows potential for 34.2% savings from price shopping if above median costs are reduced to median costs. See Appendix 5 for more detail on these results.

1.5 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 et al. (2015), Dalton et al. (2015), and Abaluck et al. (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 et al. (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 et al. (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.

[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.

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^s_m:** 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.
- **Shadow Price / Expected Marginal EOY Price, $P^e_m = E_m[P^s_{EOY}|S_m, H, X, Ins_m]$:** 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^L_m:** 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^s_m is straightforward for each consumer and each month by mapping S_m to the corresponding non-linear contract spot price. Computing P^L_t 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^e_m we use the following process:

[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.

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

[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.

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_m^e 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 Appendix.

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 4.3 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

[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 1.3. 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.

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.

The right panel in Figure 4.3 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 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.

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 4.5.

Figure 4.5 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 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

[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).

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 Appendix for these additional results.

Table 3.9 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 Appendix 5 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:

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

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 θ_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 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 3.10 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

[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 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.

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.

Table 3.10 also presents five regressions to assess the robustness of our primary specification. The results in these alternative specifications, described in more detail in Appendix 5, are similar to those from the primary specification just presented. Additionally, Appendix 5 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 et al. (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.

1.6 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 et al. (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 et al. (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.

Chapter 2

Intermediation and Vertical Integration in the Market for Surgeons

2.1 Introduction

Physician agency plays a central role in demand for health care. Patients typically do not know what care to get or where to get it, and therefore often must rely on a primary care physician (PCP) to serve as their agent. Given this role, the incentives that PCPs face can alter treatment decisions and fundamentally shape health care markets (Arrow 1963, McGuire 2000). One critical role the PCP plays is as an intermediary to specialty care providers. A referral from their PCP is second only to insurance coverage in determining where patients receive specialty care (Ziemba, Allaf and Haldeman, 2017), and, for some health insurance plans, an explicit referral is a formal prerequisite for coverage. Given the substantial dispersion in the price of care across providers (Wennberg, 1996), even conditional on quality (Baicker, Chandra and Skinner, 2012), a good referral is valuable: Going to the 'wrong' specialist could increase a patient's final expenses by thousands of dollars. Although the PCP is ostensibly the agent when they make referrals, their decisions also affect the profits of the specialty care providers they send patient to, making control over their referral patterns a highly-prized asset for specialists.

The primary way in which specialty care providers have gained control of PCP referral patterns has been to form large vertically-integrated health systems. This impetus has led to an extensive degree of consolidation in U.S. health care (Kocher and Saini 2011, Capps et al. 2017). Assessing the welfare consequences of consolidation requires policymakers to consider a central trade-off between productive efficiency and allocative efficiency. Vertical integration can improve the productive efficiency of care, through provider coordination and performance incentive provision (Burns and Pauly, 2002). However, it may also distort PCPs' incentives to serve as good agents for patients, by encouraging them to engage in self-dealing, steering the allocation of referrals towards in-house specialists even when those specialists are less efficient than external ones. These forces may also interact in complex ways: A highly-efficient integrated organization may improve welfare by steering patients who would otherwise seek care at inefficient providers.

This setting mirrors a broad trend across many industries of large firms that both supply goods and own intermediation platforms that direct consumers to those goods. Amazon, for example, both owns the largest U.S. internet shopping platform and distributes a number of consumer goods. Understanding the antitrust implications of such arrangements requires authorities to balance productive efficiencies against anticompetitive effects. Courts have lacked a unified framework to do so, with antitrust cases on intermediation resulting in starkly different rulings in the E.U.[1] and the U.S..[2]

This paper explores the welfare effects of vertical integration in health care. Doing so

[1] E.g., in 2017, the European Commission fined Google $2.7 billion for steering search consumers towards Google Shopping over competitors. The ruling primarily concerned the effect on exclusion of competition in quantity terms, rather than in efficiency or consumer welfare terms.

[2] E.g., in *Ohio v. American Express Co.*, the Supreme Court ruled that contract terms between American Express and merchants that controlled merchants' ability to steer consumers between cards were permissible, since plaintiffs had not successfully proven damages to both consumers and merchants.

requires that we separately identify the extent to which integration generates productive efficiencies from the extent to which it enables anticompetitive actions such as steering. We do so by studying referrals from PCPs to orthopedic joint specialists in Massachusetts, a state dominated by a number of large integrated health systems. We combine an administrative dataset containing the near universe of medical claims from private health insurers in Massachusetts with novel data measuring physician vertical affiliations. We observe that vertical integration is even more pervasive than previously suggested, with nearly every PCP and orthopedist sharing a vertical tie with at least one member of the other group. Nearly two-thirds of referrals are to integrated orthopedists.

We start by developing a simple model of PCP referral behavior and cost outcomes, that incorporates incentive provision both by integrated systems and by health insurers. We show that the effect of vertical integration on cost outcomes is ambiguous, and depends on the relative magnitudes of productive efficiencies and steering incentives, as well as unobservable competitive substitution patterns and the costliness of affiliated specialists. We show that a simple comparison of referral volumes and cost outcomes at integrated PCPs compared to unintegrated PCPs will not be sufficient to separate estimate these factors, so we must instead model referral choice and cost outcomes separately.

We begin by measuring heterogeneity in cost outcomes among orthopedists. We define the 'cost' of an orthopedist as their effect on total spending on health care incurred in the year following a patient's first visit. We document substantial cost dispersion. Moving a patient from the average orthopedist to one who incurs one standard deviation greater expenses would increase expected costs by nearly 30%, with estimates of orthopedist-specific effects ranging from 50% cost reductions to 85% cost increases relative to the mean orthopedist. We find that vertical integration does indeed generate efficiencies, reducing expected costs by nearly 6%.

We then examine how the allocation of referrals is determined by PCP incentives. In the time period we study, Massachusetts insurers introduced their own incentives to encourage PCPs to contain costs, in the form of "global budget" capitation contracts. These contracts force PCPs to bear a share of the cost of their referral choices. Using panel variation in new data on the use of capitation contracts across insurers (as in Ho and Pakes 2014), we find that they induce PCPs to refer patients to orthopedists who incur 6.1% lower expected one-year costs. Consistent with our model, PCPs in integrated health systems that control high-cost orthopedists respond by engaging in slightly less self-dealing, while PCPs in systems with low-cost orthopedists increase their self-dealing substantively. We interpret these results as substantive reduced-form evidence that PCP incentives drive referrals to specialists.

This motivates the estimation of the parameters of a structural model of referral choice. The model incorporates patient preferences, vertical efficiencies, and PCP incentives. In a counterfactual simulation, we find that removing efficiencies would have virtually no effect on self-dealing, whereas removing non-efficiency-driven PCP preferences for referring internally would cut self-dealing by slightly more than half. This result implies a staggering amount of anticompetitive steering. Counterintuitively, removing vertical ties between PCPs and orthopedists would nonetheless *increase* expected costs on average, in a partial-equilibrium counterfactual where we hold orthopedist costs fixed. These seemingly contradictory results are generated by the fact that PCPs' sensitivity to expected costs when making referrals is indistinguishable from zero. When there is integration, PCPs are steered internally to take advantage of efficiencies that they would be insensitive to otherwise. Moreover, steering efforts by low-cost systems seem to offset the same behavior by high-cost systems. We find that the introduction of global budget contracts does succeed in introducing cost-sensitivity into orthopedist referrals. The amount of competition induced is around two-thirds of the level needed to fully offset the loss from dis-integration.

Our results suggest a nuanced approach to evaluating vertical integration. In the current status quo, allowing consolidation may be a second-best policy, since it does generate productive cost efficiencies and can even improve allocative efficiency when low-cost specialists integrate. This is only the case, however, because status quo cost competition is so weak. Introducing policies to improve competition would preclude the need for integration. However, even the high-powered global budget contracts we observe are not quite enough to generate such competition. A policy would need to provide 42% stronger incentives than the average capitation contract we observe. However, high-powered incentives that shift substantial risk onto PCPs require paying a substantial risk premium (Holmström, 1979). These risk premia may only be affordable at integrated organizations that can use their large volume to smooth patient risk.

This paper contributes to several distinct literatures. First, we contribute to a broad literature on productivity dispersion and misallocation in health care, as well as studies of policies attempting to ameliorate that misallocation. Since the release of the Dartmouth Atlas of Health Care (Wennberg, 1996), a vast body of work has documented extensive variation across the U.S. in the cost of care for observationally identical patients, even within narrow categories of care providers. This variation cannot merely be explained by quality variation – as Baicker et al. (2012) find, the relationship between average risk-adjusted spending and mortality for a given hospital are uncorrelated. Our work follows in the vein of recent work by Finkelstein, Gentzkow and Williams (2016), Cooper, Craig, Gaynor and Van Reenen (forthcoming), and Chernew, Cooper, Larsen-Hallock and Scott Morton (2018), who find that supply-side factors explain a significant share of spending variation. The latter two in particular show that horizontal market structure and referral patterns can explain allocation of patients to high-cost medical providers. We expand on this by showing that vertical market structure also plays a large and understudied role. Additionally, prior work by Brot-Goldberg, Chandra, Handel and Kolstad (2017) and Sood, Wagner, Huckfeldt and Haviland (2013), among others, shows that demand-side cost-sharing does not work as a remedy for misallocation. We show that *does* improve cost-sensitivity, although its efficacy is small relative to cost variation.

Second, we contribute to the literature on vertical integration. Since Coase (1937), a long literature has explored the purposes of vertical integration and its ramifications for competition and welfare. The transaction cost economics theory of Williamson (1985) and Klein, Crawford and Alchian (1978) and the agency costs theory of Jensen and Meckling (1976) and Holmström and Milgrom (1994) suggest that integration allows principals to overcome contracting frictions. The implication of this work was that vertical integration is generally a force for consumers' good, since it allows merging firms to make relationship-specific investments that might be infeasible otherwise. Alternatively, agency cost reductions may allow actors within the firm to coordinate in ways that are detrimental to consumers. Integrated firms can arrange fully or partially exclusive deals between parts of the firm which foreclose on rival firms' ability to deal with their buyers or suppliers. Hart and Tirole (1990) and Ordover, Saloner and Salop (1990) discuss the theoretical case for how such effects might arise. The literature is divided, finding evidence for both the efficiency benefits (e.g. Forbes and Lederman (2010), Atalay, Hortaçsu, Li and Syverson (2017)) and competitive harms (e.g. Chipty (2001), Hastings and Gilbert (2005)) of vertical integration. Some studies, such as Hortaçsu and Syverson (2007), even suggest that integration may be welfare-neutral. This literature includes a number of papers on integration specifically in health care, where results have largely been more negative. Researchers have found that integrated specialty care providers have higher prices,[3] that integrated primary care practices have higher prices,[4] and that physicians tend to steer patients to facilities they have a vertical tie with.[5]

We build on this large body of work in a number of ways. First, in contrast to much of the health literature, we are able to separate the efficiency and foreclosure effects of integration. Our model suggests that prior results merely estimating reduced-form impacts on incurred prices and volumes may not be informative about the efficiency impact of integration. Second, in contrast with the broader empirical vertical integration literature, we study a setting with many imperfectly-competitive strategic firms both upstream and downstream, and show that the efficiency consequences of integration depend on what firms are integrating. This may explain the fact that efficiency effects in prior work have been wide-ranging. Finally, we show that when a firm acquires its own intermediaries, this can have negative effects even when, as in our model, prices do not change. This arises when high-cost firms buy referrals from their intermediaries, reducing allocative efficiency as is done in our setting. This is distinct from the potential for vertical integration to cause harm through excluding rivals or by raising equilibrium prices, and has been less-studied in the empirical literature on integration, although a similar point has been made in the broader literature on imperfect agency.[6] This channel is critical for understanding integrated platforms like Amazon and Google, where the 'downstream' intermediary does not charge prices to consumers.

Finally, we contribute to the literature on capitation and other forms of supply-side regulation in health care. This literature was spurred by the rise of managed care in the 1980s and 1990s, which combined vertical coordination with capitation contracts. Glied (2000)

[3]See Cuellar and Gertler (2006) and Baker, Bundorf and Kessler (2014) for hospitals and Baker, Bundorf and Kessler (2017) for specialist physicians.

[4]See Capps, Dranove and Ody (2018).

[5]E.g. Swanson (2013), Baker, Bundorf and Kessler (2016), and Chernew et al. (2018). Similarly, a handful of papers find that when upstream specialty care providers acquire downstream 'feeder' providers, they experience increased patient volumes. This includes Nakamura, Capps and Dranove (2007), Nakamura (2010), and Walden (2016).

[6]E.g. Afendulis and Kessler (2007), Barwick, Pathak and Wong (2017) and Egan (2018).

summarizes the early literature, which found that managed care reduced costs but often could not assess through what channel. The recent rise of Accountable Care Organizations (ACOs), which have similar features (Burns and Pauly, 2012), has led to a resurgence in this literature. Although some work in Massachusetts has found that capitation reduces costs (Ho and Pakes 2014, Song et al. 2011, 2014), results in general have varied wildly across participating health systems, with some generating large savings, and others generating spending increases (Colla, Wennberg, Meara, Skinner, Gottlieb, Lewis, Snyder and Fisher, 2012). In contrast to this recent work, we model both capitation *and* vertical integration, the two defining traits of ACOs, jointly. Our results suggest that health system identity is an important determinant of costs, and its mediation of the effects of incentives may help to explain why ACOs have been successful in some places but not others.

The rest of the paper proceeds as follows. In Section 2.2, we describe the institutions we study and the data we use. In Section 2.3, we write down a model of orthopedist referral choice that depends on incentives and potential cost outcomes. Section 2.4 presents our estimates of the extent of cost dispersion across orthopedist and vertical cost efficiencies. In Section 2.5 we present reduced-form evidence for how referral patterns depend on integration and global budgets. In Section 2.6, we estimate a structural model of referrals, while in Section 2.7 we present the results from counterfactual policy simulations based on these estimates. Section 2.8 concludes.

2.2 Setting & Data

Setting: Referrals to Orthopedic Surgeons

In this paper, we choose to focus on referrals to specialty care. Improving the choice of site of specialty care is an important part of reducing health care costs. Prior work (e.g. Chandra and Staiger (2007), Baicker et al. (2012)) has shown that the cost of care varies across healthcare providers with the same level of quality. Shifting patients across providers from high- to low-cost is a potentially fruitful way of reducing costs. Brot-Goldberg et al. (2017), for example, find that moving patients from above-median cost providers to median cost providers would generate savings of nearly 20% in their setting. Finding a way to implement this move through patient incentives, however, has been challenging. This may be because choice of specialty care is not driven by patient choice, but by the referral patterns of their PCP. A patient survey by Ziemba et al. (2017) found that PCP referrals were the second-most important factor in surgeon choice, second only to whether the surgeon accepted the patient's insurance. This should not be surprising. The typical design of HMO-style insurance plans often requires a patient's officially-designated PCP to sign an approval form before the insurer will cover specialty care, so for such patients a referral is mandatory. But even for patients not restricted in such a way, searching for specialty care providers is challenging. Public quality data is scarce and difficult to interpret for non-experts, and public cost data is only sometimes available, not necessarily correct, and not highly-used

(Brown, 2018).

We specifically analyze orthopedic surgeon choice, focusing on joint specialists. Orthopedists deal with musculoskeletal conditions and diseases, with joint specialists focusing largely on arthritis and other sources of general joint pain. Orthopedics is a particularly important specialty in the U.S., with spending on musculoskeletal issues making up nearly 8% of U.S. medical expenditures and nearly 1.3% of annual GDP (United States Bone and Joint Initiative, 2015). This high spending level has made orthopedics the medical specialty with the second-highest annual income in 2018 (Medscape, 2018), second only to plastic surgeons. We focus on joint specialists. Orthopedists can practice in one of a number of subspecialties, including joints as well as necks, spines, and feet. Orthopedists of a given subspecialty are not substitutes for one another.

Joint surgery has been a major target of Medicare cost and quality maintenance efforts, with both total hip arthroplasty and total knee arthroplasty being included in Medicare incentive programs and health care delivery innovation initiatives. The orthopedics patient is seen in a non-emergency setting, for a chronic condition, thus making formal referrals more common. Moreover, orthopedists have a fair amount of discretion over treatment decisions for a given patient who is experiencing joint pain. One option is to perform surgery, typically a total replacement of a joint with a prosthesis. Such a procedure is done in an inpatient setting, although inpatient recovery times are now relatively short, thanks to recent technological advances. The other option is to engage in non-surgical pain management, either through the use of pharmaceuticals or through the use of corticosteroid injections, which introduce anti-inflammatory medicine directly into a joint to reduce pain.

Vertical Integration in Health Care

The health care industry exhibits a number of organizational forms, ranging from outright ownership of practices as part of a tightly-organized firm to more informal collaboration agreements across practices. Much of the literature has discussed integration without formally describing organizational form, which has led to variation in definitions. Afendulis and Kessler (2007), for example, define vertical integration as a single physician who provides two goods in a vertical supply chain, in their case being diagnosis and treatment. In contrast, Capps et al. (2018) define vertical integration as hospitals' outright ownership of physician practices. We follow Capps et al. but use a broader definition: We define vertical integration as an organization made up of medical providers who supply primary care services *and* medical providers who supply secondary care services. This nests both the Capps et al. example of hospitals acquiring physician practices, as well as health systems like Atrius Health in Massachusetts, who include multispecialty physician practice groups but have no hospital facilities, as well as broad systems like Partners Health Care in Massachusetts, which owns both hospitals *and* physician practice groups that are not directly affiliated with any hospital.

This definition allows us to describe what changes when a physician is a part of an integrated system as opposed to when they are not. The vast literature on the 'make or buy' question, starting with Coase (1937), has asked why firms bring together multiple parts of the supply chain under one formal organizational structure, rather than simply undertake joint tasks at an arm's length, through contracts. This literature has spawned a number of theories for what is done differently in firms as opposed to outside of them. For our purposes, however, this question is less puzzling, as the difference arises out of explicit regulations.

Those regulations are the Anti-Kickback Statute (AKS) and the Stark Laws, two sets of laws that restrict the ability of physicians to contract with one another outside of firms. The AKS, passed in 1977, outlaws the practice of compensating physicians, both in money and in kind, knowingly and unknowingly, in exchange for referrals to other health care providers.[7] This means that, for example, an orthopedist cannot agree to share patient profits with a primary care provider who refers those patients.[8] The Stark Laws strengthened these provisions, barring physicians from making referrals to any entity (e.g. imaging facilities, hospitals, physician practices) in which that physician has any sort of financial stake, even if there is no explicit payment to that physician for referrals. Courts have interpreted these laws quite broadly, making it virtually impossible to write contracts along the vertical supply chain that involve financial transfers.

The Stark Laws contain a handful of 'safe harbor' exceptions that allow for referrals to coexist with financial arrangements. The most important one is that a physician can engage in referrals to an entity that they have a financial stake in if that financial stake is a "bona fide employment relationship." Thus, if a primary care provider is employed by or contracted by a health system, they can refer patients to orthopedists within the same system legally. However, payments must be at "fair market value," and the system still cannot pay for referrals. In practice, however, integrated systems can hide referral incentives within other physician performance incentives. A recent lawsuit filed by a urologist against Steward Health Care in Massachusetts claimed that Steward engaged in many such practices to punish him for not doing enough referring to Steward facilities. These included soft incentives, such as disciplinary action and public shaming, as well as hard incentives, including withholding a $25,000 bonus, culminating in his eventual termination. (Kowalczyk, 2018) Steward's lawyer asserted that these practices are "legal and extremely common."

The implication of the AKS and Stark Laws are that integration has explicit ramifications in U.S. health care that are not as obviously present in other settings: Integrated firms allow for incentive contracts to be written between parts of the supply chain where they would be illegal otherwise. This means that orthopedists within a system can indeed share profits with PCPs that arise from referrals, albeit indirectly through system transfers. Given that 9% of physician compensation in 2014 was from incentive payments (Medical Group Management Association, 2014), systems have broad scope to induce such transfers.

[7]The AKS only explicitly restricts self-referrals that result in federal reimbursement. However, this has been interpreted by the courts to cover any service that is reimbursed by Medicare or Medicaid, *even if it was paid for by a non-governmental party.*

[8]The AKS's passing was driven by the rise of arrangements like this between facilities and physicians, coupled with the practice of such physicians making unnecessary referrals to those facilities to capture reimbursements from the newly established Medicare and Medicaid programs.

Therefore, integration can induce steering of patients to within-system providers, even when this steering is against the patient's best interest. The ability to induce steering has been cited by critics as a major potential reason for why health systems have acquired primary care practices at an increasing rate (Capps et al., 2018).

It is important to note that the ability to write incentive contracts can induce positive benefits as well as the aforementioned negative ones. As in Klein et al. (1978), integration can solve quality incentive provision problems that would be difficult to resolve outside of the firm, particularly in this setting where contracting is legally difficult. For example, a health system with specialists may want the PCPs who handle their patients to adopt information technology, in order to facilitate coordination and transfer of patient files. However, IT adoption is costly and this coordination may not have enough private benefit to the PCPs to generate adoption. Under integration, the system can write a contract that helps the PCP internalize the positive externalities that their IT adoption would generate, which may improve patient care and reduce costs. There

These coordination benefits are the typical ones provided by health systems for why they choose to integrate, in contrast to the potentially anti-competitive steering benefits (Burns and Pauly, 2002). Which of these two benefits dominates is an empirical question that we will attempt to answer from Section 2.5 on.

Global Budget Capitation Contracts

The main policy instrument we study in this paper is the global budget capitation contract. Although the terms 'capitation' and 'global budget capitation' are used to describe a number of similar but distinct provider compensation schemes, we focus on a particular type of contract for this paper. The contracts we study are formed between a group of physicians (which may be a single practice, a group of practices, or a large health system containing many practice groups) and an insurer. The terms of these contracts apply to the set of patients who are insured by the insurer, and for whose care the physician group agrees to manage. Such contracts are typically only applied for patients in HMO plans, where a patient must have an official designated primary care physician, as opposed to patients in PPO plans who typically do not.

These contracts have two primary features: Global budgets and capitation. They generally take on the following form: Each year, the insurer sets a target 'global' budget for all of the patient's spending, including care provided by the contracting physician group and by other providers the patient sees. Throughout the year, the insurer reimburses medical providers who service the patient, both those who are party to the contract and those who are not, at typical reimbursement rates. At the end of the year, the insurer computes the total medical expenses incurred by the patient for the year at all providers. If that amount is below the budget, the insurer pays a share of the difference between the budget and the realized spending to the physician group. However, if the patient's spending exceeds the budget, the physician group must pay the difference to the insurer.[9] Often, these contracts also include some sort of lump-sum payment from the insurer to the physician group, as well as incentive payments to ensure high quality of care. The total losses are potentially unbounded, and so the insurer often additionally requires that the physician group hold reinsurance to insure against the extreme tail risk they are exposed to. In Figure 4.6, we visually display a hypothetical capitation contract and the payoffs it provides to PCPs as a function of patient spending.

This contractual form forces the physician group to bear some share of the cost (to the insurer and/or patient) of their treatment decisions. For example, if a PCP covered under a global budget contract opts to send a patient to an expensive orthopedist instead of a cheaper one, she will have to pay some share of the cost difference when the budget is reconciled at the end of the year. This explicitly solves the 'moral hazard in search' problem described in the introduction: PCPs under these contracts now have greater incentive to bear search costs so that they can find lower-cost specialists to refer to, in order to either capture savings or incur less penalties, with the incentive to do so increasing in the share of risk that is born by the PCP. Ellis and McGuire (1986) note that the design of these contracts mirrors the design of cost-sharing incentives for patients, inducing the trade-off described by Zeckhauser (1970): High-powered incentives for cost control reduce moral hazard at the cost of forcing risk-averse PCPs to bear financial risk for the component of patient spending which is out of their control.

A clear potential outcome of capitation schemes like this is the scope for reductions in care quality, including movements from high-price, high-quality medical providers to low-price, low-quality providers, or to no care at all. Despite this, Cutler (1995) finds virtually no change in care quality as a result of Medicare's introduction of prospective payment (an episode-based capitation scheme) for hospital reimbursement. In our setting, the incentives to cut back on quality are lower for two reasons. First, in the short run, if reductions in quality are likely to induce complications or hospitalizations, the PCP will bear the cost of sending patients to a low-quality provider in the form of sharing the costs of these adverse events. Second, in the longer run, contracting is done at a longer horizon.[10] To the extent that patients remain with the same insurer and PCP, reduced quality today may translate into increased shared costs in future time periods.

The broad use of global budget contracts originated with the rise of managed care insurance plans in the 1980s and 1990s. The managed care backlash of the 1990s, however, moved patients towards less-restrictive PPO insurance plans, which primarily reimburse physicians on a fee-for-service basis. This remained the state of affairs until the passage of the Affordable Care Act in 2010 codified and subsidized a number of so-called alternative payment mechanisms, such as Accountable Care Organizations. This spurred adoption of such mech-

[9] This describes 'two-sided' contracts. Some arrangements are 'one-sided' in that the insurer pays the physician group if spending is below budget, but the physician does not pay the insurer if spending is above budget, and instead no transfer is made. In our setting, two-sided contracts are more common.

[10] The largest insurer in Massachusetts, Blue Cross Blue Shield of MA, entered into global budget contracts with physicians on a five-year basis.

anisms by private insurers, as well. In 2018, after eight years, nearly 33 million lives were covered by such an arrangement (Muhlestein, Saunders, Richards and McClellan, 2018). In Massachusetts, global budget contracts came to popularity beginning with Blue Cross Blue Shield of MA's Alternative Quality Contract in 2009. Positive preliminary results (Song, Safran, Landon, He, Ellis, Mechanic, Day and Chernew, 2011) encouraged other insurers and providers to experiment as well, leading to substantial adoption in the time period we study. However, the exact long-run benefits are still unknown.

Data

Our primary dataset is the Massachusetts All Payer Claims Dataset (APCD), Version 4.0, from the Massachusetts Center for Health Information and Analysis (CHIA). Massachusetts state law required all commercial insurers to submit data on every processed health care claims to CHIA. Version 4.0 covers the years 2010 to 2014.[11] The APCD is extremely comprehensive: Out of 6.66 million individuals in Massachusetts in 2012, it contains data for 6.47 million. This includes a number of diverse health insurance products, such as employer-sponsored health insurance, insurance purchased on the individual and small business state exchanges, Medicaid, Medicare Advantage, and supplemental insurance Due to the broad coverage of the APCD, for a given orthopedist, we observe the near-universe of their patients. The largest two populations we miss are beneficiaries covered by public Medicare Parts A and B, and the uninsured.

Although the APCD contains a number of components, we only use the medical claims component. This contains all line-item claims incurred by all covered health insurance enrollees and their dependents. It includes the total payment made by both the insurer and the beneficiary, the identity of the medical provider who submitted the claim, as well as detailed codes indicating the diagnosis and procedure associated with the claim. The APCD includes a personal identifier that links beneficiaries longitudinally across the dataset even if they switch insurance plans, even across insurers, as well as a number of demographic details about the beneficiary, such as age, gender, and residential zip code. This data has been used in a number of prior papers, such as Ericson and Starc (2015), and is very similar in structure to APCDs from other states, including Colorado (Liebman, 2018), New Hampshire (Brown, 2018), and Utah (Handel, Holmes, Kolstad and Lavetti, 2018).

We link the APCD to a number of datasets containing information on both PCPs and orthopedists. We use the National Plan and Provider Enumeration System (NPPES) to link physicians to their business and practice addresses, and their medical specialties, on the physician's National Provider Identifier (NPI). We use historical snapshots of the system for each year to capture the relevant address at a given time. We link orthopedists to quality scoring data created by ProPublica. This data uses Medicare data from 2009-2013 to estimate risk-adjusted complication rates for hip and knee replacements performed by

[11] After the Supreme Court ruling in *Gobeille v. Liberty Mutual*, self-funded ERISA health insurance plans were no longer required to submit data. This ruling was passed in March 2016, postdating the years our data covers, and therefore does not affect our data quality.

the orthopedist. These rates serve as our primary measure of surgeon quality. We measure this externally, rather than using the APCD, because Medicare beneficiaries are much more likely to receive one of these surgeries than the relatively younger population in our sample.

We measure physicians' organizational ties using the Massachusetts Provider Database (MPD), a dataset created and maintained by Massachusetts Health Quality Partners (MHQP). The MPD matches to physician NPI, and contains detailed information on the practice(s), medical group(s)[12], and physician contracting network(s)[13] that the physician belongs to. We note that belonging to a physician contracting network indicates affiliation but not necessarily direct ownership or employment. The MPD is collected as a byproduct of insurer risk contracting, which assures its accuracy above similar survey datasets, and makes it well-suited for our analysis of risk contracts. For more detail, see Massachusetts Health Quality Partners (2017).

Finally, we measure the presence of global budget capitation contracts using auxiliary data from CHIA. From 2012-2014, CHIA collected data from insurers on their use of alternative payment mechanisms. In these years, insurers primarily paid physicians through either fee-for-service arrangements or global budget contracts.[14] The CHIA auxiliary data contains data on the total number of beneficiary-months covered under different arrangements, for each combination of year, insurer, market segment (e.g. Medicare Advantage, employer-sponsored), and plan type (HMO or PPO). This generates 97 groups, for which we can compute an expected probability of being covered under a global budget contract for beneficiaries of insurance plans in each group. Table 3.13 presents summary statistics for global budget contract utilization across our primary sample. As expected, global budget contracts are almost exclusively used in HMO plans, where beneficiaries typically have an officially-designated PCP, and almost never used in PPO plans.[15] We note that there is significant variation across HMO plans sold by the three largest insurers, with Blue Cross using them heavily compared to the moderate use by HPHC and Tufts. There is also a secular increase in their use across time.

[12] A medical group, in the MPD, nests practices. MHQP defines it as "A "parent" provider organization that may include multiple practices and practice sites. These can be single specialty or multi-specialty organizations and may exist within a broader network structure." Examples include groups such as the Brigham and Women's Hospital Physician Organization and the Cambridge Health Alliance.

[13] A physician contracting network, in the MPD, nests medical groups. MHQP defines it as "An organization of medical groups and/or practice sites with an integrated approach to quality improvement that enters into contracts with payers on behalf of its provider members." Examples include Partners Community Health Care and Atrius Health.

[14] Other arrangements make up less than 1% of reported beneficiary-months.

[15] Only two of our PPO cells report nonzero usage of global budget contracts. The insurer-market segment pairs report zero usage of these contracts in other years, so we infer that this is a reporting error and treat these shares as zero.

Sample Selection

Our primary goal is to measure referral behavior, which we define as a hand-off from a PCP to a specialist. This goal guides the construction of our primary analysis sample.

We first begin by defining our sample of specialists of interest. We define a joint surgeon as any individual physician listed in NPPES with a specialty of orthopedic surgery,[16], whose practice address is located in Massachusetts, and who we observe performing at least five total hip and/or knee replacement surgeries over the full course of our data.[17]

We then build our sample of referrals. We do this by finding all individuals treated by one of our sample of orthopedists, and finding their first office visit with that surgeon.[18] We index all analysis relative to this visit. We drop from this search process any claims covered by a non-primary insurance plan, any claims covered under a plan from a non-standard insurance market segment (such as Tricare or worker's compensation plans), and any patients under the age of 18. After these restrictions, we end up with a sample of 222,380 individuals in the years 2012-2014.

Next, we assign individuals to their relevant PCPs. For each patient, we look at all medical claims filed on their behalf in the twelve months prior to their orthopedist visit. We assign them a PCP according to the physician with a primary care specialty (general internal medicine, or family medicine) with the highest number of office visit claims in those twelve months. This follows the standard in the literature (see e.g. Agha, Ericson, Geissler and Rebitzer (2018)). We drop individuals to whom we cannot assign a PCP. This leaves us with 167,183 remaining individuals.

Then we link our data to the MPD, matching on both PCP NPI and surgeon NPI. Most PCPs and surgeons are represented in the data, although not all, resulting in us retaining 4,115 PCPs out of 5,550, and 206 orthopedists out of 258. This cut leaves us with 126,956 individuals in our sample. Table 3.12 displays the changes in basic patient summary statistics as we make these sample restrictions. Our final sample is slightly older, and slightly more female than the initial sample, but is quite similar otherwise.

Finally, we construct our outcome variables. We pull all claims from the twelve months following the initial office visit (including the claims from that visit). The total cost of those claims is the measure of cost outcomes that we use in our analysis. We also determine whether or not the patient received an orthopedic surgery within this time period. We do so by checking for the presence of any claim with an orthopedic-surgery-associated CPT

[16] Specifically, we restrict to those who list a primary specialty with a code beginning in '207X', but exclude those with primary specialties in foot and ankle, hand, spine, and pediatrics.

[17] Specifically, we count up the number of claims for a given surgeon with CPT procedure code '27130' (total hip arthroplasty) or '27447' (total knee arthroplasty).

[18] Specifically, we find their first incurred claim with a CPT procedure code of '9920X' or '9921X', for X between 1 and 5. These codes are used to indicate a standard evaluative physician visit in an office setting. Normally, '9920X' are used to indicate a new patient whereas '9921X' are used to indicate an established patient. In situations where the provider is in a multispecialty practice unit (e.g. a practice that contains both PCPs and orthopedists), visits will be coded as established even if the patient has never seen that provider before, so we include both.

code. The list of these codes is given in Appendix 6. We also use claims from the twelve months prior to the initial office visit (the same claims used to determine the patient's PCP) to measure the patient's prior health status. We do so by constructing indicators for a variety of chronic conditions. We specifically choose the 31 chronic conditions that are used to construct the Elixhauser Comorbidity Index, using the presence of certain ICD-9 diagnosis codes of the patient's claims from the year prior as indicators for that condition. The Elixhauser Comorbidity Index is commonly used in the health economics and health services literature to help adjust for underlying patient risk in predictive models of patient mortality, and we use it in our analysis as a control for patient health.

We present summary statistics for our final sample in the third column of Table 3.12. Demographically, our sample is more likely to be older, and female, than the average Massachusetts resident, a demographic that reflects the typical patient with an orthopedic issue. Around three-quarters of our sample live in the Boston Hospital Referral Region (a geography that is somewhat larger than the greater Boston metropolitan area). The same number are enrolled in employer-sponsored insurance, with the rest in Medicaid or Medicare Advantage.

The average 1-year spending among our patients following their first visit with their chosen orthopedist is $12,218. This is much higher than the 2014 U.S. average of $8,045 per person, and even higher than the 2014 Massachusetts average of $10,559 (Massachusetts Health Policy Commission, 2018), but this should not be surprising given that our patient population is older than the median American, and our sample restrictions condition on seeing any doctor, which is likely to raise expected utilization. Indeed, nearly a fifth of our sample receives some kind of orthopedic surgery in that year, which we should expect to raise costs significantly.

The most stark statistic in this table is that 96% of patients see a PCP who is integrated with at least one orthopedist, a profound level of integration. In contrast to other work on hospital-physician integration, this means that we cannot simply use the set of unintegrated PCPs as a control group relative to integrated PCPs: There are too few of them to serve effectively. In our analytic sections we discuss assumptions we make to bypass this issue. This profound level of integration generates substantial self-dealing, with nearly two-thirds of referrals from PCPs being to an orthopedist who he is integrated with. This highlights how critical it is to understand the cost implications of vertical integration.

2.3 Model

We introduce a simple model of PCP referral behavior to motivate our empirical analysis. We model referral choice as a function of patient preferences, cost and quality outcomes, and the PCP's financial incentives. We first describe the general model, and then introduce simplifications and auxiliary assumptions that allow us to map the theoretical model to an estimable empirical model.

The timing of our model is as follows. First, a patient i experiences a health shock (e.g. increased joint pain), and sees a PCP j. The PCP initially evaluates the patient, and decides whether or not they need to see an orthopedist. We consider the choice of specialist that occurs at this point to be where our model begins, although we present results on the extensive margin decision in Section 2.5. We assume that orthopedist choice occurs as a result of a joint decision-making process between the patient and PCP. Our model is agnostic about the nature of this joint decision-making, but we assume it produces pair-specific preferences that are policy-invariant.[19] For ease of exposition, we refer to this decision as a referral being made by the PCP. The PCP observes patient preferences, his own incentives, and a signal of expected patient cost outcomes. He then chooses a specialist k^*. Finally, the cost outcome Y_{ik^*} is realized. Y represents the total costs generated by the specialist, including her own charges for services as well as charges for services that are ancillary to her own, such as anesthesiologist charges incurred during surgery or the cost of recommended imaging.

We model cost outcomes Y_{ik} as:

$$Y_{ijk} = g(X_i, k, V_{jk}, \upsilon_{ijk})$$

where X_i are patient characteristics and V_{jk} is an indicator that is 1 when PCP j and orthopedist k are vertically integrated and 0 otherwise. That is, costs are dependent on patient characteristics, orthopedist identity, whether or not the patient's PCP j is integrated with the orthopedist, and a cost shock, υ_{ijk}, which is realized after the orthopedist is chosen. Our notion of the efficiencies from vertical integration are represented by the value $\mathbb{E}[g(X_i, k, 1, \upsilon_{ijk}) - g(X_i, k, 0, \upsilon_{ijk})]$, the expected change in costs when j and k are integrated as opposed to unintegrated, holding all else equal. As described in Section 2.2, these efficiencies may represent direct reductions in spending thanks to coordination, e.g., non-duplication of imaging, or the spending reduction effects of improved care quality. One important note is that, although the efficiencies we measure are not the reduction in input costs often cited as a justification for mergers, they are the relevant efficiencies that are typically described in antitrust cases, as they represent the product of input costs reductions and the extent to which input cost reductions are passed through to consumers. We cannot disentangle these two factors, but their product alone is sufficient to discuss consumer welfare.

When cost outcomes are realized, the patient's value of treatment is fully realized. We assume that the patient has some preference over both the cost of the orthopedist to them, as well as other characteristics of that orthopedist. We write that value down in the following form:[20]

$$v_{ijk} = f(X_i, r_i\mathbb{E}[Y_{ijk}], k)$$

[19] One potential model described by this would be a model where the PCP unilaterally makes the specialist choice with a fixed altruistic weight on the patient's preferences, which is the way we frame our model. Another equivalent model is one where the patient and PCP engage in Nash bargaining over the specialist choice with fixed bargaining weights.

[20] Our notation appears to embed the assumption that patients are risk-neutral over potential costs, since value is a function of expected cost. This assumption is not strictly necessary, but allows for ease of notation.

Note here that the patient's preferences over costs only apply to $r_i\mathbb{E}[Y_{ijk}]$, where r_i is the patient's coinsurance rate – what share of the total bill they have to pay themselves.

Next, we consider a model of choice given potential cost outcomes. We assume that the outcome of patient-PCP joint decision-making admits a utility representation, which we denote as u_{ijk} for patient i and PCP j's choice utility for orthopedist k. We model u_{ijk} as a weighted sum of patient value v_{ik} and PCP financial payoffs:

$$u_{ijk} = \underbrace{\Psi_j f(X_i, r_i\mathbb{E}[Y_{ijk}], k)}_{\text{Patient preferences}} + \underbrace{\mathbb{E}[B_{ij} - b_{ij}Y_{ijk}] + V_{jk}T_{ijk}}_{\text{PCP preferences}}$$

The referring PCP places a weight of one on her own financial payoffs, and an altruistic weight of Ψ_j on the patient's value of k. The PCP receives two sets of financial payoffs: First, she receives a capitated budget B_{ij} for each patient, and must pay a penalty b_{ij} for each dollar spent on patient care. Second, she receives a payment T_{ijk} from his system when he refers a patient to an orthopedist she is integrated with. We model this as a piece-rate payment. As we described in Section 2.2, incentives are often implicit threats, or bonuses that cannot legally be tied to referral behavior. We think of T_{ijk} as representing the average expected dollar-equivalent difference in these instruments between referring and not referring.

We assume that idiosyncratic decision shocks, ϵ_{ijk}, also are present, which drive otherwise observationally identical patients to different orthopedists. We are agnostic about the source of such shocks. They may come from randomness in i's propensity to follow j's referral, or from random frictions in the collaborative decision-making process, or random taste shocks to either party.. The PCP refers each patient to k^*, where $k^* = \arg\max_k u_{ijk} + \epsilon_{ijk}$. Therefore, we can describe the probability of i being referred by j to k as

$$s_{ijk} = \int_{\mathbb{R}^K} 1\{u_{ijk} + \epsilon_{ijk} \geq \max_{k' \neq k} u_{ijk'} + \epsilon_{ijk'}\} dF(\vec{\epsilon_{ijk}})$$

the probability that i and j's total choice utility for k is higher than all other options k'.

Misallocation

First, we can use this model to describe sources of misallocation. If we believed that there was no misallocation of referrals to orthopedists, then there would be no reason to make policy to influence them. Our setting has two distinct potential sources of misallocation.

The first is the classic moral hazard problem of Arrow (1963): When patients are insured, they do not internalize the full cost of their choices, since the insurer pays a share of it. Therefore, if patients are rational, the optimal allocation of patients to orthopedists is the one that maximizes

$$v_{ik}^* = f(X_i, \mathbb{E}[Y_{ik}], k)$$

i.e. the choices that patients would make if they made the choice of orthopedist, and faced the full cost of that choice. The patient's most-preferred option is instead the one that maximizes v_{ik}, which will put excessive weight on non-cost factors. This is the first source of misallocation: Too much is spent on orthopedists who have a higher level of perceived quality or other differentiating characteristics than would be preferred if the patient had to pay the cost.

The second source of misallocation, which we focus on in this paper, comes from the PCP's incentives not being aligned with patient preferences. The T_{ijk} term encourages PCPs to steer patients towards integrated orthopedists. This steering is unlikely to be positively related to cost savings, and may perhaps be related to cost *increases*, since the system may benefit from orthopedists who incur greater costs. Additionally, if Ψ_j is relatively lower, and patient preferences are not internalized, more weight will be put on PCP preferences, or on idiosyncrasies (ϵ_{ijk}) that may or may not be related to value.

In the paper, we measure the effect of policy and market structure on costs, rather than the more utilitarian welfare measure, v^*. We do this for three reasons: First, we are unable to measure v^*. As we describe in Section 2.6, patient preferences are not separately identified from PCP altruism weights Ψ_j, particularly if we are worried that PCPs may not internalize patient preferences over characteristics in the same proportion as the patient. Second, suggested by prior work such as Brot-Goldberg et al. (2017) (and demonstrated in our analysis in Section 2.6), cost-sensitivity of treatment choice is so low, both due to insurance and other reasons, that improving it, and thus lowering costs, is likely to be a first-order welfare improvement even if it requires reductions in quality. Third, prior work has shown that patients tend to be poor judges of treatment quality, so patient preferences may not truly be a good measure of welfare.

For these reasons, we focus on evaluating potential policies by their effect on costs, measured as

$$C_{ij} = \sum_k s_{ijk} Y_{ijk}$$

Testable Implications

In this section, we derive testable implications of our model for the effects of policies on allocation and outcomes. We derive measures of the impact of the introduction of global budget contracts and of vertical integration. We then discuss how the two might interact. We employ a partial-equilibrium framework, where orthopedists do not respond to policies by changing their practice styles, prices, or forms of differentiation. Models of pricing responses to integration can be found in Cuellar and Gertler (2006) and Capps et al. (2018).

Global Budgets

First, we examine the impact of global budgets, which is relatively simple. Introducing global budgets introduces a capitation payment B_{ij}. This payment does not influence referral choice since it is paid no matter what orthopedist is chosen. However, the penalty, $b_{ij} Y_{ijk}$, does.

Global budget contracts effectively make PCPs more sensitive to costs. The effect of their on any given orthopedist's patient share will be proportional to that orthopedist's relatively costliness compared to other close alternatives. To a first order approximation, this is

$$\Delta^{GB} s_{ijk} \approx \frac{\partial s_{ijk}}{\partial u_{ijk}} \left(-b_{ij} Y_{ijk}\right) + \sum_{k' \neq k} \frac{\partial s_{ijk}}{\partial u_{ijk'}} \left(-b_{ij} Y_{ijk'}\right)$$

That is, the effect on an orthopedist's probability of being referred i by j depends on two things: First, PCPs disprefer k proportionally to the penalty they receive, but they also disprefer other choices k' proportionally to the penalty they receive for those choices. Note that the influence of the utility of an alternative k' on the probability of k being chosen, $\frac{\partial s_{ijk}}{\partial u_{ijk'}}$, depends on the product of two factors: The effect of $\mathcal{M}_k = \max_{k' \neq k} u_{ijk}$, the maximum utility of all alternatives, on the probability of k being chosen, $\frac{\partial s_{ijk}}{\partial \mathcal{M}_k}$, and the probability that k' is the most-preferred of the alternatives, $P(u_{ijk'} = \mathcal{M}_k)$. Note that the probability of k being chosen is $P(u_{ijk} - \mathcal{M}_k)$, so $\frac{\partial s_{ijk}}{\partial \mathcal{M}_k} = -\frac{\partial s_{ijk}}{\partial u_{ijk}}$. Making this replacement we have that

$$\Delta^{GB} s_{ijk} \approx \frac{\partial s_{ijk}}{\partial u_{ijk}} b_{ij} \left[-Y_{ijk} + \sum_{k' \neq k} P(u_{ijk'} = \mathcal{M}_k) Y_{ijk'} \right]$$

So the effect of global budgets on a given orthopedist's market share is proportional to the extent to which k incurs fewer costs than the weighted average of other orthopedists, weighted by their status quo patient share.

Now that we have the effect of global budgets on orthopedist choice, we can derive the effect on expected costs:

$$\begin{aligned}
\Delta^{GB} C_{ijk} &= \sum_k \Delta^{GB} s_{ijk} Y_{ijk} \\
&= \mathrm{Cov}(\Delta^{GB} s_{ijk}, Y_{ijk}) + \underbrace{\mathbb{E}[\Delta^{GB} s_{ijk}]}_{=0} \mathbb{E}[Y_{ijk}] \\
&= \mathrm{Cov}(\Delta^{GB} s_{ijk}, Y_{ijk})
\end{aligned}$$

That the effect of global budgets on costs is equivalent to the relationship between its effect on an orthopedist's patient share and their cost.[21] From our prior analysis, we have a representation of $\Delta^{GB} s_{ijk}$, and can thus show that

$$\frac{\partial \Delta^{GB} s_{ijk}}{\partial Y_{ijk}} \approx -\frac{\partial s_{ijk}}{\partial u_{ijk}} b_{ij} < 0$$

and so, unsurprisingly, global budgets should reduce expected costs. We can see that global budgets are more effective when the share of risk that the PCP must bear, b_{ij}, is higher,

[21] The average change in market share over orthopedists, $\mathbb{E}[\Delta^{GB} s_{ijk}]$, must be zero since the sum of changes must be zero, since the sum of market shares is always 1.

as well as when choices are more responsive to choice utility (i.e., when $\frac{\partial s_{ijk}}{\partial u_{ijk}}$ has a greater magnitude, more weight is placed on incentives).

Another factor to note is that the effect of global budgets depends on the status quo market shares. This will affect both $\frac{\partial s_{ijk}}{\partial u_{ijk}}$ (how sensitive choices are to changes in utility) and $\sum_{k' \neq k} P(u_{ijk'} = \mathcal{M}_k) Y_{ijk'}$ (the average cost of alternatives). Both of these objects are difficult to assess from summary statistics alone without strong theoretical restrictions.

Vertical Integration

Next, we can examine the impact of j joining a system M. System affiliation can affect outcomes in two ways: It reduces the expected cost of patients of j who are referred to orthopedists within M, and it additionally gives j an additional preference for M's orthopedists above and beyond this cost reduction. First, we refer to $\mathbb{E}[g(X_i, k, 1, \upsilon_{ijk}) - g(X_i, k, 0, \upsilon_{ijk})]$, the cost reduction from vertical integration, as the term $-\eta_{ijk}$. We can again use a first-order approach, to find that

$$\begin{aligned}
\Delta^{VI} s_{ijk} &\approx \frac{\partial s_{ijk}}{\partial u_{ijk}} V_{jk} \left(T_{ijk} - \Psi_j \frac{\partial f}{\partial Y} \eta_{ijk} \right) + \sum_{k' \neq k} \frac{\partial s_{ijk}}{\partial u_{ijk'}} V_{jk'} \left(T_{ijk'} - \Psi_j \frac{\partial f}{\partial Y} \eta_{ijk'} \right) \\
&= \frac{\partial s_{ijk}}{\partial u_{ijk}} \left[V_{jk} \left(T_{ijk} - \Psi_j \frac{\partial f}{\partial Y} \eta_{ijk} \right) - \sum_{k' \neq k} P(u_{ijk'} = \mathcal{M}_k) \left(T_{ijk'} - \Psi_j \frac{\partial f}{\partial Y} \eta_{ijk'} \right) \right]
\end{aligned}$$

If we assume that $T_{ijk} = T, \eta_{ijk} = \eta$ for all k, then this becomes clear:

$$\Delta^{VI} s_{ijk} \approx \frac{\partial s_{ijk}}{\partial u_{ijk}} \left(T - \Psi_j \frac{\partial f}{\partial Y} \eta \right) \left[V_{jk} - \sum_{k' \neq k} P(u_{ijk'} = \mathcal{M}_k) V_{jk'} \right]$$

i.e., if j joins a system, k's market share increases if they are part of that system, and decreases if they are not. The effect increases with the size of the incentive $T - \Psi_j \frac{\partial f}{\partial Y} \eta$. It is also moderated by share decreases if the most-preferred alternatives are also being integrated. The preference weighting is important: If, for example, the PCP integrates with a system that contains all the orthopedists he typically refers to the most, integration will only have a minimal effect on referral patterns.

Now, we can use this to examine the effect of vertical integration on costs. The difference in costs will be

$$\Delta^{VI} C_{ijk} \approx \sum_k \Delta^{VI} s_{ijk} (Y_{ijk} - \eta V_{jk}) - \sum_k s_{ijk} \eta V_{jk}$$

Which includes both the reallocation across orthopedists, as well as the effect of making integrated orthopedists less expensive. Using a similar method as above, we can see that this decomposes to

$$\Delta^{VI} C_{ijk} \approx \text{Cov}\left(\Delta^{VI} s_{ijk}, Y_{ijk}\right) - \eta \text{Cov}\left(\Delta^{VI} s_{ijk}, V_{jk}\right) - \sum_{k} s_{ijk} \eta V_{jk}$$

There are three terms here. The first is similar to the effect of global budget contracts: It measures the extent to which choice differences induced by vertical integration reallocate patients to higher- or lower-cost orthopedists. Unlike that result, this one is not easy to sign. Intuition suggests that, ceteris paribus, integration with a system that contains high-cost orthopedists will increase expected costs, and vice versa. However, this is not the complete story: If the PCP integrates with a system that is generally low-cost, but its orthopedists compete most strongly with even lower-cost options (e.g., if there is some kind of market segmentation), then integration may reduce allocative efficiency despite it being with a low-cost system.

The second and third terms are more intuitive, and are, combined, a measure of how much efficiencies reduce costs. The third term is the savings in costs from integration for patients who the PCP would refer to (counterfactually) integrated orthopedists even if there were no incentive to. The second term, $\text{Cov}(\Delta^{VI} s_{ijk}, V_{jk})$ is simply a measure of how much integration shifts patients towards integrated orthopedists, which is also always positive.

The sign of this effect is not a given, since the first term can be positive or negative, and even when positive it can outweigh the latter terms or not.

Interaction

Finally, we can ask how the effects we have described above change when they interact. In Section 2.5, we will look at how integrated systems shape the effect of global budget contracts, so we will examine that theoretically here. A more formal analysis is relatively intractable. Instead, we will describe this in a more abstract way. Recall that

$$\Delta^{GB} s_{ijk} \approx \frac{\partial s_{ijk}}{\partial u_{ijk}} b_{ij} \left[-Y_{ijk} + \sum_{k' \neq k} P(u_{ijk'} = \mathcal{M}_k) Y_{ijk'} \right]$$

First, the efficiencies brought on by integration will lower Y_{ijk} for integrated doctors, meaning that global budgets will *increase* self-dealing through this channel. This may end up lowering expected costs, however, by reallocating patients to orthopedists who are now less expensive than in the counterfactual.

More complicated are the effects of integration on $\frac{\partial s_{ijk}}{\partial u_{ijk}}$ and $P(u_{ijk'} = \mathcal{M}_k)$. If the status quo incentives to self-deal are large, then $\frac{\partial s_{ijk}}{\partial u_{ijk}}$ for unintegrated orthopedists will be relatively low (i.e., even large incentives will not increase the referrals of patients to unintegrated low-cost orthopedists), and so global budget contracts with only induce reallocation within the system, where patients under such a contract are moved towards lower-cost internal specialists only, even when there are better external options. Note that this is only possible when there is sufficient variation in costliness within the system–if all orthopedists in a system have identical costs, and steering is present, capitation may not have an effect at all, except for the handful of patients who would already be sent externally.

Summary

The point of this discussion was to find a way to use our model to determine what parameters . What we find is somewhat disarming–the effect of vertical integration on costs is highly ambiguous even in partial equilibrium, without pricing responses. This comes from the $\text{Cov}\left(\Delta^{VI} s_{ijk}, Y_{ijk}\right)$ term, which measures the reallocative effects of integration. This has ambiguous sign, depending on which orthopedists are integrating, and with whom those orthopedists are competing. This heterogeneity is important to consider, and may explain the wide variety of results in the literature, where researchers have typically studied settings with a single integrated firm rather than many heterogeneous ones.

One important thing to note is that our model shows that a simple comparison of the outcomes of integrated PCPs to unintegrated PCPs will not allow T and η to be separately identified, even if such comparisons did not have endogeneity issues. The effect of integration on patient volumes at newly-integrated orthopedists can be positive both due to high T or due to high η. Adding the effect on spending to this will not necessarily help, as spending reductions can come from high η *or* high T and integration with orthopedists who primarily steal business from higher-cost orthopedists.

This means that we have to estimate the parameters of our model explicitly, rather than rely on reduced-form effects to guide us. This is especially true given that we do not have many unintegrated PCPs to compare to.

This motivates the empirical strategy we follow for the rest of the paper. Because the effects of vertical integration on volumes and costs depend on competitive substitution patterns, we must model those patterns directly, which we do in Section 2.6, taking our choice model to the data. This model requires orthopedist costs as a key input. We estimate these, including vertical efficiencies, in Section 2.4. We can then use these to estimate how PCP incentives change referral patterns, which we do in Sections 2.5 and 2.6.

2.4 Orthopedist Costs

As described in Section 2.3, our model of orthopedist referral choice is built on top of a model of potential cost outcomes at orthopedists. Therefore, before we can explore how PCP incentives affect allocation, we must first estimate those outcomes. This section describes that process. First, we describe our estimation process, and what assumptions we make in order to simultaneously identify orthopedist effects on spending and vertical efficiencies. After presenting our results, we discuss sources of heterogeneity across orthopedists, including variation in extensive margin decisions about whether or not to perform surgery. Finally, we present some analysis of the extent to which patients sort across orthopedists by sickness.

Cost Dispersion

We begin by describing how we estimate orthopedist heterogeneity. Our workhorse model of outcomes is one in which outcomes for patient i depend on the orthopedist k they see, whether or not k is vertically integrated with their PCP j, patient characteristics X_i, and an error term:

$$\log Y_i = \gamma_{k(i)} + \eta V_{j(i)k(i)} + \delta X_i + \upsilon_i$$

Our parameters of interest are the set of γ_k, and η. We will interpret γ_k as the risk-adjusted cost to a patient as a consequence of being referred to orthopedist k. η, on the other hand, is our primary measure of vertical efficiencies. We can interpret it directly as the amount of spending that is conserved when the PCP j and orthopedist k are integrated. Our primary outcome measure Y_i will be total medical spending incurred by the patient (and paid by either the insurer or patient) in the year following the first orthopedist visit. We follow the literature in modeling this as a log-linear function of observables, as the distribution of health expenses approximates a lognormal distribution. We limit our analysis to the years 2012-2013, because for patients who are referred in 2014 we do not observe a full year of claims following their visit.

We estimate this model using OLS. One key assumption allow us to identify γ_k and η: That there is no sorting on related unobservables, i.e., that patients do not select orthopedists based on knowledge of potential match-specific components of cost, conditional on observables. This rules out situations where patients who are unobservably complex are referred to specific orthopedists over others. This seems relatively restrictive, but we find it acceptable. Our controls are rich enough that we observe most of the major sources of cost heterogeneity that are also ex ante observable. In addition, this assumption allows for sorting on health match effects (e.g., patients with hip problems matching to hip specialists), as long as that sorting does not affect costs. It also rules out similar selection in the decision of whether or not to send the patient to an integrated orthopedist, including "cherry-picking" behavior. Swanson (2013) finds little evidence for cherry-picking in the case of cardiology patients being referred to physician-owned hospitals, so we feel comfortable assuming away this behavior. We explore potential sorting in Section 2.4.

Our framework also rules out the ability of productive PCP-orthopedist 'teams' to form, allowing PCPs to refer to orthopedists who they have positive match effects with, as in Agha et al. (2018). In practice, the volumes of most PCP-orthopedist pairs in our data are so low[22] that it would be hard for any such specialization to build up.

This no-sorting assumption also assumes that integration is not endogenous. Since we only observe a single snapshot of integration status, studying how integration occurs is beyond the scope of this paper. Our assumption in this regard is the following: PCPs and orthopedists cannot choose their integration status based on the potential match effects on

[22] In Table 3.19, we show that the average PCP in our data refers 30.7 patients to 9.2 unique orthopedists, implying that the average 'team' volume is 3.3 patients conditional on having any patients.

cost from their integration. In theory, endogenous matching could go either way–systems might form based on ability to cut costs, but they also might form based on the ability to upcharge patients. We leave the analysis of this issue to future work.

We include a rich set of controls in our analysis, which we add sequentially to demonstrate coefficient stability. We begin by adding controls for patient demographics, including dummies for age (bracketed into 18-44, 45-54, 55-64, and 65+), gender, year, and indicators for 31 different chronic conditions.[23] Next, we add in dummies for the patient's insurer, insurance market segment, and plan type. Finally, we add dummies that indicate whether the patient's PCP j belonged to one of each of the eight largest integrated health systems. This generates three regressions, each with its own set of γ_k and η values. These are displayed in the first three columns of Table 3.15.We see that our estimates of the distribution of γ_k change by a relatively small amount as controls are added.

Our initial estimates of γ_k imply substantial orthopedist variation, with a move from the average orthopedist to an orthopedist who is one standard deviation more costly inducing a 30.4% increase in 1-year total costs. At our sample's average spending level of \$12,218, this would increase 1-year spending by around \$3,714. A worry, however, is that sampling variation and small patient panel sizes for some orthopedists may generate measurement error in γ_k which will cause its variance to be overestimated. We handle this issue by using a shrinkage procedure from the empirical Bayes literature. Empirical Bayes shrinkage generates new fixed effect estimates that are a weighted sum of the OLS estimate and zero, with weights determined by the variance of the estimates relative to the variance of the estimators. This method has been used in similar ways both in the hospital quality literature (McClellan and Staiger (1999), Chandra, Finkelstein, Sacarny and Syverson (2016)) as well as the education quality literature (Kane and Staiger (2001), Rose, Schellenberg and Shem-Tov (2018)). We describe this procedure in more detail in Appendix 6. The resulting variation from these, our preferred estimates, is described in the fourth column of Table 3.15. The shrinkage procedure reduces estimated variation by a small amount.

Table 3.15 also contains our estimates of η, the measure of vertical efficiencies. Our preferred estimate is -0.058, implying that referrals to an orthopedist from a vertically-integrated PCP reduce spending by nearly 6% relative to referrals from an unintegrated PCP. Again, at our sample average level of spending, this would reduce 1-year spending by \$708. This falls within the range of the sparse set of prior estimates of vertical efficiencies in the literature–it is higher than Hortacsu and Syverson's (2007) estimate of zero for total factor productivity in the cement industry, but much smaller than Forbes and Lederman's (2010) estimate of 25% reductions in departure delay times in the airline industry. It is substantial, although small relative to the variation across orthopedists. The standard error of η is fairly consistent across models, and rejects a null hypothesis of no efficiencies.

To demonstrate the full distribution of expected costs faced by patients, we generate two plots in Figure 4.7. The upper plot is a histogram of the $\gamma_{k(i)}$ faced by each patient

[23]These chronic condition measures are the component conditions that make up the Elixhauser Comorbidity Index. We describe the Index and how we construct it in Appendix 6.

i. The distribution is right-skewed, with a number of extremely high-cost orthopedists on the right tail but with a substantial amount of variation throughout. 6.9% of patients see an orthopedist whose expected effect is to increase spending by over 50%. Even at more moderate parts of the distribution, half of patients see an orthopedist whose expected effect is to increase costs by at least 7%. In the lower plot, we add the cost-reducing vertical efficiency. Although it does shift the distribution lower to an extent, one can see that it does little to the overall shape of the distribution, implying that although efficiencies do conserve on costs, they may be a drop in the bucket compared to aggregate patterns.

In the first two columns of Table 3.17, we break down orthopedist variation at the integrated system level. For each of the eight top systems, we present the mean and standard deviation of γ_k for the orthopedists within that system. We see variation across systems, including higher-cost systems such as Partners, UMass, and Lahey, as well as lower-cost systems such as Steward, Atrius, and Baycare. More surprisingly, the variation within some systems is nearly as large as the overall variation. Partners and Beth Israel, for example, have as much variation within their own surgeons as there is across the state-wide distribution. With the exception of Lahey, the other systems contain significant variation as well. In Figure 4.8, we plot kernel density plots of orthopedist fixed effects for three systems–Atrius, Partners, and Steward–to highlight their differences. Atrius is relatively smaller, and concentrated around the mean spending level, albeit with a few high-expense orthopedists. Steward is the lowest-cost systems, although it too exhibits substantive variation. Partners is the largest and second-most expensive system, containing the highest-cost orthopedists but having the most variation of any system. This figure symbolizes how we model systems in Section 2.3: That when a PCP integrates with a system, they are integrating with a large set of orthopedists, whose costs may follow a complex distribution.

Sources of Orthopedist Heterogeneity

Observing the orthopedist dispersion that we estimate, a natural question that arises is where this cost variation comes from. Orthopedists can be more costly in a variety of ways: They can charge higher prices; they can perform or recommend more services; or they can choose more expensive service or facility options when choosing to perform a service. This potential heterogeneity is highly multidimensional. We focus on a single, salient distinction: Whether an orthopedist is expensive because they do surgery at higher rates or for other reasons.

We implement this decomposition in a fairly simple way. We begin by constructing an indicator $surg_i$, which represents whether or not i received an orthopedic surgery within a year after their first orthopedic visit. We describe how this is coded in Appendix 6. We then estimate the following two regression models:

$$
\begin{aligned}
surg_i &= \delta^{surg} X_i + \gamma_{k(i)}^{surg} + \eta^{surg} V_{j(i)k(i)} + \upsilon_i^{surg} \\
\log Y_i &= \delta^{other} X_i + \gamma_{k(i)}^{other} + \eta^{other} V_{j(i)k(i)} + \theta surg_i + \upsilon_i^{other}
\end{aligned}
$$

We estimate these using OLS, under the same assumptions as were employed by our baseline cost model. Four sets of parameters are important in these models: $\gamma_{k(i)}^{surg}$, which is k's differential propensity to do surgery conditional on patient observables, $\gamma_{k(i)}^{other}$, k's propensity to incur costs conditional on surgery, η, the vertical integration effects, and θ, the effect of surgery on costs. We perform the same sequential control process as done for our main cost outcomes model, including empirical Bayes shrinkage on $\gamma_k^{surg}, \gamma_{k)}^{other}$.

We present our estimates in Table 3.16. Dispersion in both surgery propensity and other costs are each substantial. By our estimates, patients who see an orthopedist with surgery propensity one standard deviation above the mean are just over 10 percentage points more likely to receive at least one surgery, a substantial increase. Variation in other costs is substantial as well. Receiving a surgery increases costs by 156% on average. We describe the covariance between these two measures of orthopedist costs in a scatterplot in Figure 4.9. We can see that although the two covary to an extent, there are some orthopedists who do many surgeries but incur only moderate costs otherwise, and some orthopedists who incur large expenses but do few surgeries.

To understand the extent to which variation comes from the decision to do surgery, we decompose the variance. First, we note that $\gamma_k = \theta\gamma_k^{surg} + \gamma_k^{other}$, since those are the only two sources of orthopedist costs. Using this, we can see that

$$\text{Var}(\gamma_k) = \text{Var}(\theta\gamma_k^{surg}) + \text{Var}(\gamma_k^{other}) + 2\ \text{Cov}(\theta\gamma_k^{surg}, \gamma_k^{other})$$

We perform this variance decomposition explicitly by calculating the variance of each component and taking its ratio with respect to the variance of γ_k. The results of this exercise are presented in Table 3.18. We can see that the surgery decision alone explains around 30% of the variation in γ_k.

Robustness Check: Patient Sorting

One concern with our analysis is that we assume away patient sorting to orthopedists on unobservables. This is a concern for two sets of parameters: γ_k and η. Ideally, we could use an instrument to shift identical patients across orthopedist. Unfortunately, we do not have an adequate instrument available. Instead, we take an approach inspired by Altonji, Elder and Taber (2005). We describe the extent to which patients who are observably sicker appear to sort towards different orthopedists. If sorting on observable sickness is similar to sorting on unobservable sickness, the former can give us a sense of how strong we expect the latter to be.

We use two forms of observable patient sickness. The first is simply the Elixhauser Comorbidity Index, which is a count measure of the patient's number of chronic illnesses. The second is δX_i, where δ is our estimated effect of patient covariates from our cost model in Section 2.4. Figure 4.10 shows a binned scatterplot, where the average patient value of these sickness measures for bins of patients are plotted against the average γ_k of the orthopedist patients in those bins were referred to. Our plots suggest that sorting on comorbidities is essentially nonexistent–the line of best-fit has a slope close to zero. We do find positive sorting of patients with high expected costs towards orthopedists with high expected costs. However, this sorting is fairly weak, and the slope of the best-fit line implies that a move of 1 in patient-driven expected log costs increases the expected log cost of the orthopedist seen by a mere 0.03, roughly a tenth of a standard deviation of the orthopedist distribution. Given that a standard deviation of the distribution of patient-driven expected log costs is 0.49, if the relationship between unobservable sickness and selection is roughly the same as observable sickness, then the variation in unobservable sickness would have to be twenty times as large as the variation in observable sickness to explain our distribution of orthopedists costs through sorting alone.

We repeat the same exercise, but instead analyze the impact on the extent of internal referrals. This gives a reduced-form measure of what Swanson (2013) calls 'cherry-picking.' Swanson finds little evidence for this phenomenon in sorting to hospitals. The results from our exercise are given in Figure 4.11. Again, we find that sorting does not seem to depend on the patient's Elixhauser index. We do find that patients with higher expected costs are slightly more likely to be referred internally, but this effect is very small relative to the distribution of patient costs. Given that we find that internally-steered patients have lower costs, that suggests that patients would need the sort on unobservable sickness in opposite to the way they sort on observable sickness. The variation in such sickness would also have to be nearly ten times larger than the observable variation if the magnitudes of their effects were the same.

We view these two exercises as suggesting that, although selection on unobservables may be present, we should not be excessively worried about it biasing our cost model estimates.

2.5 Reduced-Form Evidence on Referrals

In this section, we present reduced-form evidence on what influences referral patterns. We begin by documenting referral patterns among PCPs in our data, and how PCP systems may influence referral behavior. Then, we document PCP responses to patients who are under global budget contracts. We show that global budget contracts induce PCPs to refer to lower-cost orthopedists. We validate that this induces differential propensities to self-deal. We then show that the reallocative effects of global budget contracts are different across systems, highlighting the complexity of the interactions between incentives and integration.

PCPs and Referral Patterns

We begin by documenting how PCPs play a role in referral patterns. In Table 3.19, we provide some summary statistics on referrals at the PCP level. The average PCP in our data sees around 30 patients over the course of the three years of our data. This, however, is highly skewed, with a substantial share of PCPs seeing very few patients. Therefore we also show the same statistics for a subsample of PCPs who have at least 40 referrals, which restricts us to 1,064 out of 4,038 PCPs. Of this subsample, the average number of referrals is nearly 80 over three years.

PCPs send patients to just over 9 unique orthopedists on average (16 for high-volume PCPs). This implies at least a decent amount of diversity in referral behavior. This could be skewed, however, if a PCP refers, for example, to 8 orthopedists once and 1 orthopedist many times. To quantify the exact dispersion, we follow Agha et al. (2018) and quantify a 'referral Herfindahl-Hirschman Index (HHI).' For a given PCP, we compute each orthopedist's share of that PCP's referrals. The sum of squared shares is the PCP's referral HHI. For the average PCP, who refers to 9 unique orthopedists, the lower bound of this value is $\frac{1}{9} \approx 0.11\bar{1}$. For that PCP, at the average number patients referred of 30, the upper bound is $0.54\bar{6}$. We calculate an average of approximately 0.33, which is in the middle of these two values. This reassures us that PCPs seem to tailor referrals, rather than have a single specialist of choice. For high-volume PCPs, average referral HHI is 0.22, suggesting that the 0.33 value is likely driven by extremely low-volume PCPs whose referral HHI has a high lower bound.

Again, we note that nearly all PCPs are integrated with at least one orthopedist. The average PCP is linked to over 25 orthopedists, although this high number is driven by Partners–Partners contains 63 orthopedist, so any Partners PCP is linked to at least that many. We see that a substantial share of referrals are internal referrals, with an average PCP sending nearly two-thirds of their patients to integrated orthopedists. Finally, PCPs have diverse tastes in what orthopedists they refer to, with some PCPs sending patients to extremely high-cost orthopedists and some sending to low-cost ones. One might worry that this result is a product of low-volume PCPs who have little experience with the market. However, variation in referral patterns is substantial even among high-volume PCPs, with the 75th percentile PCP sending patients to orthopedists who are 16% more expensive than the orthopedists sent to by the 25th percentile PCP.

We further explore heterogeneity across PCPs by summarizing different referral patterns across PCPs in different systems. Table 3.20 displays four statistics for PCPs in each of the eight large integrated health systems. All analysis in this table is done at the referral level, so PCPs are implicitly weighted by their volume. Firstly, we display the rate of self-dealing. Although two-thirds of all referrals in our data are internal referrals, this varies across systems, with Baycare and Partners engaging in the most self-dealing, and Beth Israel and NEQCA doing the least. We then present the average expected log cost of the orthopedists referred to by system PCPs. PCPs from systems with expensive orthopedists send their patients to expensive orthopedists, which is not surprising given the extensive self-dealing. In the following two columns, we break this out into the average expected cost of orthopedists conditional on an internal or external referral. We should not be surprised that low-cost systems have higher expected costs when referring externally than high-cost systems, since the conditional statement excludes their own specialists. What is more surprising is that there is still some heterogeneity within costly systems. For example, PCPs in UMass, the most expensive system, refer to more expensive external orthopedists than Partners, another expensive system. This may be suggestive of some kind of difference in organizational referral strategy, in that some systems encourage their PCPs to be more sensitive to potential costs than others. This strategy may affect system responses to cost-saving incentives, as well.

The Effect of Global Budgets

The prior results showed that PCPs are heterogeneous, in ways that might be related to their incentives. However, those results may have also been driven by patient differences. In this section, we show that changes to PCPs' incentives do affect referral patterns, by measuring the impact of global budget contracts. We study how these contract reallocate patient referrals across orthopedists.

Our baseline regression is the following:

$$(\gamma_{k(i)} + \eta V_{j(i)k(i)}) = \beta_i GBShare_i + \zeta X_i + \epsilon_i$$

That is, we regress the expected log 1-year cost of the orthopedist k that patient i was referred to by their PCP on whether or not i was covered under a global budget contract.[24] β, in this model, represents $\frac{\partial C_j}{\partial GB}$ as described in Section 2.3. Ideally, we would be able to observe whether i was covered under a global budget contract exactly. However, as described in Section 2.2, we only observe, for each patient, the probability they are covered. We use this probability, represented by $GBShare_i$, as the primary regressor. Although using $GBShare_i$ is less efficient than observing contract status, it will produce unbiased estimates of the effect of global budget contracts.[25]

We estimate this via OLS. Identification of β relies on panel variation in the use of global budget contracts across insurer, market segment, plan type, and year combinations, with fixed effects for each component part to absorb inherent differences across the groups. This identification strategy mirrors that of Ho and Pakes (2014) in their study of capitation in California, and is valid under the assumption that patients who are more likely to be covered by a global budget contract are not differentially referred to higher- or lower-cost orthopedists for unobservable reasons. The clearest violation of this assumption would be if being covered by a global budget causes patients to sort towards PCPs who have different referral patterns. We argue that this is reasonable to rule out, as patient costs would not depend directly on the use of global budgets, and patients are likely not even aware of how their physicians are reimbursed. Another violation of this assumption would be if patients who were differentially needy or had preferences for different orthopedists were differentially likely to be covered by a global budget contract. We explore this in Section 2.5.

Our initial estimate is presented in the first column of Table 3.21. We control for a rich set of observables, including age, gender, insurer, insurance type, and patient insurance market segment. Interpreting the coefficient estimate, a move from standard fee-for-service reimbursement to global budget reimbursement causes PCPs to refer patients to surgeons who are approximately 6% less costly in the year following the first visit. At the average level

[24] Our results are qualitatively robust to excluding the efficiency term η.

[25] This is true as long as $P(GB_i|GBShare_i)$ is independent of what orthopedist would be chosen both when $GB_i = 0$ and when $GB_i = 1$.

of 1-year spending in our sample ($12,218), this represents a modest reduction in spending of $745 per patient.

Integrated System Responses to Global Budgets

This initial estimate restricts the treatment effect of global budget contracts to be uniform across PCPs. In the second column of Table 3.21, we allow β to differ for PCPs who are part of one of the eight largest integrated health systems. We also allow PCPs for the system to have different baseline levels of cost for the orthopedists they send to, which may be driven by self-dealing or by unrelated baseline referral patterns. The results from this analysis are presented in the second column of Table 3.21. We can see differential responses across systems–UMass and Partners have the largest responses, whereas Atrius and NEQCA have relatively muted responses.[26] We do not wish to overinterpret the different magnitudes of these effects, since they may constitute either real differences in responsiveness, or differences in the extent of risk-sharing in the contracts the PCPs hold.

Next, we examine the effect of global budget contracts on self-dealing. Interviews with health systems affected by Blue Cross's global budget contracts program, the Alternative Quality Contract, suggested that they would respond to capitation by changijng their level of self-dealing (Mechanic, Santos, Landon and Chernew, 2011). We use the same basic regression structure as before, but instead regress $V_{j(i)k(i)}$, the indicator for j and k being vertically tied, on $GBShare_i$. We restrict the sample to only those patients who see a PCP who shares a vertical tie with at least one orthopedist. We first model this as a uniform effect, and then allow it to differ across health systems as in Table 3.21. Our results are presented in Table 3.22. Consistent with Mechanic et al. (2011), we find that the use of global budget contracts causes the rate of self-dealing to *increase* slightly, by around 2 percentage points. This effect masks substantial heterogeneity across systems, as presented in the second column. The increase in self-dealing is driven by relatively lower-cost systems, like Atrius and Steward. They are contrasted by declines in self-dealing by higher-cost systems like Partners and UMass.

This suggests that changes in self-dealing may be a main channel of the savings from the introduction of global budget contracts. We test this suggestion in the data by measuring the strength of different channels of system-specific cost reductions. This exercise also allows us to understand different effects of capitation across organizations, since, as we mentioned earlier, we cannot interpret the magnitudes of our interaction terms.

To understand what the potential channels are, we first note that the average costs for patients whose PCPs are part of a given system m are

[26] Baycare and Lahey's responses stand out for being approximately zero. Auxiliary data we have on provider-specific exposure to global budgets suggests that the two systems saw no change in their global budget contracting experience for the three largest insurers over the three years of our data, and so it is sensible that their PCPs should not respond to changes in aggregate usage of global budget contracts. We take these approximate zero estimates as suggestive that the responses we observe are roughly correct, and we omit discussion of Baycare and Lahey throughout the rest of the paper.

$$C_m = s_m^{in} c_m^{in} + (1 - s_m^{in}) c_m^{out}$$

where s^{in} and $(1 - s_m^{in})$ are the share of patients who are sent to orthopedists inside and outside of m, respectively, and c_m^{in} and c_m^{out} are the average expected costs of orthopedists that PCPs in m send to inside and outside of m. Given this, we can see that if we differentiate with respect to GB, we get

$$\underbrace{\frac{\partial C_m}{\partial GB}}_{\beta^0+\beta^m} = \underbrace{s_{in}\frac{\partial c_{in}}{\partial GB}}_{\text{Internal Reallocation}} + \underbrace{(1 - s_m^{in})\frac{\partial c_{out}}{\partial GB}}_{\text{External Reallocation}} + \underbrace{\frac{\partial s_{in}}{\partial GB}(c_{in} - c_{out})}_{\text{Cross-Org Reallocation}}$$

that is, when PCPs of system m respond to global budget use, they can do so in three possible ways. First, they can take patients who were going to be referred to higher-cost orthopedists within m and instead refer them to lower-cost orthopedists within m, which we call "internal reallocation." Second, they can trake patients who were going to be referred to higher-cost orthopedists outside of m and instead refer them to lower-cost orthopedists outside of m, which we call "external reallocation." Finally, they can take patients who would have been referred to orthopedists within m and instead refer them outside of m, or vice versa, which we call "cross-organization reallocation."

We perform an explicit decomposition of $\frac{\partial C_m}{\partial GB}$, which is equal to $\beta^0 + \beta^m$, into these three component parts. We take s^{in}, c_m^{in}, and c_m^{out} from the first, third, and fourth columns of Table 3.20, respectively. We use estimates of $\frac{\partial s_{in}}{\partial GB}$ from Table 3.22. Finally, we estimate $\frac{\partial c_{in}}{\partial GB}$, $\frac{\partial c_{out}}{\partial GB}$. We do so by replicating our baseline regression of the effect of global budgets on allocation, but instead condition on internal and external regressions, respectively.

With these estimates in hand, we compute the three component parts of the decomposition. We divide each part by $\beta^0 + \beta^m$, so that we compute the share of the global budget effect driven by each of the three channels. The results from this exercise are given in Table 3.25. We find, surprisingly, that contrary to our supposition, the cross-organization channel of reallocation has a limited contribution to global budget savings. In fact, the systems for which cross-organizational reallocation has the largest effect, Atrius and Steward, are low-cost systems, and these gains come from moving patients *into* the system rather than out of it. The majority of the gains instead come from external reallocation. However, Partners is noticeably different from other systems, in that the vast majority of their savings come from internal reallocation.

The cause of this difference in system response is not obviously clear. As described in Section 2.3, the effect of system participation on response to global budgets depends on how integration changes substitution patterns. Integration can change these patterns in many ways, though–in the extent to which the system incentivizes self-dealing and the extent to which those incentives respond to global budgets. Moreover, even conditional on these strategic variables, the effect of integration may differ if the system's set of orthopedists (and therefore the set of orthopedists who a PCP in that system has an increased preference for) are relatively more expensive. Even these results alone do not allow us to separate out these sources–Partners may engage in more internal reallocation because it employs stronger agency incentives which encourage PCPs to keep patients within the Partners system, or if it engages in more internal reallocation because Partners simply owns a substantial share of orthopedists whose costs vary highly.

Extensive Margin Responses

We have primarily discussed the reallocation of patients across surgeons. However, global budgets may generate incentives for another decision margin: Whether to send a patient to an orthopedist at all. For patients where a PCP is on the margin of whether or not to refer them at all, the increased cost to the PCP of referrals may dissuade them from making a referral. Our current sample construction does not allow us to explore this question, since we do not include patients who were never referred.

We undertake an analysis of this margin by constructing an auxiliary dataset that includes both referred and unreferred patients. We do that in the following way: We begin with our primary sample. For each patient, we find the last office visit the patient had with their PCP before their first orthopedist visit. We then construct a matched sample of unreferred patients by finding all patients who also had office visits with the same PCP on the same day. Combining our main analytic sample and the matched sample, we then estimate the following regression:

$$Referred_i = \beta_i^{Ext} GBShare_i + \zeta^{Ext} X_i + \epsilon_i^{Ext}$$

where β^{Ext} represents the effect of global budgets on the probability of being referred to an orthopedist. Estimates from this regression are presented in the first column of Table 3.26. Our results suggest that global budgets reduced the probability of a referral by 4.3 percentage points, from a base of a 17.8% referral rate. One worry is that some PCPs in our sample have low patient volumes, increasing measurement error in our matching procedure and the resulting estimate. Therefore, in column 2, we re-estimate the model, restricting only to PCPs who we observe referring at least 20 patients to an orthopedist. This generates a slightly lower estimate for β^{Ext}, an effect of 3.3 percentage points from a base rate of 16.7%.

We interpret the size of these responses as fairly small. This may be due to the long-run nature of global budget contracting. Although holding back on a referral for a patient in the present may cut back on expenses, this may cause an undertreatment of the underlying medical issue. If this undertreatment has the potential to lead to complications, and thus higher expenses in the near future (e.g., if failing to treat arthritis in the present leads to a fracture from a fall in the future), nonreferral can actually be more expensive than referral. We suspect that a similar reasoning drives the effects we observe.

A similar concern is that, although the overall response is small, if the response is concentrated among global budget recipients who are unobservably different than the population at large, their exclusion from our main sample may bias our estimates in Section 2.5. To address this, in Figure 4.12 we show a binned scatterplot of $GBShare_i$ against two measures of observable sickness: The Elixhauser Index for i, and δX_i, where δ is the set of parameters from our cost model in Section 2.4. We find that patients with a higher chance of being covered by a global budget are slightly sicker, conditional on being referred. This is unsurprising, given that patients on the extensive margin are likely to be much less in need of surgical attention than the average referred patient. Therefore, to the extent that we showed in Section 2.4 that sicker patients are slightly more likely to go to a higher-cost orthopedist, our estimates of the effect of global budgets may be slightly understated.

Other Sources of Heterogeneity in Responses to Global Budgets

As a final reduced-form exercise, we return to our decomposition of surgeon effects in Section 2.4, where we showed that orthopedists can differ both in their propensity to do surgery and their propensity to incur non-surgical costs. We note again that $\gamma_k = \theta\gamma_k^{surg} + \gamma_k^{other}$. This implies that the response to global budgets can be decomposed as:

$$\beta = \theta\beta^{surg} + \beta^{other}$$

i.e., we can decompose the response to global budgets as moving to orthopedists who are lower-cost because they do less surgeries, and moving to orthopedists who are lower-cost because they incur less costs even when they do surgery. We estimate β^{surg} and β^{other} by estimating our baseline global budgets regression, but regressing on $(\gamma_{k(i)}^{surg} + \eta^{surg}V_{j(i)k(i)})$ and $(\gamma_{k(i)}^{other} + \eta^{other}V_{j(i)k(i)})$ instead. Our estimates are presented in the third through sixth columns of Table 3.21. We call $\frac{\theta\beta^{surg}}{\beta}$ the share of the global budget effect coming from the channel of reducing surgeries, and $\frac{\beta^{other}}{\beta}$ as the share coming from the 'other costs' channel. We present this decomposition in Table 3.23. The surgery costs channel represents about a quarter of the reductions in spending. This is disproportionately somewhat low compared to the extent to which surgery costs explain orthopedist variation. Surgery costs alone explain 30% of variation, and they explain more when accounting for covariance between their costs and other costs. Instead, it seems that PCPs conserve costs by referring to low-other-cost orthopedists. This may be suggestive of the idea that PCPs respond on margins that are easier to observe, and that they may have better signals for factors like prices than for surgery propensity.

2.6 Referral Choice Model

Although our reduced-form analysis describes the effect that global budget contracts had on referrals, it is unable to fully articulate the mechanisms. For example, since we observe nearly zero PCPs who are completely unintegrated, considering the counterfactual impact of integration on referral patterns and global budget effects is impossible from our reduced-form results alone.

In this section, we describe how we can estimate the parameters of a structural model of choice similar to the one that we describe in Section 2.3. That model involved a number of nonparametric objects that our data is not large enough to estimate efficiently, so we make a set of functional form assumptions, as well as assumptions about parameters that cannot be observed or estimated. Next, we describe how moments of our data serve to identify key choice parameters. Finally, we present our parameter estimates.

Functional Form and Estimation

Recall that in our model of choice from Section 2.3, we specify the choice utility for patient i at orthopedist k to be

$$u_{ijk} = \Psi_j f(X_i, r_i \mathbb{E}[Y_{ijk}], k) + \mathbb{E}[B_{ij} - b_{ij} Y_{ijk}] + V_{j(i)k} T_{ijk}$$

We make a number of assumptions to turn this into an estimable object. First, we assume that $\Psi_j = \Psi$ for all j. Referral choice may, for example, be more sensitive to a patient's distance from an orthopedist because the patient has a strong preference for close orthopedists, or because the PCP is putting higher weight on patient preferences. We cannot separately identify these, and for our purposes they are not necessary to separately identify, so we assume heterogeneity in Ψ_j away. Second, we do not directly model r_i. Cost-sharing is typically nonlinear, and not directly reported in our data. Instead, we allow for heterogeneity in cost preferences by insurance type, which we think should safely proxy for heterogeneity in cost-sharing. Third, we assume that PCPs treat global budget contracts as linear, and assume that global budget contracts are like across insurers. We feel comfortable with the linearity assumption, since PCPs likely are not able to track patient spending exactly. The assumption of contract likeness is nakedly incorrect. However, given our inability to observe the precise contract terms, we cannot do better. We therefore think of our estimates of the effect of global budgets as the average effect on cost-sensitivity at the current market-wide intensity of supply-side risk-sharing. We additionally assume that the measure of expected costs that the PCP responds to, $\mathbb{E}[Y_{ik}]$, is expected log 1-year costs, $\widehat{\gamma}_k + \widehat{\eta} V_{j(i)k}$.[27] Finally, we assume that there are unobservable idiosyncratic preference shocks at the patient-orthopedist level, ϵ_{ik}, that are i.i.d. standard Gumbel, so that our model reduces to a multinomial logit. With these assumptions, we have:[28]

$$u_{ik} = (\beta_i^0 + \beta_i^{GB} GB_i)\left(\widehat{\gamma}_k + \widehat{\eta} V_{j(i)k}\right) + T_i V_{j(i)k} + \beta^Z Z_{ik} + \epsilon_{ik}$$

[27] This assumption is primarily due to computational concerns–using our more precise estimate of $\mathbb{E}[Y_{ik}]$, which would be $\exp\left(\widehat{\beta} X_i + \widehat{\gamma}_k + \widehat{\eta} V_{j(i)k}\right)$, generated problems in estimation. For patients with high $\widehat{\beta} X_i$ values, the differences in expected costs between high- and low-cost orthopedists are magnified, and our estimation procedure had difficulty rationalizing these patients' choice of high-cost options even at low levels of cost-sensitivity.

[28] We suppress the j subscript since j, in our data, is deterministic for i.

with

$$\begin{aligned}
\beta_i^0 &= \beta^{0,M} M_{j(i)} + \beta^{0,X} X_i \\
\beta_i^{GB} &= \beta^{GB,0} + \beta^{GB,M} M_{j(i)} \\
T_i &= T^0 + T^M M_{j(i)}
\end{aligned}$$

where M_j are indicators for j's affiliation with health systems, X_i are patient characteristics, including age brackets, gender, and Elixhauser index brackets, GB_i is an indicator that is 1 when i is covered by a global budget and 0 otherwise, and V_{jk} is, again, an indicator that j and k are vertically integrated.

Next, we allow steering to vary across systems and respond to the introduction of global budgets. We allow cost-responsiveness to vary by demographics. We allow responses to global budgets to vary across systems, and allow PCPs to respond differentially to costs and to global budgets when considering inside orthopedists as opposed to outside orthopedists.

Our parameters of interest are $\boldsymbol{\beta} = \{\{\beta^0\}, \{\beta^{GB}\}, \beta^Z\}$ and $\mathbf{T} = \{\{T^0\}, \{T^{GB}\}\}$, which we estimate with data on choices

We consider a small set of orthopedist characteristics for Z_{ik}. We include γ_k^{surg}, k's propensity to do surgery conditional on patient characteristics, and we allow referral preferences over this propensity to vary to the same extent that we allow preferences over costs to vary. We include quality measures from ProPublica's Surgeon Scorecard. In particular, we include a dummy for whether the orthopedist is included at all in the scorecard for hip and knee replacement complication rates,[29] as well as linear preferences over their complication rates. We also include dummies for the orthopedist's Hospital Referral Region of practice, and allow PCPs to have differential preferences over orthopedist locations based on their own location. Finally, we allow for dummies for orthopedist system affiliation. Given the multinomial logit form, we have that the probability of choosing orthopedist k is

$$P_i(k) = \frac{\exp(u_{ik})}{\sum_{k'} \exp(u_{ik'})}$$

As in our reduced-form analysis, we do not observe GB_i. In that analysis, we replaced it with $GBShare_i$, the probability that i was covered by a global budget contract. In a linear model, this is sensible. In a nonlinear model like this one, that strategy will not necessarily produce the true effect of global budgets. Instead, we integrate over the distribution of GB_i. Therefore, our probability of choice is given by

$$\begin{aligned}
P_i(k|\boldsymbol{\beta}, \mathbf{T}, \mathbf{X}) &= GBShare_i \cdot P_i\left(k|\boldsymbol{\beta}, \mathbf{T}, \mathbf{X}^{-GBShare}, GB = 1\right) \\
&+ (1 - GBShare_i) \cdot P_i\left(k|\mathbf{T}, \mathbf{X}^{-GBShare}, GB = 0\right)
\end{aligned}$$

[29] ProPublica's scoring is based on data from Medicare. Due to CMS requirements barring users from reporting data on small cells, 53% of orthopedists in our data do not have a hip replacement complication rate reported and 25% do not have a knee replacement complication rate reported.

We estimate $\boldsymbol{\beta}, \mathbf{T}$ via maximum likelihood estimation, maximizing the log-likelihood function:

$$\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{T}} = \arg\max_{\boldsymbol{\beta},\mathbf{T}} LL(\boldsymbol{\beta}, \mathbf{T}|\mathbf{X}) = \arg\max_{\boldsymbol{\beta},\mathbf{T}} \sum_i \sum_k 1\{i \text{ choose } k\} \log P_i(k|\boldsymbol{\beta}, \mathbf{T}, \mathbf{X})$$

Identification

We have two sets of parameters to estimate: $\boldsymbol{\beta}$ and $\mathbf{T}$. Identification of $\boldsymbol{\beta}$ is relatively straightforward. β^0 is identified from the covariance between patient shares and expected costs for unintegrated orthopedists, and how that covariance varies across patient demographic bins. The global budget responses, β^{GB}, are identified from the relative strength of these covariances across insurer-year-market segment-plan type groups who have a greater share of patients who are covered under global budget contracts. Similar to β^0, β^Z is identified from the covariance of orthopedist non-cost characteristics and patient shares.

Identification of T is more complex. We observe complex integration arrangements, where nearly all PCPs are integrated with at least one orthopedist. Therefore, our identification cannot rely on differences in shares between PCPs who are integrated with a given orthopedist and those who are not integrated with any orthopedist, since that comparison is unavailable. Instead, we make the following assumption, which is embodied in our functional form: Integration does not differentially affect preferences for orthopedists in other systems, by their system. This is distinct from integration affecting preferences–we allow PCPs in different systems to have different β^0 and different β^{GB} values. This means that we can use, for example, the relative choices of Partners PCPs across different non-Partners orthopedists to identify β^0 and β^{GB}, and use the comparison against their choices of Partners orthopedists to identify T.

Results

We present parameter estimates in Table 3.27. The first row is the average, over patients, of each of the three main parameters in our data:

1. β^0, the baseline PCP sensitivity to costs

2. β^{GB}, the extent to which global budgets increase sensitivity to costs

3. T, the incentive for PCPs to steer patients to integrated orthopedists net of vertical efficiencies

We also provide standard errors for each of these parameter averages. We compute standard errors by bootstrap, using 40 bootstrap runs.[30]

[30] As of this draft, we are currently in the process of increasing the number of runs.

These parameters are denominated in utility units, so interpreting their magnitudes directly is impossible. One easy way to benchmark them is against the standard deviation of idiosyncratic preferences ϵ_{ik}. Our model assumes that they are distributed standard Gumbel, which has a standard deviation of $\frac{\pi}{6} \approx 1.28$. This serves as a baseline relative scaling factor for the rest of our parameters.[31] First, we can examine β^0. Recall that β parameters are multiplied by $\gamma_k + \eta V_{j(i)k}$. Dividing 1.28 by the average value of β^0, we can see that a PCP is indifferent between two otherwise-identical orthopedists, where one is granted a one standard deviation increase in idiosyncratic preference, and the other has expected log costs that are 64 lower. This is nearly fifty times larger than the difference between the most costly and least costly orthopedist in our data, implying that cost-sensitivity is essentially nonexistent in the absence of global budget incentives. This varies by system, with Steward's PCPs having an average β^0 of -0.10 and Partners having a value of 0.04, but all of these are statistically and economically indistinguishable from zero.

Under global budgets, we can see that cost sensitivity increases dramatically, by 0.56. This is still relatively small, however. Given that a one standard deviation change in costs is 0.294, which is quite high, it is still true that PCPs consider equivalently a one standard deviation change in ϵ_{ik} and a 7.5 standard deviation change in $(\gamma_k + \eta V_{jk})$. This is about 150% of the difference between the most and least costly orthopedists we record, implying that even capitation cannot bring cost competition to the forefront.

In contrast, the average value for T is 1.63, equal to around 1.27 times the standard deviation of ϵ_{ik}. T is one of the most important factors in orthopedist choice–in our analysis, it is second only to Boston patients' unwillingness to travel to Western Massachusetts for care.

A first look at our results suggests that steering is deeply important. This does not directly tell us about the effect of steering on self-dealing. We examine that, as well as the effect on allocation generally, in the next section.

2.7 Counterfactuals

The parameter estimates presented in Section 2.6 suggest that 1) global budgets matter in that they increase sensitivity to 1-year costs; and 2) vertical integration matters in that our estimates imply that PCPs prefer integrated orthopedists over identical unintegrated orthopedists who incur substantially lower costs. The parameter estimates alone do not tell us the exact magnitude of these results. Instead, we quantify those magnitudes in this section by using counterfactual simulations. In these simulations, we change the use of global budget contracts and/or the integration status of PCPs and orthopedists, and compute orthopedist choice probabilities. We can then compute the impact on the extent of self-dealing, expected costs, and potential competition.

[31] Recall that the scale of idiosyncratic preferences in the standard logit model are not identifiable separately from the scale of preferences for observables.

Self-Dealing

We begin by studying what drives vertical referrals in the first placce. There are two potential drivers of self-referrals: Sensitivity to cost efficiencies (a combination of η and β^0), and steering (T). We simulate choice probabilities $P_i(k)$ under three regimes: The status quo of vertical integration; a setting where $\eta = 0$, i.e., there are no efficiencies; and a setting where $V_{jk} = 0$ for all j, k, i.e., there is no integration. For each, we compute the probability of internal referral,

$$\frac{1}{I} \sum_{i,k} V_{j(i),k} \cdot P_i(k|\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{T}}, \mathbf{X})$$

where I is the number of patients. The results of this are given in the first row of Table 3.28. Removing efficiencies only reduces the self-referral rate from 62% to 61%. In contrast, disintegrating completely reduces the referral rate to orthopedists by their formerly integrated PCPs by over half, to 25%. This suggests that, regardless of whether efficiencies have positive welfare impacts, they are not the driver of the extensive self-dealing we observe in the data.

In the second through seventh rows, we display these counterfactuals, for the more specific case of the self-referral rates of PCPs from a given system to orthopedists of that system. We can see that the effects of disintegration vary. The changes come in part from the relative strength of the system's steering, the relative favorability of the surgeons in a given system, as well as the other local options. For example, UMass retains substantial market share from its integrated PCPs even in the absence of integration, because its orthopedists are largely located in Western Massachusetts, where other options are scarcer than in Boston.

Expected Costs

Next, we study the ramifications of integration and global budgets for costs. Here, we consider four potential outcomes from two binary policy instruments. For the first policy instrument, we again consider the idea of disintegrating all vertical ties by setting V_{jk} to 0 for all j, k, compared to the status quo. For the second, we change which patients are exposed to global budgets. We consider two possibilities: One where all patients are covered by a global budget contract ($GBShare_i = 1$ for all i), and one where no patients are covered ($GBShare_i = 0$). We then compute orthopedist expected log cost effects,

$$\frac{1}{I} \sum_{i,k} \left(\widehat{\gamma}_k + \widehat{\eta} V_{j(i)k}\right) \cdot P_i(k|\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{T}}, \mathbf{X})$$

under each regime. Differences in this value across two policy counterfactuals will approximate the percent change in spending between the two. The first row of Table 3.29 plots this measure under our four simulated policies. Unsurprisingly, we see that introducing global budgets (comparing the first and third columns) reduces expected log costs, from 0.097 to

0.064, a similar magnitude to that estimated in our reduced-form exercise. More surprisingly, when we remove integration (comparing the first and second columns), expected costs *increase* from 0.09 to 0.141. Adding global budgets to this disintegration policy reduces costs to 0.110. This effect is slightly larger than under integration, but not large enough to return expected costs to their status quo level.

This result is puzzling: Since integration steering incentives are in competition with incentives to reduce costs, one might expect that the absence of the former would strengthen the influence of the latter. One explanation is that disintegration also removes efficiencies, which will raise costs. In Table 3.30 we compare full disintegration (again, the third and fourth columns) to a counterfactual simulation where we remove only efficiencies (the first two columns). We see that the removal of efficiencies increases costs substantially, comparing the first column of Table 3.30 to the first column of Table 3.29. Still, even without efficiencies on the table, we still see that removing vertical ties (comparing the first and third columns of Table 3.30) increases expected costs. Looking at the heterogeneity in both tables gives an easy explanation as to why: Disintegration lowers costs at high-cost systems like Partners and UMass, but raises them at lower-cost systems like Atrius and Steward. This comes from the fact that there is virtually no cost sensitivity among PCP referrals, so even when steering is absent, referrals will still not be sent to low-cost orthopedists. In fact, for Atrius and Steward, their steering incentives are the strongest force pushing patients towards low-cost orthopedists, as we can see by how costs leap when Atrius is disintegrated. Moreover, the reason why efficiencies matter for costs but not for self-dealing comes from the same factor: PCPs do not internalize substantial efficiencies in their referral choices.

A natural question that arises is how much cost-sensitivity is needed to restore competition. Defining this is challenging. We choose an easy benchmark: If a policymaker was to forcibly disintegrate all organizations, how much cost-sensitivity would they have to induce to return expected costs to the status quo? In this way, we define our cost-sensitivity parameter as a measure of the extent of competition, similar to the use of price elasticities of demand in more standard product markets.

We do so from a baseline of no global budgets. We replace our estimated β^0 with $\beta^0 + \Delta$, and compute Δ such that, with Δ and no integration, expected costs are equal to 0.095. This requires a Δ of approximately 0.797, approximately 42% greater than the estimated effect of global budget contracts. This implies that policymakers would need to put even stronger incentives for cost-sensitivity if a harsher vertical antitrust policy was undertaken. However, these incentives would require insurers or policymakers to shift more risk onto PCPs, requiring hefty risk premiums to be paid (Holmström, 1979). Moreover, dis-integration will reduce the size of PCPs' firms, making it hard for them to spread risk throughout their organizations. This will make it costlier for PCPs to bear the same amount of risk and require they be paid a greater risk premium.

Given the lack of competition we observe, the present state of vertical integration represents an odd sort of second-best arrangement for some patients. Integrated firms do produce real efficiencies, although those efficiencies are not large relative to the wide variation in the cost of care. Moreover, the steering efforts by low-cost firms like Steward do improve allocative efficiency even though they are anticompetitive.

However, this is second-best to simply having more cost competition. Not only would that improve allocative efficiency, other work suggests it would have other positive effects that integration does not incorporate. Work has shown that competition tends to improve productivity, both directly by treatment (in e.g. Backus (2014)) and indirectly through demand-driven selection pressures (in e.g. Chandra et al. (2016)). In contrast, in a setting like ours, the competition for orthopedists is essentially generated entirely through competition in the market for PCPs instead. If orthopedist productivity in a system is not related to the attractiveness of the system's PCPs, selection pressures will not necessarily be efficiency-enhancing in the long run, and may even be efficiency-reducing if, for example, systems with costlier orthopedists engage in more vertical integration.[32]

Moreover, although some of the steering is beneficial, it also forecloses on rivals' demand. Both in standard models of price-setting, as well as intrafirm bargaining models such as the 'Nash-in-Nash' model of Ho and Lee (2017), this will raise prices relative to nonintegration, since specialists will have increased bargaining power. This channel is the explanation that Baker et al. (2014) and Capps et al. (2018) give for why hospital-physician integration raises hospital prices.

Understanding the viability of competition- and efficiency-enhancing policy measures depends on the balance between effects on allocative efficiency and productive efficiency, and these countervailing forces. We leave the study of this balance to future work.

2.8 Conclusion

Improving allocative efficiency in U.S. health care is a difficult task. The most influential agents in the decision of where a patient will seek care, PCPs, do not generally receive direct incentives to act on their patients' behalf. We find, however, that they do appear to receive incentives from suppliers of specialty care who they are integrated with. The raw data alone displays this, given the internal referral rate of nearly two-thirds.

Incentives to engage in such self-dealing can come either from efficiencies (that integration brings some reduction in treatment costs), or from upstream specialists paying downstream PCPs to steer patients. Our empirical results suggest both are relevant, but that only the steering component drives self-dealing. However, the combination of efficiencies and a lack of cost-sensitivity among referring PCPs means that disintegrating vertical systems, and thus getting rid of this steering, will not improve efficiency, but in fact make it worse, thanks to a quirk of the present setting: That anticompetitive steering by low-cost systems offsets the same actions by high-cost systems.

[32] A rent-seeking theory of the firm would suggest that integration happens when there are larger quasi-rents to be split, which would be the case for higher-cost systems. We leave the validation of that hypothesis to future work. In our setting, Partners has been engaging in rapid acquisition in recent years, suggesting the empirical validity of this hypothesis.

One way to alleviate this problem is for insurers to introduce direct incentives towards PCPs, to offset this steering behavior. Our results show that global budget capitation schemes do achieve their intended goal, in that PCPs respond by sending patients to specialists who incur 6% lower costs. However, this does not solve the overarching incentive problem: Global budgets also seem to slightly increase self-dealing among lower-cost systems, and the induced increase in cost-sensitivity is not enough to break the market power of high-cost systems. It is also not enough to restore partial-equilibrium allocative efficiency in the absence of integration – our counterfactuals show that expected orthopedist costs are slightly *higher* under global budgets and no integration, compared to status quo integration and no global budgets.

All of our results, taken together, suggest a serious competition problem for specialty care. Although our results suggest the partial-equilibrium allocative efficiency of vertical integration, given the strong steering, it is not hard to imagine that integration is raising prices, as shown by Baker et al. (2014) and Capps et al. (2018). Capitation may help to restore competition, although our results suggest that the optimal level of incentives might need to be much higher than the average incentive used at present. Our results on vertical structure imply that a more vigorous antitrust policy is desperately needed. Unfortunately, physician practice acquisitions are individually generally so small that they fall below Hart-Scott-Rodino thresholds for reporting to federal antitrust agencies. However, as Capps, Dranove and Ody (2017) recommends, state authorities have the ability to pursue these cases. Even for existing systems, where actions to break them up (such as divestiture) may be difficult, greater monitoring of physician pay may be necessary so that systems cannot skirt the Stark laws.

Our results also have ramifications for Accountable Care Organizations (ACOs), an organizational form codified by the Affordable Care Act. ACOs combine both vertical coordination and cost-controlling incentives similar to the capitation contracts we study. Our results suggest that the combination serves to reduce competition more than integration alone, since we find that capitation slightly increases the level of self-dealing. Therefore, policymakers must be careful when approving new ACOs. An ACO constructed with inefficient specialists may result in increased costs even with high-powered incentives, as patients may end up stuck with the specialists in the ACO. This is particularly likely, as the variation in orthopedist costliness is much wider than the cost reduction generated by vertical efficiencies.

Finally, our results may also suggest a new understanding of how integrated systems work. A popular question in health care policy has been why fully-integrated systems, like Kaiser Permanente in California, where the insurer is integrated with physicians *and* hospitals, have been successful at keeping costs low. One reason may be the efficiencies we find are larger when all parties are coordinated. One might also think that Kaiser employing its physicians allows it to use strong incentives to direct physician care decisions. Our results suggest that is untrue: Insurers are perfectly able to exert incentives over PCPs, but our results suggest that even high-powered incentives may not have a large effect. Instead, we suggest that the success of Kaiser may be that its ownership of PCPs cuts off the influence of any specialty care providers who may wish to use PCPs to increase costs. It may be fruitful for future work to consider how integration between insurers and PCPs may be able to help contain cost growth in health care.

This paper should not be seen as the end of work in this vein. We are unable to directly observe the contracts both between insurers and physicians, as well as between physicians and the systems they are affiliated with. For one, our estimates only show the effect of global budgets at their current average level, whereas a richer analysis might be able to use variation in incentive strength. In addition, we estimate differences in the strength of steering incentives across systems. Future analysis should seek to discover whether these differences are generated by something about the system's structure and profitability, or from differences in managerial ability.

Chapter 3

Tables

Table 3.1: 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}.

	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%
% Male - Emp. & Dep.	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

Table 3.2: 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.

	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.

Table 3.3: 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).

	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.

Table 3.4: 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.

	Model		
	(4) Early Switcher DID	(5) Ext. Validity Truven Weighted DID	(6) 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.

Table 3.5: 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.

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.

Table 3.6: 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.

	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%

Table 3.7: 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 $t_{-3} - t_{-2}$.

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

Table 3.8: 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.

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

Table 3.9: 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.

	% 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%

Table 3.10: Results for regressions examining consumer responses to non-linear contract prices in the HDHP.

	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)	–
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 Appendix 5

Table 3.11: Characteristics of our samples of orthopedists. The first column contains summary statistics for all orthopedic joint specialists who we identify, while the second column computes the same statistics for only those specialists who we are able to link to the Massachusetts Provider Database (MPD). Data on orthopedist age, gender, and whether they are sole practitioners are taken from Medicare's National Provider and Plan Enumeration System. Data on integration is drawn from the MPD. We define vertical integration as sharing a practice, medical group, or contracting network with any primary care provider (PCP). We define affiliation with a large system as being affiliated with any of the eight contracting networks given in Table 3.14. An orthopedist's number of total arthroplasty surgeries is given by the number of patients who file an insurance claim for a procedure they performed with the CPT codes '27130' or '27447.'

	Full Sample	Matched to MPD
% Male	95.9%	97.6%
% In Boston HRR	77.1%	78.1%
% In Sole Practice	12.1%	13.0%
% Matched to MPD	80.1%	100%
% Vertically Integrated with Any PCP	-	97.0%
% Affiliation with Integrated Health System	-	75.2%
No. of total hip/knee arthroplasty surgeries per year		
Mean	59	65
25th	11	13
75th	69	73
N	258	206

Table 3.12: Characteristics of our samples of patients. The first column cotnains the full set of patients we enumerate in Section 2.2. The second column restricts that sample to only patients who can be matched to a PCP. The third column restricts to only patients who can be matched to a PCP, and whose PCP and orthopedist can both be matched to the Massachusetts Provider Database (MPD). A patient is defined as being in the Boston area if they reside in a zip code within the Boston Hospital Referral Region as defined by Wennberg (1996). A patient is defined as having received a surgery if they file a claim with a procedure code given from the list described in Appendix 6.

	Full Sample	Matched to PCP	Matched to MPD
% Male	45.0%	42.7%	42.6%
Average Age	48.9	50.4	50.6
% In Boston Area	75.9%	75.4%	75.5%
% Covered by Employer-Sponsored Insurance	76.9%	75.7%	77.5%
% Receives Surgery Within 1 Year	18.6%	19.3%	17.7%
Avg. 1-Year Post Spending (2012,2013)	$14,013	$12,935	$12,218
N	222,380	167,183	124,131
PCPs	-	5,550	4,038
Surgeons	262	258	206

Table 3.13: Shares of patients covered by global budget capitation contracts, for different patient insurance coverage categories.

Category	Share of Patients in GB Contract
Employer-Sponsored HMO	0.65
Employer-Sponsored PPO	<0.01
Blue Cross Employer-Sponsored HMO	0.86
HPHC Employer-Sponsored HMO	0.46
Tufts Employer-Sponsored HMO	0.58
2012	0.37
2013	0.42
2014	0.47

Table 3.14: Shares of PCPs and orthopedists who are affiliated with one of the eight largest integrated health systems in Massachusetts.

Health System	PCP Share	Orthopedist Share
Atrius	0.09	0.06
Baycare	0.04	0.05
Beth Israel	0.11	0.09
Lahey	0.04	0.03
NEQCA	0.08	0.11
Partners	0.22	0.30
Steward	0.10	0.17
UMass	0.07	0.04

Table 3.15: Estimated parameters from risk-adjusted cost model. γ_k are orthopedist-specific intercepts for expected log 1-year costs, which we define as the 'cost' of seeing k. η represents the effect of having a PCP who is integrated with the orthopedist seen on log 1-year spending, which we interpret as the cost efficiencies generated (and passed through to patients) of PCPs and orthopedists being integrated. In each column, we add additional sets of controls. In the fourth column, we apply an empirical Bayes shrinkage procedure on the γ_k estimates, described in Appendix 6. The units of the rows are in log costs, which can be approximately interpreted as percent changes.

Standard Deviation of γ_k	0.339	0.308	0.304	0.294
η (Vertical Efficiencies)	-0.034	-0.043	-0.058	-0.058
	(0.008)	(0.008)	(0.008)	(0.008)
Patient Demographic Controls	X	X	X	X
Patient Insurance Controls		X	X	X
System Controls			X	X
Shrinkage Estimator				X
R^2	0.14	0.17	0.17	

Table 3.16: Estimated parameters from risk-adjusted two-outcomes model. This table presents the same parameters as Table 3.15, except for two new models. In the first, we estimate orthopedist and integration effects for the likelihood of a patient being treated by an orthopedic surgery. In the second, we estimate orthopedist effects on cost as in Table 3.15, but add an additional control for whether or not the patient received a surgery. θ is the estimated coefficient for that control variable.

Surgery Propensity				
Standard Deviation of γ_k^{surg}	0.104	0.104	0.104	0.103
η (Vertical Efficiencies)	-0.010	-0.010	-0.012	-0.012
	(0.002)	(0.002)	(0.002)	(0.002)
Cost Conditional on Treatment				
Standard Deviation of γ_k^{other}	0.267	0.239	0.233	0.206
η (Vertical Efficiencies)	-0.018	-0.027	-0.039	-0.039
	(0.007)	(0.007)	(0.007)	(0.007)
θ (Cost Effect of Surgery)	1.566	1.558	1.559	1.559
	(0.007)	(0.007)	(0.007)	(0.007)
Patient Demographic Controls	X	X	X	X
Patient Insurance Controls		X	X	X
System Controls			X	X
Shrinkage Estimator				X

Table 3.17: This table presents distributions of γ_k fixed effects for orthopedists within the eight larg health systems in Massachusetts. The first two columns represent the mean and standard deviation of the fixed effects for orthopedists within a given system, while the third and fourth columns represent the average fixed effects for our decomposition of costs into surgery and non-surgery causes. The fifth column displays a count of the number of orthopedists affiliated with a system.

System	Mean γ_k	SD γ_k	Mean $(\theta\gamma_k^{surg})$	Mean γ_k^{other}	Num. Orthopedists
Atrius	-0.09	0.20	0.02	-0.10	13
Baycare	-0.15	0.14	0.04	-0.18	11
Beth Israel	0.07	0.32	0.03	0.05	20
Lahey	0.14	0.02	-0.00	0.16	7
NEQCA	-0.10	0.17	-0.02	-0.05	24
Partners	0.14	0.29	0.02	0.13	63
Steward	-0.20	0.14	-0.05	-0.13	36
UMass	0.08	0.24	0.01	0.08	9

Table 3.18: Decomposition of surgeon effects. The first column is based on estimates from our standard cost model; the second column is based on estimates after we apply empirical Bayes shrinkage. The first row presents $\frac{\mathrm{Var}(\theta\gamma_k^{surg})}{\mathrm{Var}(\gamma_k)}$. The second presents $\frac{\mathrm{Var}(\gamma_k^{other})}{\mathrm{Var}(\gamma_k)}$, and the third presents $\frac{2\cdot\mathrm{Cov}(\theta\gamma_k^{surg},\gamma_k^{other})}{\mathrm{Var}(\gamma_k)}$.

	Unshrunken	Shrunken
Variance Component of Surgery Costs	28.7%	30.0%
Variance Component of Other Costs	59.0%	49.6%
2x Covariance Component	12.6%	20.4%

Table 3.19: This table presents referral patterns of primary care physicians in our data. Each observation is a PCP. The second column analyzes a subsample of PCPs who have referred at least 30 patients to orthopedists in our data.

	All PCPs	High-Volume PCPs
Avg. # of Referrals	30.7	79.3
Avg. # Unique Orthopedists Referred To	9.17	16.3
Avg. Referral HHI	0.33	0.22
Share Vertically Integrated	0.95	0.95
Avg. # Orthopedists Integrated With	25.5	26.2
Avg. Share of Internal Referrals	0.63	0.65
Expected Log Cost of Orthopedists Referred		
10th Percentile	-0.05	-0.06
25th Percentile	0.04	0.01
50th Percentile	0.14	0.09
75th Percentile	0.28	0.17
90th Percentile	0.41	0.27
N	4038	1064

Table 3.20: Referral patterns of primary care physicians in our data.

System	Internal Ref. Rate	Avg. $\gamma_k + \eta V_{jk}$ of Referred Orthopedist		
		All Referrals	Internal Referrals	External Referrals
All PCPs	0.65	0.08	0.05	0.13
Atrius	0.58	0.06	-0.00	0.17
Baycare	0.82	-0.01	-0.01	0.00
Beth Israel	0.49	0.11	0.13	0.09
Lahey	0.66	0.21	0.25	0.12
NEQCA	0.49	0.03	-0.01	0.08
Partners	0.76	0.15	0.17	0.09
Steward	0.68	-0.02	-0.10	0.15
UMass	0.56	0.18	0.22	0.13

Table 3.21: Regressions of patient allocation to surgeons (based on cost and surgery propensity) on global budget contract utilization.

	$\gamma_{k(i)} + \eta V_{j(i)k(i)}$ (1)	$\gamma_{k(i)} + \eta V_{j(i)k(i)}$ (2)	$\gamma^{surg}_{k(i)} + \eta^{surg} V_{j(i)k(i)}$ (1)	$\gamma^{surg}_{k(i)} + \eta^{surg} V_{j(i)k(i)}$ (2)	$\gamma^{other}_{k(i)} + \eta^{other} V_{j(i)k(i)}$ (1)	$\gamma^{other}_{k(i)} + \eta^{other} V_{j(i)k(i)}$ (2)
β^0	-0.061	-0.041	-0.008	-0.003	-0.047	-0.036
	(0.004)	(0.005)	(0.001)	(0.002)	(0.003)	(0.003)
β^{Atrius}		-0.006		-0.000		-0.006
		(0.006)		(0.002)		(0.004)
$\beta^{Baycare}$		0.042		0.020		0.010
		(0.010)		(0.003)		(0.007)
$\beta^{BethIsrael}$		-0.012		-0.007		-0.001
		(0.007)		(0.002)		(0.005)
β^{Lahey}		0.039		-0.000		0.040
		(0.009)		(0.003)		(0.006)
β^{NEQCA}		0.004		-0.006		0.015
		(0.006)		(0.002)		(0.004)
$\beta^{Partners}$		-0.019		-0.005		-0.010
		(0.005)		(0.002)		(0.003)
$\beta^{Steward}$		-0.018		-0.011		-0.001
		(0.006)		(0.002)		(0.004)
β^{UMass}		-0.030		-0.009		-0.015
		(0.009)		(0.003)		(0.006)
ζ^{Atrius}		-0.002		-0.002		0.001
		(0.003)		(0.001)		(0.002)
$\zeta^{Baycare}$		-0.077		0.006		-0.088
		(0.006)		(0.002)		(0.004)
$\zeta^{BethIsrael}$		0.056		-0.009		0.070
		(0.004)		(0.001)		(0.002)
ζ^{Lahey}		0.117		-0.012		0.137
		(0.005)		(0.002)		(0.003)
ζ^{NEQCA}		-0.032		-0.017		-0.005
		(0.003)		(0.001)		(0.002)
$\zeta^{Partners}$		0.085		-0.011		0.103
		(0.003)		(0.001)		(0.002)
$\zeta^{Steward}$		-0.075		-0.022		-0.040
		(0.003)		(0.001)		(0.002)
ζ^{UMass}		0.136		0.033		0.085
		(0.005)		(0.001)		(0.003)
N	124,131	124,131	124,131	124,131	124,131	124,131

Table 3.22: Regressions of PCP tendency to refer to integrated surgeons on global budget contract utilization.

	$V_{j(i)k(i)}$	
	(1)	(2)
β^0	0.021	0.083
	(0.009)	(0.011)
β^{Atrius}		0.005
		(0.012)
$\beta^{Baycare}$		-0.078
		(0.020)
$\beta^{BethIsrael}$		-0.048
		(0.015)
β^{Lahey}		-0.033
		(0.019)
β^{NEQCA}		-0.112
		(0.013)
$\beta^{Partners}$		-0.108
		(0.011)
$\beta^{Steward}$		-0.030
		(0.012)
β^{UMass}		-0.186
		(0.019)
ζ^{Atrius}		-0.143
		(0.007)
$\zeta^{Baycare}$		0.127
		(0.012)
$\zeta^{BethIsrael}$		-0.196
		(0.008)
ζ^{Lahey}		-0.043
		(0.011)
ζ^{NEQCA}		-0.167
		(0.007)
$\zeta^{Partners}$		0.105
		(0.006)
$\zeta^{Steward}$		-0.006
		(0.007)
ζ^{UMass}		-0.076
		(0.010)
N	119,273	119,273

Table 3.23: Decomposition of effect of global budgets into reallocation to surgeons who perform less surgeries and surgeons who incur less costs conditional on surgery. Computation for “Surgery Channel” is $\frac{\theta\beta^{surg}}{\beta}$, computation for “Other Costs Channel” is $\frac{\beta^{other}}{\beta}$. β estimates for "Total" row come from third and fifth columns in Table 3.21. Estimates for system-specific rows come from fourth and sixth) regressions in Table 3.21.

	GB Effect	Surgery Channel	Other Costs Channel
Total	-0.061	22.5%	77.5%
Atrius	-0.048	7.87%	88.4%
Beth Israel	-0.054	28.1%	68.6%
NEQCA	-0.037	37.5%	57.3%
Partners	-0.061	20.1%	76.9%
Steward	-0.060	34.9%	62.0%
UMass	-0.072	25.8%	71.5%

Table 3.25: Decomposition of organization-specific savings from global budgets into three categories. "Internal" is the share of savings from reallocating patients from higher-cost within-organization surgeons to lower-cost within-organization surgeons. "External" is the share of savings from reallocating patients from higher-cost outside-organization surgeons to lower-cost outside-organization surgeons. "Between" is the share of savings from moving patients from within-organization surgeons to outside-organization surgeons (or vice versa). A negative value implies that the firm *increased* costs on that margin.

		Savings From Reallocation Method		
	GB Effect	Internal	External	Cross-Organization
Atrius	-0.048	-6.41%	58.2%	32.6%
Beth Israel	-0.054	2.97%	98.9%	-2.72%
NEQCA	-0.037	25.8%	80.3%	-7.34%
Partners	-0.061	74.0%	26.2%	3.13%
Steward	-0.060	35.4%	22.1%	22.5%
UMass	-0.072	27.6%	56.5%	12.5%

Table 3.26: Results from our analysis of extensive margin responses. The dependent variable in these regressions is a binary indicator for whether the patient was referred or not. β measures the effect of global budget contracts on the share of patients who are referred.

	$Referred_i$	
	(1)	(2)
$\beta^{Ext,0}$	-0.037	-0.033
	(0.001)	(0.002)
$\beta^{Ext,Atrius}$		0.010
		(0.002)
$\beta^{Ext,BethIsrael}$		-0.009
		(0.003)
$\beta^{Ext,NEQCA}$		-0.002
		(0.002)
$\beta^{Ext,Partners}$		-0.007
		(0.002)
$\beta^{Ext,Steward}$		-0.000
		(0.002)
$\beta^{Ext,UMass}$		-0.004
		(0.003)
N	1,471,139	1,471,139

Table 3.27: Parameter estimates from our model of orthopedist referral choice. The first row presents average values of β^0 (PCP sensitivity to costs), β^{GB} (the effect of global budget contracts on cost-sensitivity), and T (the strength of steering incentives) over patients in our data. The subsequent rows present average values of these parameters for patients in our data whose PCP is part of a given system. All parameters are measured in utility units. Standard errors are computed via bootstrap and are given in parentheses.

	β^0	β^{GB}	T
Average	-0.02	-0.56	1.63
	(0.05)	(0.05)	(0.01)
Atrius	-0.01	-0.63	2.65
	(0.05)	(0.08)	(0.02)
Beth Israel	-0.02	-0.35	1.82
	(0.05)	(0.12)	(0.03)
NEQCA	-0.04	-0.69	1.48
	(0.05)	(0.11)	(0.02)
Partners	0.04	-0.61	1.81
	(0.05)	(0.07)	(0.02)
Steward	-0.10	-0.80	1.40
	(0.05)	(0.10)	(0.02)
UMass	-0.07	-0.50	1.04
	(0.05)	(0.17)	(0.04)

Table 3.28: The percentage of patients who are referred to orthopedists who are vertically tied to their PCP, in three counterfactual simulations. In all three simulations, no patients are covered by global budget contracts. The first column presents results for the integration status quo, the second presents results when we remove efficiencies ($\eta = 0$), and the third presents results when we remove vertical ties ($V_{jk} = 0$).

	Internal Referral Rate		
	Status Quo	No Efficiencies	No Integration
Total	62.8%	61.9%	25.6%
	(0.1)	(0.1)	(0.1)
Atrius	57.1%	55.5%	9.40%
	(0.4)	(0.4)	(0.1)
Beth Israel	49.1%	48.4%	15.2%
	(0.5)	(0.5)	(0.1)
NEQCA	49.6%	47.7%	23.3%
	(0.4)	(0.3)	(0.2)
Partners	76.9%	75.7%	38.1%
	(0.2)	(0.2)	(0.2)
Steward	67.3%	66.7%	42.2%
	(0.4)	(0.4)	(0.3)
UMass	56.9%	56.2%	34.8%
	(0.6)	(0.7)	(0.6)

Table 3.29: Average γ_k values for orthopedists who patients are referred to, in four counterfactual simulations. In the first and second columns, integration is left at status quo levels, while in the third and fourth, we remove all vertical ties ($V_{jk} = 0$). In the first and third columns, no patients are covered by global budgets, whereas in the second and fourth columns, all patients are covered.

	Avg. γ_k of Orthopedist Chosen			
	No Global Budgets		Global Budgets	
	Status Quo	No Integration	Status Quo	No Integration
Total	0.095	0.141	0.067	0.110
	(0.001)	(0.002)	(0.001)	(0.002)
Atrius	0.086	0.180	0.052	0.138
	(0.003)	(0.004)	(0.002)	(0.003)
Beth Israel	0.127	0.153	0.101	0.128
	(0.004)	(0.004)	(0.005)	(0.005)
NEQCA	0.048	0.078	0.018	0.044
	(0.003)	(0.003)	(0.003)	(0.003)
Partners	0.169	0.171	0.132	0.131
	(0.002)	(0.003)	(0.003)	(0.003)
Steward	-0.007	0.075	-0.039	0.038
	(0.002)	(0.003)	(0.003)	(0.003)
UMass	0.198	0.182	0.170	0.153
	(0.005)	(0.006)	(0.006)	(0.006)

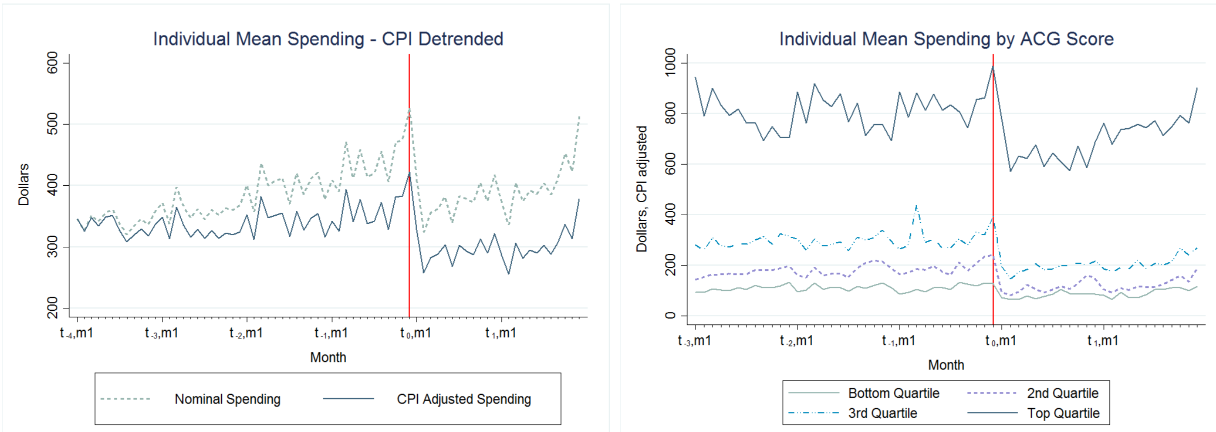
Table 3.30: Recreation of Table 3.29, however, in all simulations we have removed efficiencies ($\eta = 0$).

	Avg. γ_k of Orthopedist Chosen			
	No Global Budgets		Global Budgets	
	Status Quo	No Integration	Status Quo	No Integration
Total	0.131	0.141	0.104	0.110
	(0.001)	(0.002)	(0.001)	(0.002)
Atrius	0.121	0.180	0.089	0.138
	(0.003)	(0.004)	(0.002)	(0.003)
Beth Israel	0.156	0.153	0.130	0.128
	(0.004)	(0.004)	(0.005)	(0.005)
NEQCA	0.078	0.078	0.047	0.044
	(0.003)	(0.003)	(0.003)	(0.003)
Partners	0.213	0.171	0.174	0.131
	(0.002)	(0.003)	(0.003)	(0.003)
Steward	0.032	0.075	0.003	0.038
	(0.002)	(0.003)	(0.002)	(0.003)
UMass	0.230	0.182	0.199	0.153
	(0.005)	(0.006)	(0.006)	(0.006)

Chapter 4

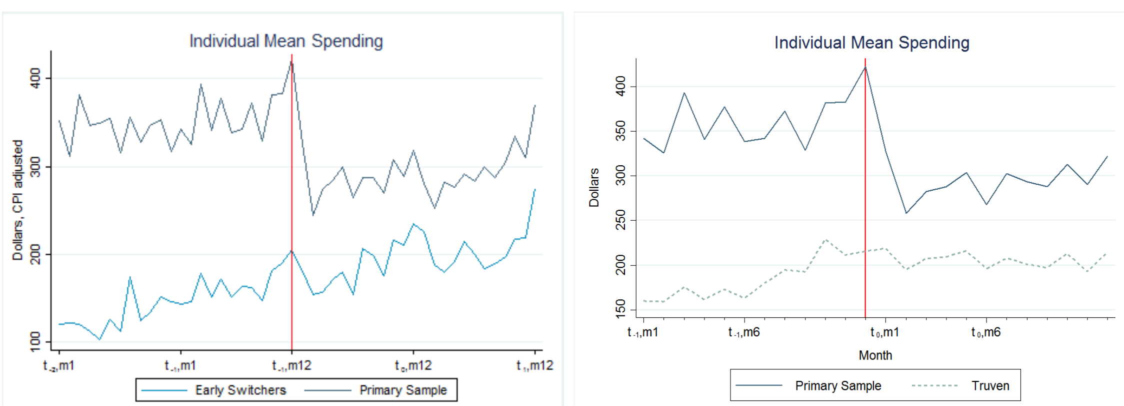
Figures

Figure 4.1: Incremental Spending Time Series



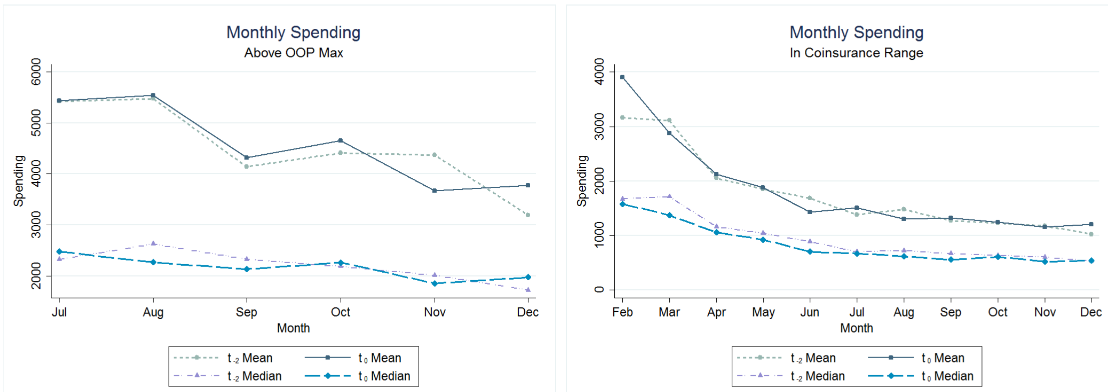
This left panel in this figure plots mean monthly spending by individuals in our primary sample over the six years in our data, both adjusted and unadjusted for age and price trends. The right panel plots adjusted spending for individuals in a given month, by ACG predictive health index quartile (the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

Figure 4.2: Difference-in-Differences Time Series Analysis: Early Switchers and Truven Control Group



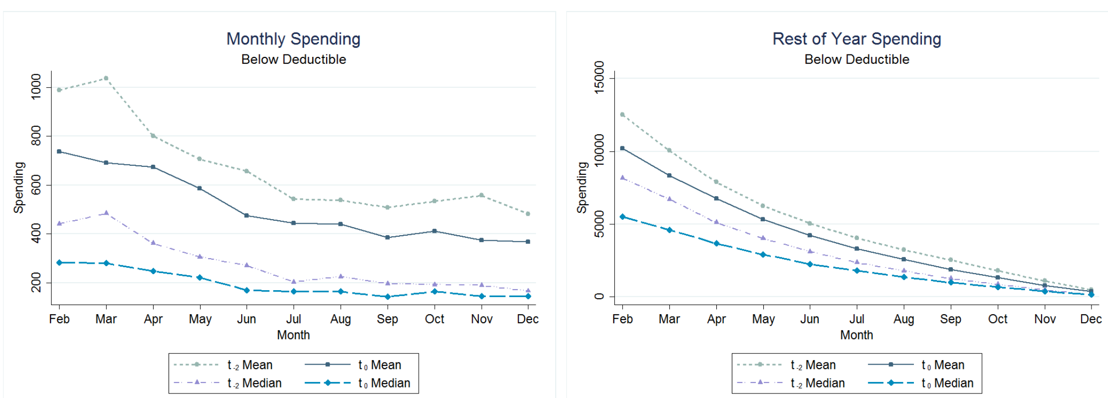
The left panel presents unweighted spending over time for early switchers to the HDHP alongside our primary sample. The right panel presents spending for our primary sample alongside spending for the weighted control group formed from Truven MarketScan data.

Figure 4.3: Incremental Spending for Employees Over Out-of-Pocket Maximum and in Coinsurance Arm



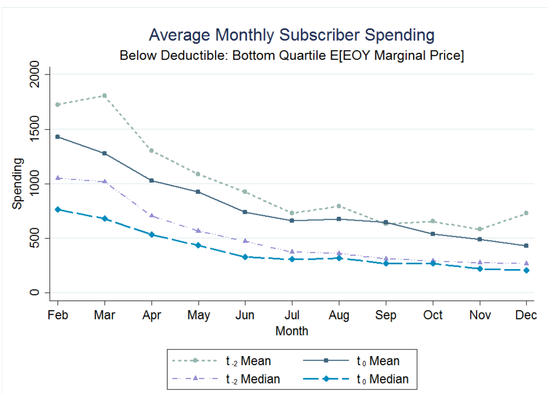
The left panel in this figure shows incremental spending for the next month, for families who have passed the out-of-pocket maximum by the start of a given month in t_0, compared to t_{-2} incremental spending for equivalent quantiles of pre-period consumers. The right panel presents the analogous figure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

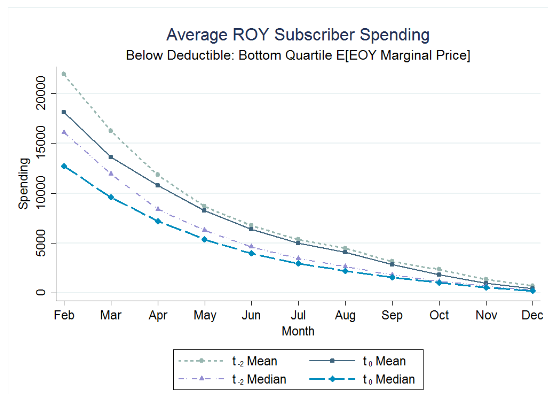
Figure 4.4: Incremental and Rest-of-Year Spending for Employees Under Deductible



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 4.5: 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.

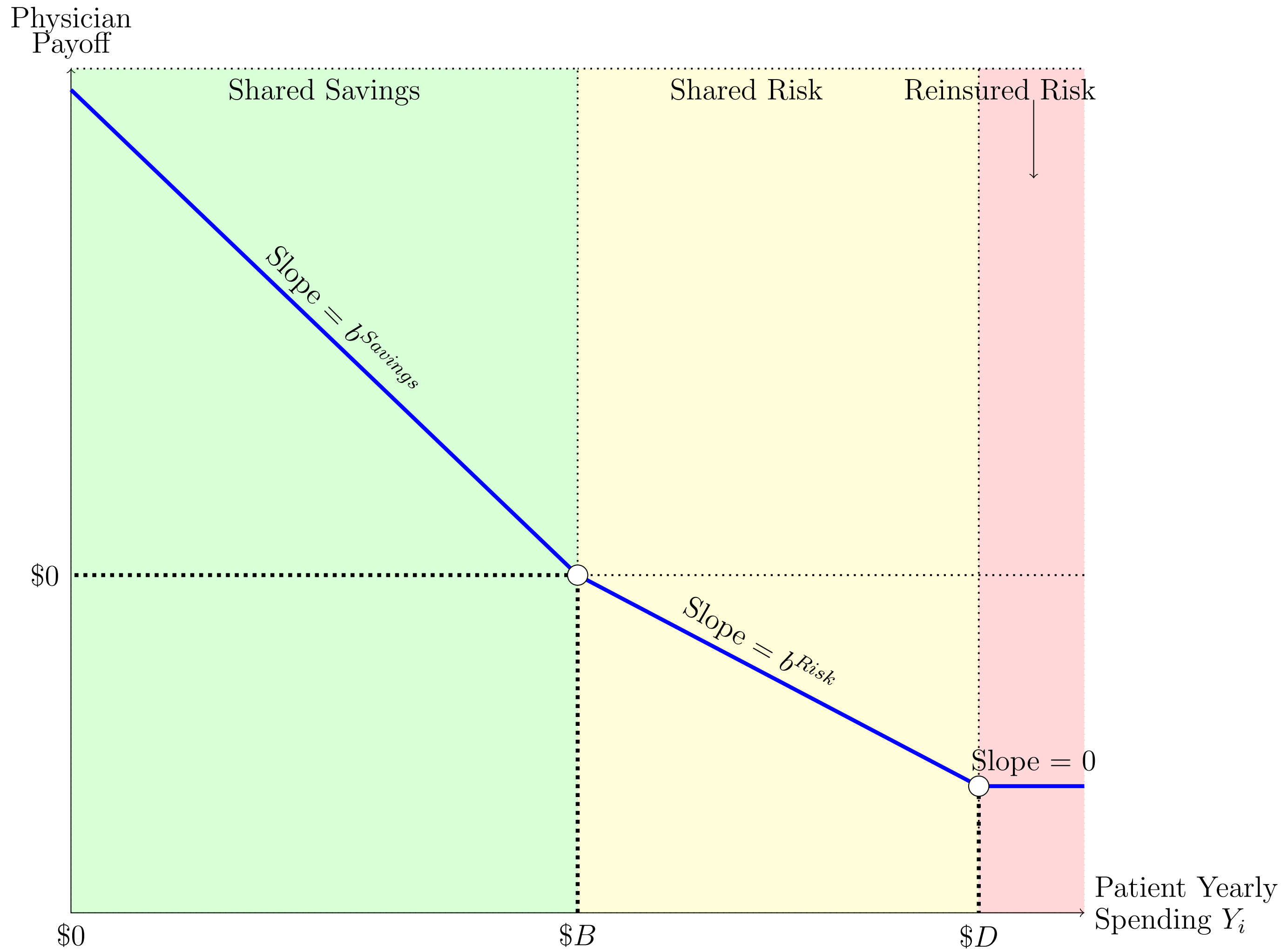


Figure 4.6: The blue line in this figure depicts a hypothetical global budget contract. The x-axis is the amount of total medical expenditures a given patient incurs during a year, whereas the y-axis is the total dollar amount transferred from the insurer to the patient's primary care provider. The contract involves a budget B, and risk-sharing rates $b^{Savings}$ and b^{Risk}. If the patient's spending for the year is below $\$B$, the PCP is in the green "Shared Savings" region, and receives $\$b^{Savings}$ from the insurer for each dollar of relative savings. If the patient's spending is above B, the PCP is in the yellow "Shared Risk" region, and must pay the insurer $\$b^{Risk}$ for each dollar of excess spending. Some insurers require that PCPs must carry reinsurance against tail patient spending risk. This reinsurance typically takes on the form of a deductible contract. As such, when the patient's spending goes above the deductible $\$D$, the remaining risk is borne by the reinsurer–every dollar that the PCP must pay the insurer is instead paid by the reinsurer.

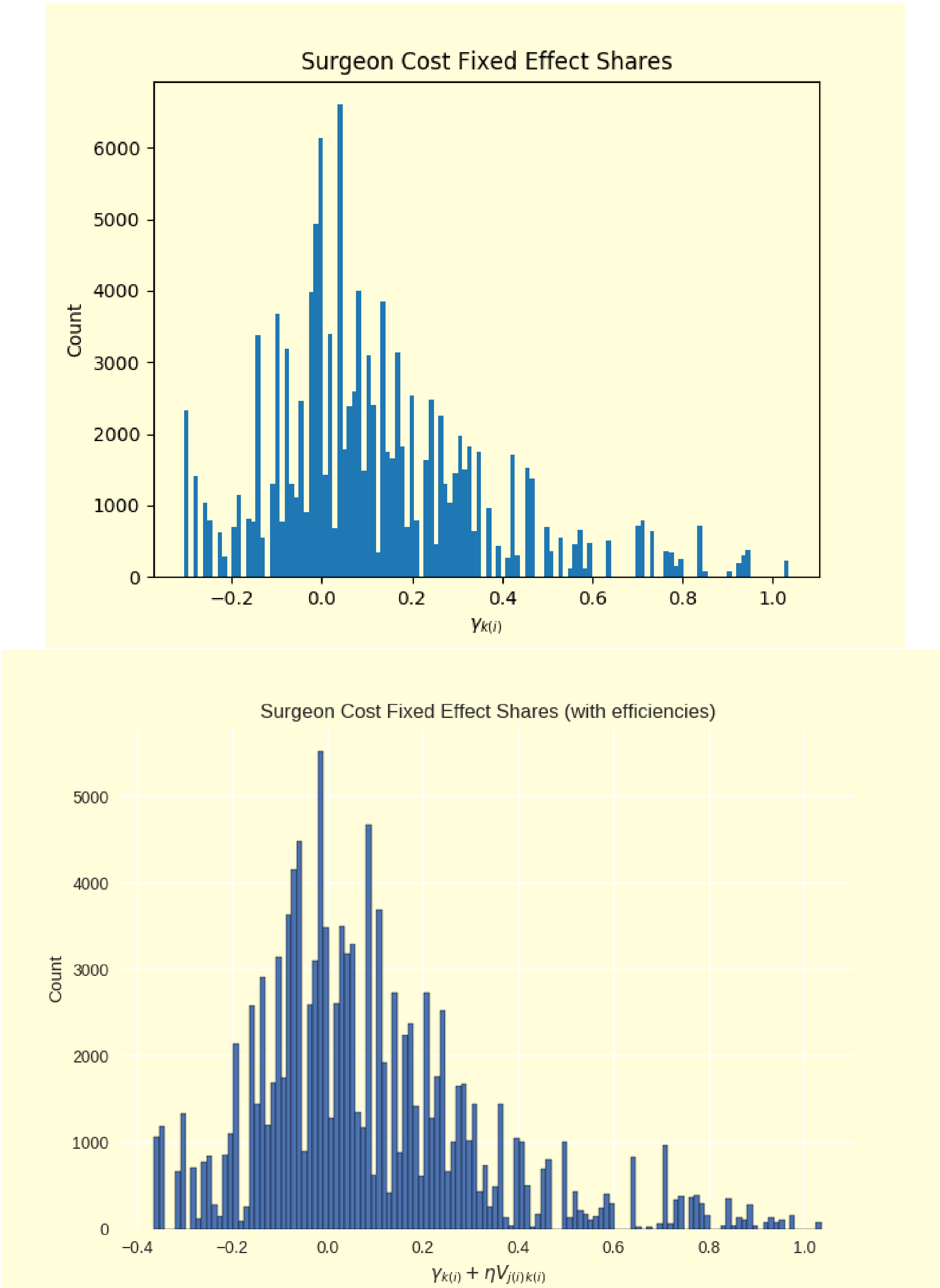


Figure 4.7: Distribution of surgeon fixed effects $\gamma_{k(i)}$ and expected surgeon costs $\gamma_{k(i)} + \eta V_{j(i)k(i)}$, by patient.

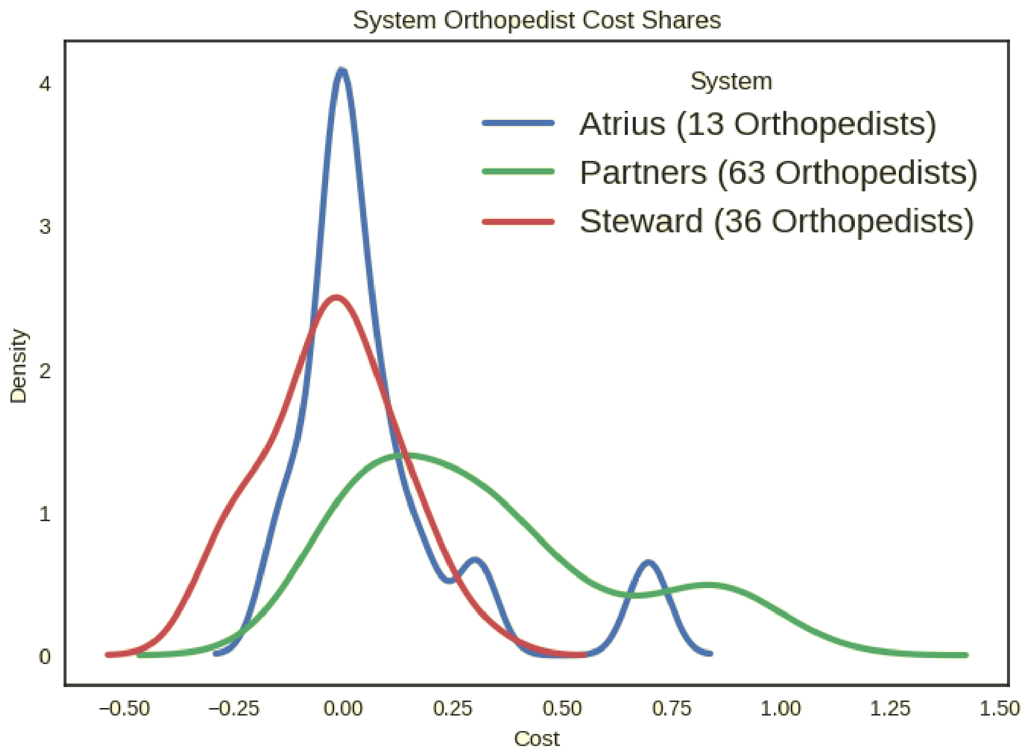


Figure 4.8: Distribution of surgeon fixed effects $\gamma_{k(i)}$ for orthopedists in three major health systems: Atrius, Partners, and Steward..

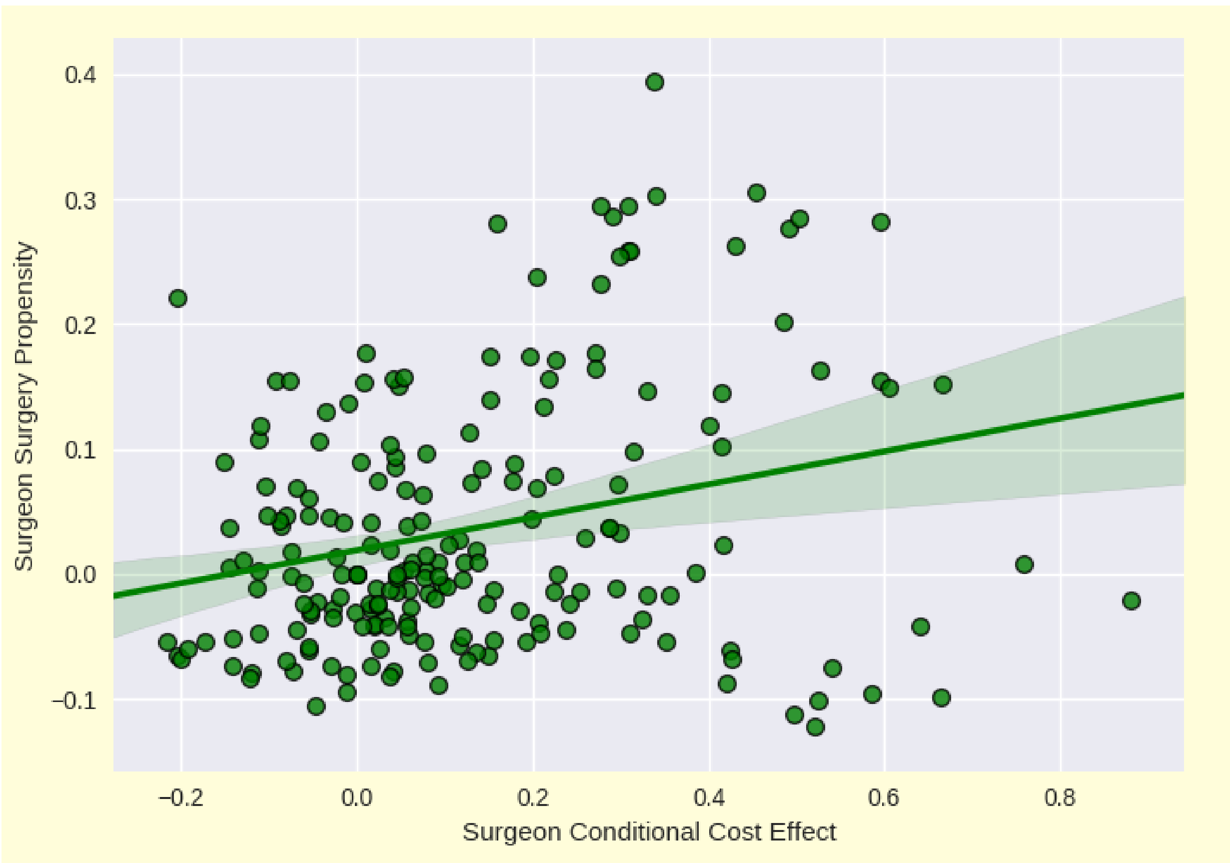


Figure 4.9: Scatterplot depicting the joint distribution of γ_k^{surg}, an orthopedist k's propensity to do surgery, against γ_k^{other}, his propensity to incur costs conditional on a surgery decision.

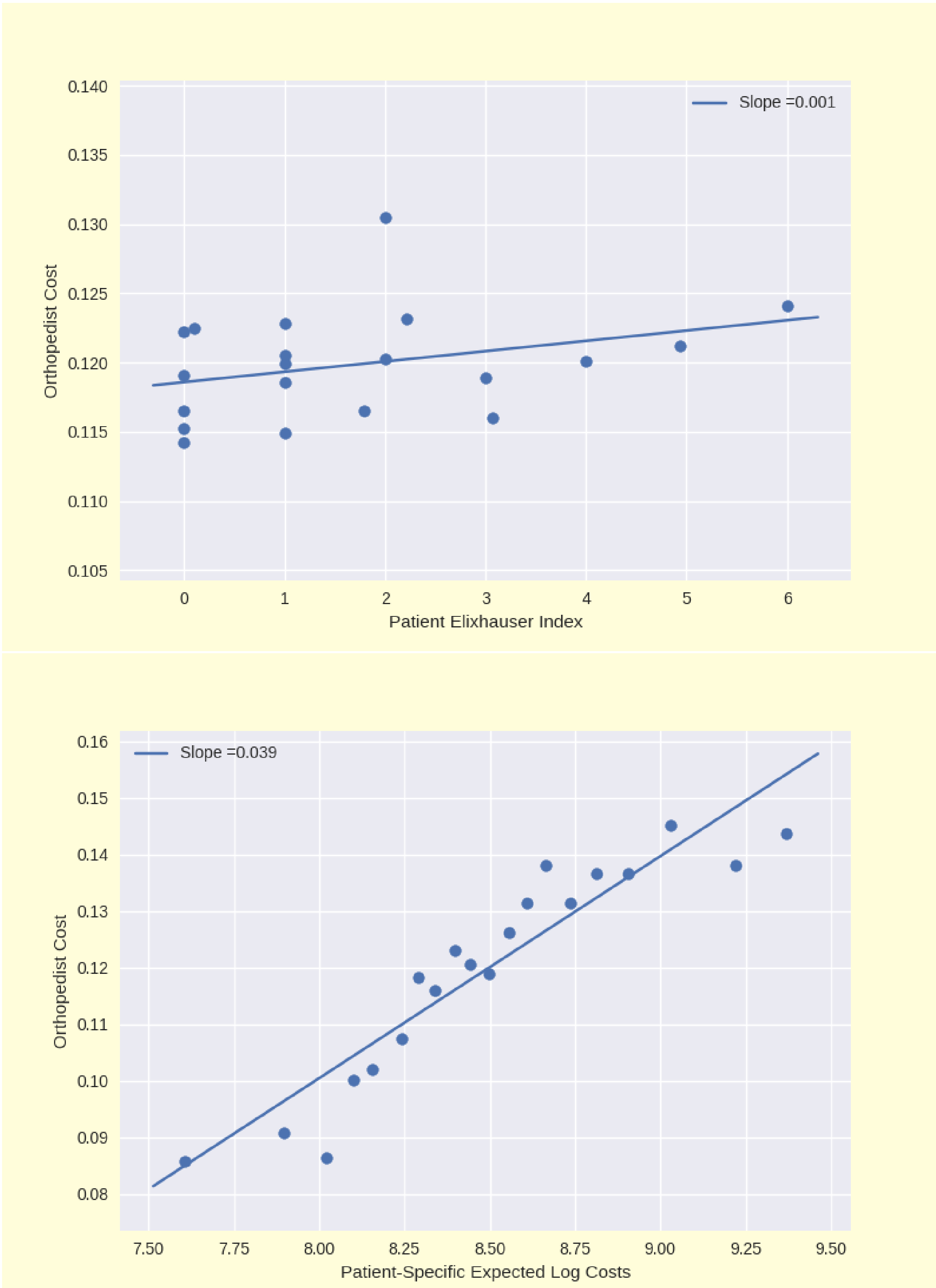


Figure 4.10: Binned scatterplot depicting the average γ_k of the orthopedist chosen for patients, binned by Elixhauser index (top figure) and δX_i (bottom figure).

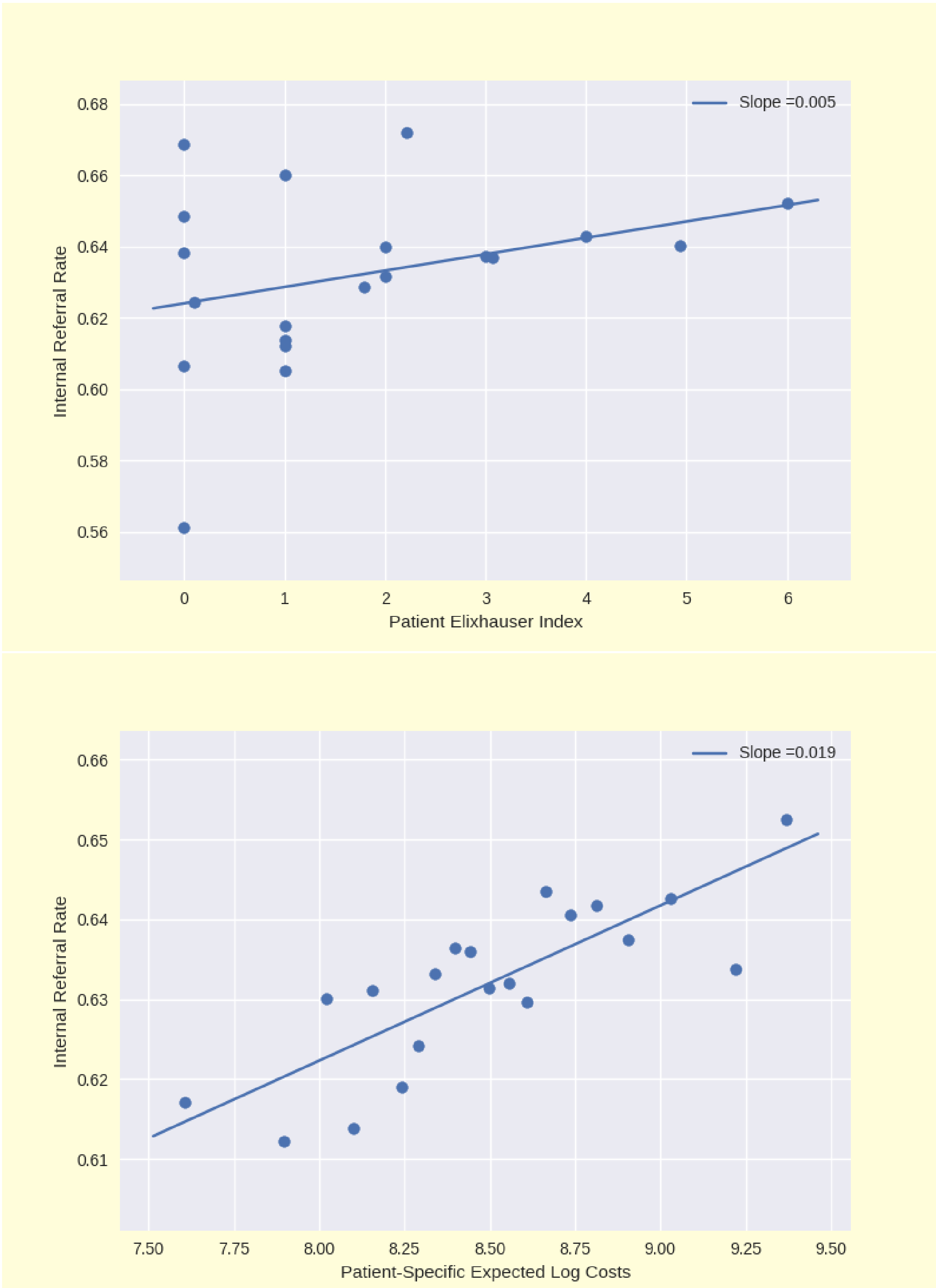


Figure 4.11: Binned scatterplot depicting the rate of referrals to integrated orthopedists, binned by Elixhauser index (top figure) and δX_i (bottom figure).

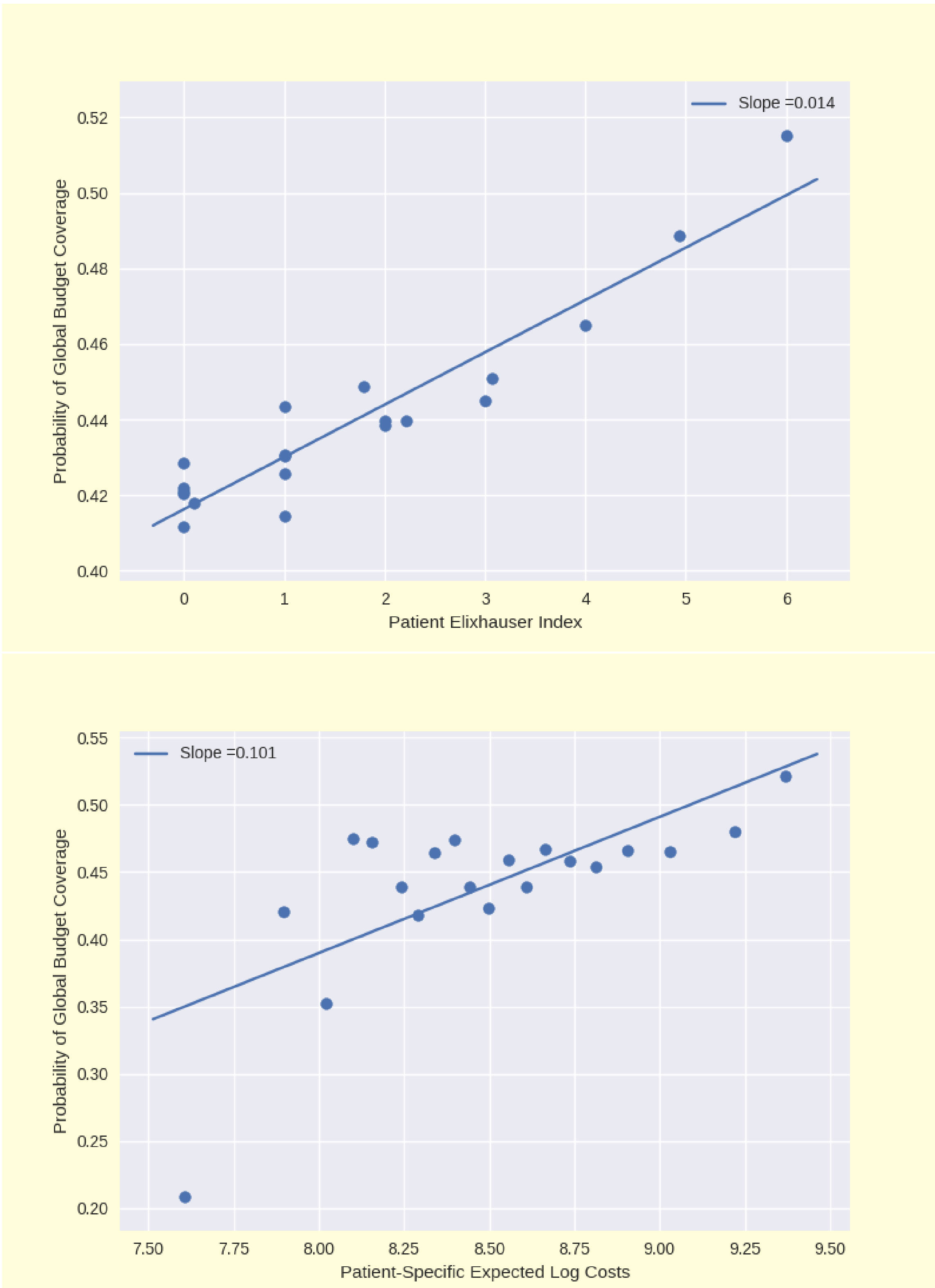


Figure 4.12: Binned scatterplot depicting the average probability of being covered by a global budget capitation contract, binned by Elixhauser index (top figure) and δX_i (bottom figure).

Bibliography

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

Afendulis, Christopher C. and Daniel P. Kessler, "Tradeoffs from Integrating Diagnosis and Treatment in Markets for Health Care," *American Economic Review*, 2007, *97* (3), 1013–1020.

Agha, Leila, Keith Marzilli Ericson, Kimberley H. Geissler, and James B. Rebitzer, "Team Formation and Performance: Evidence from Healthcare Referral Networks," February 2018. NBER Working Paper No. 24338.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

Anastov, Pavel and Tom Baker, "Putting Health Back Into Health Insurance Choice," *Medical Care Research Review*, August 2014, *71* (4), 337–355.

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

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

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 1963, *53* (5), 941–973.

Atalay, Enghin, Ali Hortaçsu, Mary Jialin Li, and Chad Syverson, "How Wide Is the Firm Border?," 2017. NBER Working Paper No. 23777.

Backus, Matt, "Why is Productivity Correlated with Competition?," 2014.

Baicker, Katherine, Amitabh Chandra, and Jonathan S. Skinner, "Saving Money or Just Saving Lives? Improving the Productivity of US Health Care Spending," *Annual Review of Economics*, 2012, *4* (1), 33–56.

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

Baker, Laurence C., M. Kate Bundorf, and Daniel P. Kessler, "Vertical Integration: Hospital Ownership Of Physician Practices Is Associated With Higher Prices And Spending," *Health Affairs*, 2014, *33* (5), 756–763.

_ , _ , and _ , "Does Health Plan Generosity Enhance Hospital Market Power?," *Journal of Health Economics*, 2015, *44*, 54–62.

_ , _ , and _ , "The Effect of Hospital/Physician Integration on Hospital Choice," *Journal of Health Economics*, 2016, *50* (1), 1–8.

_ , _ , and _ , "Does Multispecialty Practice Enhance Physician Market Power?," September 2017. NBER Working Paper No. 21104.

Barwick, Panle Jia, Parag Pathak, and Maisy Wong, "Conflicts of Interest and Steering in Residential Brokerage," *American Economic Journal: Applied Economics*, 2017, *9* (3), 191–222.

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

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes, "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review*, 2006, *96*, 988–1012.

Blinder, Alan S., "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 1973, *8* (4), 436–455.

Brot-Goldberg, Zarek C., Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad, "What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics," *Quarterly Journal of Economics*, 2017, *132* (3), 1261–1318.

Brown, Zach Y., "An Empirical Model of Price Transparency and Markups in Health Care," 2018.

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*, 2011, *17* (3), 222–230.

Burns, Lawton R. and Mark V. Pauly, "Integrated Delivery Networks: A Detour On The Road To Integrated Health Care?," *Health Affairs*, 2002, *21* (4), 128–143.

_ **and** _ , "Accountable Care Organizations May Have Difficulty Avoiding The Failures Of Integrated Delivery Networks Of The 1990s," *Health Affairs*, 2012, *31* (11), 2407–2416.

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

Capps, Cory, David Dranove, and Christopher Ody, "Physician Practice Consolidation Driven By Small Acquisitions, So Antitrust Agencies Have Few Tools To Intervene," *Health Affairs*, 2017, *36* (9), 1556–1563.

_ **,** _ **, and** _ , "The Effect of Hospital Acquisitions of Physician Practices on Prices and Spending," *Journal of Health Economics*, 2018, *59*, 139–152.

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

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson, "Healthcare Exceptionalism? Performance and Allocation in the U.S. Healthcare Sector," *Journal of Political Economy*, 2016, *106* (8), 2110–2144.

_ **and Douglas O. Staiger**, "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy*, 2007, *115* (1), 103–140.

_ **, 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*, 2007, *26* (2), 195–203.

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

_ **, Zack Cooper, Eugene Larsen-Hallock, and Fiona Scott Morton**, "Are Health Care Services Shoppable? Evidence from the Consumption of Lower-Limb MRI Scans," July 2018. NBER Working Paper No. 24869.

Chipty, Tasneem, "Vertical Integration, Market Foreclosure, and Consumer Welfare in the Cable Television Industry," *American Economic Review*, 2001, *91* (3), 428–453.

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

Coase, Ronald H., "The Nature of the Firm," *Economica*, 1937, *4* (16), 386–405.

Colla, Carrie H., David E. Wennberg, Ellen Meara, Jonathan S. Skinner, Daniel Gottlieb, Valerie A Lewis, Christopher M. Snyder, and Elliott S. Fisher, "Spending Differences Associated With the Medicare Physician Group Practice Demonstration," *Journal of the American Medical Association*, 2012, *308* (10), 1015–1023.

Cooper, Zack, Stuart V. Craig, Martin Gaynor, and John Van Reenen, "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured," *Quarterly Journal of Economics*, forthcoming.

Cuellar, Alison E. and Paul J. Gertler, "Strategic Integration of Hospitals and Physicians," *Journal of Health Economics*, 2006, *25* (1), 1–28.

Cutler, David, "From the Affordable Care Act to Affordable Care," *Journal of the American Medical Association*, 2015, *314* (4), 337–338.

Cutler, David M., "The Incidence of Adverse Medical Outcomes Under Prospective Payment," *Econometrica*, 1995, *63* (1), 29–50.

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

Egan, Mark, "Brokers vs. Retail Investors: Conflicting Interests and Dominated Products," 2018.

Einav, Liran, Amy Finkelstein, and Mark Cullen, "Estimating Welfare in Insurance Markets Using Variation in Prices," *Quarterly Journal of Economics*, 2010, *125* (3), 877–921.

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

_ , _ , Stephen P. Ryan, Paul Schrimpf, and Mark R. Cullen, "Selection on Moral Hazard in Health Insurance," *American Economic Review*, 2013, *103* (1), 178–219.

Elixhauser, Anne, Claudia Steiner, D Robert Harris, and Rosanna M. Coffey, "Comorbidity measures for use with administrative data," *Medical care*, 1998, pp. 8–27.

Ellis, Randall P. and Thomas G. McGuire, "Provider Behavior Under Prospective Reimbursement: Cost Sharing and Supply," *Journal of Health Economics*, 1986, *5* (1), 129–151.

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

Ericson, Keith M. and Amanda Starc, "Measuring Consumer Valuation of Limited Provider Networks," *American Economic Review: Papers & Proceedings*, 2015, *105* (5), 115–119.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams, "Sources of Geographic Variation in Health Care: Evidence From Patient Migration," *Quarterly Journal of Economics*, 2016, *131* (4), 1681–1726.

_ , 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*, 2012, *127* (3), 1057–1106.

Forbes, Silke J. and Mara Lederman, "Does Vertical Integration Affect Firm Performance? Evidence from the Airline Industry," *RAND Journal of Economics*, 2010, *41* (4), 765–790.

Glied, Sherry, "Managed Care," in "Handbook of Health Economics," Vol. 1 2000, pp. 707–753.

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

Handel, Benjamin, Jonathan Holmes, Jonathan Kolstad, and Kurt Lavetti, "Insurer Innovation and Health Care Efficiency: Evidence from Utah," 2018.

Handel, Benjamin R., "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*, 2013, *103* (7), 2643–2682.

_ and Jonathan T. Kolstad, "Health Insurance for "Humans:" Information Frictions, Plan Choice, and Consumer Welfare," *American Economic Review*, August 2015, *105* (8), 2449–2500.

Hart, Oliver and Jean Tirole, "Vertical Integration and Market Foreclosure," *Brookings Papers on Economic Activity: Microeconomics*, 1990, *1990*, 205–286.

Hastings, Justine S. and Richard J. Gilbert, "Market Power, Vertical Integration, and the Wholesale Price of Gasoline," *Journal of Industrial Economics*, 2005, *53* (4), 469–492.

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*, 2016, *46*, 33–51.

_ , 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*, 2012, *31* (5), 1009–1015.

Ho, Kate and Ariel Pakes, "Hospital Choices, Hospital Prices and Financial Incentives to Physicians," *American Economic Review*, 2014, *104* (12), 3841–3884.

_ and Robin Lee, "Insurer Competition in Health Care Markets," *Econometrica*, 2017, *85* (2), 379–417.

Holmström, Bengt, "Moral Hazard and Observability," *Bell Journal of Economics*, 1979, *10* (1), 74–91.

_ and Paul Milgrom, "The Firm as an Incentive System," *American Economic Review*, 1994, *84* (4), 972–991.

Hortaçsu, Ali and Chad Syverson, "Cementing Relationships: Vertical Integration, Foreclosure, Productivity, and Prices," *Journal of Political Economy*, 2007, *115* (2), 250–301.

Ito, Koichiro, "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing," *American Economic Review*, 2014, *104* (2), 537–563.

Jensen, Michael C. and William H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics*, 1976, *3* (4), 305–360.

Kaiser Family Foundation, "Focus on Health Reform: Summary of the Affordable Care Act," 2011. `http://kaiserfamilyfoundation.files.worldpress.com/2011/04/8061-021.pdf`.

_ , "Employee Health Benefits: 2014 Summary of Findings," August 2015.

_ , "Preventive Services Covered by Private Health Plans Under the Affordable Care Act," 2015. `http://kff.org/health-reform/fact-sheet/preventive-services-covered-by-private-health-plans/`.

Kane, Thomas J. and Douglas O. Staiger, "Improving School Accountability Measures," 2001. NBER Working Paper No. 8156.

Keeler, Emmett B. and John E. Rolph, "The Demand For Episodes of Treatment in the Health Insurance Experiment," *Journal of Health Economics*, 1988, *7* (4), 337–367.

Klein, Benjamin, Robert G. Crawford, and Armen A. Alchian, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, 1978, *21* (2), 297–326.

Kocher, Robert and Nikhil R. Saini, "Hospitals' Race to Employ Physicians–The Logic behind a Money-Losing Proposition," *New England Journal of Medicine*, 2011, *364* (19), 1790–93.

Kowalczyk, Liz, "Steward Health Care pressured doctors to restrict referrals outside chain, suit says," *Boston Globe*, May 24 2018.

Lieber, Ethan M.J., "Does it Pay to Know the Prices in Health Care?," *American Economic Journal: Economic Policy*, 2017, *9*, 154–179.

Liebman, Eli, "Bargaining in Markets with Exclusion: An Analysis of Health Insurance Networks," 2018.

Liebman, Jeffrey B. and Richard J. Zeckhauser, "Schmeduling," October 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*, 2010, *77* (1).

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

Massachusetts Health Policy Commission, "2017 Annual Health Care Cost Trends Report," March 2018.

Massachusetts Health Quality Partners, "MHQP 2017 Massachusetts Provider Database (MPD)," 2017. `http://www.mhqp.org/products_and_tools/?content_item_id=226`.

McClellan, Mark and Douglas Staiger, "The Quality of Health Care Providers," 1999. NBER Working Paper No. 7327.

McGuire, Thomas G., "Physician Agency," in "Handbook of Health Economics," Vol. 1 2000, pp. 461–536.

Mechanic, Robert E., Palmira Santos, Bruce E. Landon, and Michael E. Chernew, "Medical Group Responses To Global Payment: Early Lessons From The 'Alternative Quality Contract' In Massachusetts," *Health Affairs*, 2011, *30* (9), 1734–1742.

Medical Group Management Association, "MGMA Physician Compensation and Production Survey: 2014 Report Based on 2013 Data," 2014.

Medscape, "Medscape Physician Compensation Report 2018," 2018. `https://www.medscape.com/slideshow/2018-compensation-overview-6009667`.

Muhlestein, David, Robert Saunders, Robert Richards, and Mark McClellan, "Recent Progress In The Value Journey: Growth Of ACOs And Value-Based Payment Models In 2018," 2018. *Health Affairs Blog*. `https://www.healthaffairs.org/do/10.1377/hblog20180810.481968/full/`.

Nakamura, Sayaka, "Hospital Mergers and Referrals in the United States: Patient Steering or Integrated Delivery of Care?," *Inquiry: The Journal of Health Care Organization, Provision, and Financing*, 2010, *47* (3), 226–241.

_ , Cory Capps, and David Dranove, "Patient Admission Patterns and Acquisitions of "Feeder" Hospitals," *Journal of Economics & Management Strategy*, 2007, *16* (4), 995–1030.

Nevo, Aviv, John L. Turner, and Jonathan W. Williams, "Usage-Based Pricing and Demand for Residential Broadband," *Econometrica*, 2016, *84* (2), 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*, 1973, *14* (3), 693–709.

Ordover, Janusz A., Garth Saloner, and Steven C. Salop, "Equilibrium Vertical Foreclosure," *American Economic Review*, 1990, pp. 127–142.

Rose, Evan, Jonathan Schellenberg, and Yotam Shem-Tov, "The Effects of Teacher Quality on Criminal Behavior," 2018.

Schwartz, Aaron, Bruce Landon, Adam Elshaug, Michael Chernew, and Michael McWilliams, "Measuring Low-Value Care in Medicare," *JAMA Internal Medicine*, 2014, *174* (7), 1067–1076.

Song, Zirui, Dana Gelb Safran, Bruce E. Landon, Yulei He, Randall P. Ellis, Robert E. Mechanic, Matthew P. Day, and Michael E. Chernew, "Health Care Spending and Quality in Year 1 of the Alternative Quality Contract," *New England Journal of Medicine*, 2011, *365* (10), 909–918.

_ , Sherri Rose, Dana G. Safran, Bruce E. Landon, Matthew P. Day, and Michael E. Chernew, "Changes in Health Care Spending and Quality 4 Years into Global Payment," *New England Journal of Medicine*, 2014, *371* (18), 1704–1714.

Sood, Neeraj, Zachary Wagner, Peter J. Huckfeldt, and Amelia M. Haviland, "Price Shopping in Consumer Directed Health Plans," *Forum for Health Economics & Policy*, 2013, *16* (1), 35–53.

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

Swanson, Ashley, "Physician Investment in Hospitals: Specialization, Incentives, and the Quality of Cardiac Care," December 2013.

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.

United States Bone and Joint Initiative, "The Burden of Musculoskeletal Diseases in the United States: Prevalence, Societal and Economic Cost, Third Edition," 2015.

Vera-Hernandez, Marcos, "Structural estimation of a principal-agent model: moral hazard in medical insurance," *RAND Journal of Economics*, 2003, *34* (4), 670–693.

Walden, Emily, "Can Hospitals Buy Referrals? The Impact of Physician Group Acquisitions on Market-Wide Referral Patterns," November 2016.

Wennberg, John E., *The Dartmouth Atlas of Health Care in the United States*, American Hospital Association, 1996.

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," 2014. `https://www.whitehouse.gov/sites/default/files/docs/fact_sheet_implementing_the_affordable_care_act_from_the_erp_2013_final1.pdf`.

Williamson, Oliver, *The Economic Institutions of Capitalism*, Free Press, 1985.

Zeckhauser, Richard, "Medical Insurance: A Case Study of the Tradeoff between Risk Spreading and Appropriate Incentives," *Journal of Economic Theory*, 1970, *2* (1), 10–26.

Ziemba, Justin B., Mohamad E. Allaf, and Dalal Haldeman, "Consumer Preferences and Online Comparison Tools Used to Select a Surgeon," *JAMA Surgery*, 2017, *152* (4), 410–411.

Chapter 5

Appendix For “What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Spending”

Robustness Checks of Primary Sample Construction

The main sample we use throughout the paper is constructed so as to ensure we can analyze long-term trends in spending. We constructed a similar sample using weaker restrictions to show that our sample restrictions are innocuous in terms of their effects on the final result. Our primary sample is restricted to only include employees who were enrolled in a health insurance plan at the firm for all years between t_{-4} and t_1, the entire span of our data. Our alternate sample is only restricted to employees who were enrolled between t_{-2} and t_0, which includes employees who may have left the firm in t_1, or joined it in t_{-4} or t_{-3}. Summary statistics for our main sample and this alternative are given in the first two columns of Table A1. This new sample includes approximately 8,000 additional employees and 10,000 additional dependents. These excluded employees are relatively younger, and have smaller families (mostly those employees who joined the first during t_{-4} or t_{-3}), but the overall mix of ages among them and their dependents changes only slightly. Most importantly, the distribution of health spending is nearly identical.

Another concern with our approach is that, since employees were aware of the policy change well in advance, they might make the decision to leave the firm in advance of being required to switch into a health insurance plan with cost-sharing. To test this, we plot the hazard rate of employees and their dependents exiting the firm's insurance coverage for each month in our data. We see no meaningful change in employee exit either around the announcement date for the plan switch (October of year t_{-3}) or the implementation date (January of year t_0). There is some incremental dependent attrition at the implementation date, but not enough to meaningfully impact our main results.

For those few who do exit in response to the change, one might expect them to be relatively sicker, which might induce a selection bias into our results. To examine this, we look at employees who exited the firm in t_{-1}, the year before the change. Summary statistics for this group of 1,153 employees are given in the third column of Table A1. This group of employees and their dependents does differ somewhat on demographic variables. Moreover, on average, this group spends approximately \$700 more in t_{-1} than individuals in our main sample. However, this difference seems to be driven by the upper tail of a small number of individuals, as the medians of the two spending distributions are nearly identical, and the 75th percentiles are different by a minor amount.

Given these similarities, we feel comfortable using our main sample restrictions throughout the paper.

Intertemporal Substitution Analysis

In our analysis, we measure the extent to which employees increase spending in t_{-1} above expectations by substituting care that would otherwise have been obtained in the future if not for the policy change. To measure this ‘excess mass’, we first try to predict from prior years what spending would have been during t_{-1}, then measure the disparity. We run a regression as described in the main text in Section 3, for which the results are given in Table A2. We then calculate the excess mass as the difference between the true mean monthly individual spending amount and the predicted level. This measurement of excess mass is given in Table A3.

We note that, starting in December, excess mass is positive and high for December, November, and October (the three months with the largest excess mass among months in t_{-1}), before it drops down to nearly zero in September. There are some other outlier months across t_{-1} (March and August both have unusually high spending levels), however, as shown in Figure A2, the number of claims in those months is fairly reasonable relative to the trend. Careful investigation of those months (which cannot be shown due to individual privacy issues) uncovers that spikes in mean spending in those two months are primarily driven by a very small handful of unusually high-cost consumers. We take these combined trends as evidence that the majority of intertemporal substitution behavior is coming from care substituted into the last three months of t_{-1}.

One issue is that deviations from trend can occur both because of intertemporal substitution, as well as because of some nonzero draw of the unobservable idiosyncratic error term, $\bar{\epsilon}_t$. To account for our uncertainty over this term, we construct a confidence interval around our excess mass computation. We note that the mean squared error (MSE) of a regression is a consistent estimator of the variance of $\bar{\epsilon}$ in our model. Assuming that errors are not serially correlated, the standard deviation of the sum of the error terms for October, November, and December is $\sqrt{3 \cdot MSE}$, which in our case is approximately equal to 26.16. We multiply this term by 1.96 to get the 95% confidence interval for excess mass used in Table 4.

Treatment Effect Standard Errors

We compute the standard errors for all estimates presented in Section 3 with a block bootstrap method. We take a sample (with replacement) of N individuals from our primary sample, where N is the number of individuals in the sample, including their spending levels for both t_{-1} and t_0. We then compute the percent change in spending between t_{-1} and t_0 for this new sample. Importantly, to compute standard errors for our anticipatory-spending-adjusted estimates, we generate an 'excess mass' estimate for this new sample (in the method described in Appendix 5), and then use it to adjust our estimates accordingly. We repeat this procedure 1000 times, retaining 1000 sampled estimates of the treatment effect. We take the square root of the variance of these 1000 estimates, and use that as our estimate of the standard error of our treatment effect estimates.

Elasticity Estimates

A typical metric used to compare price sensitivity estimates in medical spending is the arc elasticity of total medical spending with respect to the price consumers face. As discussed in Aron-Dine et al. (2013), describing a non-linear insurance contract by one price is an oversimplification, since consumers face many potential true marginal prices throughout the contract and also face different marginal prices based on their respective health risks. The notion that it is difficult for one price to represent an insurance contract for a population is supported in our Section 1.5 analysis, which shows that consumers face very different prices throughout the year and that they respond to spot prices instead of true expected marginal prices.

Nevertheless, for comparison purposes, in Tables 3.3 and 3.4 we present the semi-arc elasticity of total medical spending with respect to price:

$$\frac{(q_{t_0} - q_{t_{-1}})/(q_{t_0} + q_{t_{-1}})}{(p_{t_0} - p_{t_{-1}})/2}$$

Here, q_t is mean individual total medical spending in year t, and p_t is the single 'price' of insurance coverage for the population in year t. We follow the literature here, and take the single price of the HDHP in t_0 to be the proportion of medical spending that consumers in the overall population would have paid for if t_{-1} medical spending occurred under the HDHP plan design. This is .219 in the primary sample in our setting. The price of the PPO in t_{-1} is 0 since consumers do not pay anything for health care on the margin in the PPO. 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, arc elasticity just represents the % quantity change so is not a satisfactory descriptive statistic.[1] The semi-arc elasticity represents the

[1]The arc elasticity in our context would be $\frac{(q_2-q_1)/(q_2+q_1)}{(p_2-p_1)/(p_2+p_1)}$. If p_1 is 0, then the denominator of this fraction always equals 1, and so the arc elasticity really only gives the arc change in quantity, regardless of the magnitude of the price change.

change in quantity, normalized by the baseline quantity, divided by the change in price.[2]

As Tables 3.3 and 3.4 reveal, the semi-arc elasticity for our primary causal treatment effect estimate lies in the range [-0.59, -0.69], averaging over both post-period years, while those from the other approaches in these tables lie between -0.57 and -1.32. We compare these estimates to two of the main estimates cited in the RAND Health Insurance Experiment, which compare two pairs of consumer groups: (i) those with 100% or 84% actuarial value plans or (ii) those with 84% or 69% actuarial value plans.[3] We use statistics from Keeler and Rolph (1988) to compute semi-arc elasticities of -2.11 and -2.26 respectively for these two estimates. Our semi-arc elasticity estimates range between one-quarter and one-half of those for RAND. Though, by this metric, consumers are less price sensitive in our setting, we note that the economic magnitudes of our treatment effect estimates are still substantial (regardless of the elasticity measures / comparison) and that there are many potentially important differences between our setting and the RAND setting.

Early Switcher Difference-In-Differences

Our primary sample includes individuals who were in the PPO prior to the required switch, and thus those that were actively required to join the HDHP in t_0. As discussed in Section 1.2, approximately 85% of consumers at the firm fall into this category and were required to switch into the HDHP. In this section, we use consumers who voluntarily switched to the HDHP earlier, in either t_{-2} or t_{-1}, as a control group for the treatment effect analysis just described. By incorporating an additional control group, we estimate a differences-in-differences specification where we compare the change in spending from t_{-1} to t_0 in our primary sample, where consumers were required to switch plans, to the control group where consumers were enrolled in the HDHP in both years. We focus on the t_{-1} - t_0 two-year period for this analysis to remove confounds that could manifest over longer time horizons: as shown in the earlier analysis, t_{-2} statistics are similar to t_{-1}, and t_0 similar to t_1.

Figure A3 plots the mean individual monthly spending from t_{-4} - t_1 for (i) our primary sample (ii) individuals who switched to the HDHP at the beginning of t_{-2} (6,255 individuals) and (iii) individuals who switched to the HDHP at the beginning of t_{-1} (5,528 individuals). We note that the early switcher samples are balanced, in the sense that employees are present from t_{-4} - t_1, and that prior to joining the HDHP these employees and their dependents were enrolled in the PPO.

[2] In general, as with the arc-elasticity measure, one might want to normalize the price change as well to reflect differences in scale (e.g. comparing changes of $5 to $10 versus $5000 to $10000). In our setting, this is not an issue because we define price as the share of firm-wide costs that fall on the employee, following past work on moral hazard (see e.g. Manning, Newhouse, Duan, Keeler and Leibowitz (1987)). Since this percentage is a relative measure already, this scaling issue does not arise when using the semi-arc elasticity measure.

[3] The 84% actuarial value contract has a 25% coinsurance rate up to an out-of-pocket maximum of $1000 while the 69% actuarial value plan has a 95% coinsurance up to a $1000 out-of-pocket maximum.

The figure clearly illustrates that early switchers are, on average, healthier than those in our primary sample who are required to switch for t_0. In addition, the figure shows a relative drop for mean spending for t_{-2} switchers in t_{-2}, for t_{-1} switchers in t_{-1}, and for t_0 required switchers in t_0. Figure A4 plots median spending over time for these different cohorts, and shows a similar pattern with much less noise since the median is a more robust statistic.

The fact that early switchers are healthier suggests that, in order to use them as a meaningful comparison group for the primary sample, we need to form a modified primary sample that matches the population of early switchers based on health status. For this analysis, we pool the two groups of early switchers (t_{-2} and t_{-1}) since we will be analyzing the spending change from t_{-1}-t_0. To measure health status in a predictive sense, we leverage the Johns Hopkins ACG software, which assigns each individual a predictive score, based on their past year of detailed claims data, for the upcoming health year. 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.[4]

We quantify the health status of early switchers with the observed distribution of individual-level ACG health status predictions for the year t_{-1}. We characterize this distribution with ventiles (20 equal sized buckets) of this predictive score, and weight the primary sample observations to match this distribution. Each ventile has, by definition, 5% of the early switcher sample. Thus, if 8% of the primary sample is contained in one of the early switcher ventiles, those individuals are weighted by $\frac{.05}{.08} = \frac{5}{8}$ in the weighted primary sample. We construct weights in this manner across the health status distribution to match the primary sample to the early switcher sample based on health status.

Figure A5 plots mean monthly individual-level spending for the pooled sample of early switchers and for our health-status weighted primary sample through t_0. The figure clearly illustrates that, prior to the switch in t_{-1}, when the two samples are in different plans, the HDHP consumers spend approximately 25% less than PPO consumers. In t_0, when both groups are in the HDHP, they spend almost identically (which also indicates successful matching on health status). Column 4 in Table 3.3 presents the quantitative difference-in-differences t_{-1}-t_0 spending reduction due to the HDHP switch implied by this figure:

$$[\bar{y}^{WPS}_{AS,t_0} - \bar{y}^{WPS}_{AS,t_{-1}}] - [\bar{y}^{ES}_{CPI,t_0} - \bar{y}^{ES}_{CPI,t_{-1}}]$$

Here, $\bar{y}^{S}_{M,T}$ refers to mean individual spending in year T under model M for sample S. Model AS refers to the model with both anticipatory spending and age/CPI adjustments. Model CPI refers to the model adjusting for age/CPI adjustments.[5] Sample WPS refers to the weighted primary sample, while sample ES refers to the early switcher sample.

[4] 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 on these predictive algorithms.

[5] We adjust for anticipatory spending in the weighted primary sample, which switches for t_0, and not for the early switcher sample, which remains in the HDHP over these two years. Even if there is some anticipatory spending for some HDHP consumers in December in a given year, it should be the same cross-sectionally (detrended) in t_{-1} and t_0.

Truven MarketScan Difference-in-Differences

Truven Data. In Section 3, we use data from Truven Analytic's MarketScan commercial claims database both as a control group, and to construct weights for an externally valid estimate. In this appendix section, we describe the data in more detail, and display an alternative version of the above exercises where we use Truven data with linked income as part of our matching procedure.

The Truven MarketScan database is a nationally representative individual-level commercial database that collects health insurance claims from a number of large insurers across the U.S., and includes data for both the insurance policyholder and their dependents. Much like our own firm's data, it includes claim-line-level data on the universe of medical visits and prescription drug usage for those individuals whose data it collects. It includes identifiers for the insurance carrier, the employer (where applicable), and the specific plan the individual is enrolled in, although no details are given about the cost-sharing characteristics of that plan. It also includes basic demographics, including as age and gender.

We restrict the Truven sample we use to individuals receiving private health insurance (i.e., not Medicaid or Medicare) during the years t_{-1} and t_0, as well as to only individuals who live in the state where the majority of our firm's workers reside. This leaves us with roughly 600,000 individuals in this sample in each year. For each individual, we compute their total incurred spending for each month by adding up the allowed expenditures claimed by them for both medical visits and prescription drugs. As in our primary sample, we deflate spending in the Truven sample according to the medical CPI given by the BLS to account for medical price inflation. This data is used in the analysis of Section 3.

We account for the vastly different demographics between the Truven data and the employees and their dependents in our dataset. Normally, one would account for this in a regression framework by using demographics as control variables. However, such a regression would produce the average treatment effect, averaged across both our primary sample and the Truven sample. Relative to this framework, our baseline treatment effects are measuring the average treatment effect on the treated (ATOT), since they average only for our firm's employees and dependents, so the estimates would be incomparable. Therefore, we instead follow, e.g., Bitler, Gelbach and Hoynes (2006), and, rather than use control variables in a regression, we reweight the Truven control group with propensity scores so that it, demographically, resembles our main firm's data. We take a nonparametric approach to constructing propensity scores, by dividing our main sample into cells based on age (in 5-year bins, with a cell for those aged 65 and older) and gender. We compute the proportion of individuals in our main sample in each cell, and then weight each observation in the Truven data by its equivalent cell weight in the main sample data. We then compute the treatment effect as:

$$ATOT = [\bar{y}_{t_0}^{PS} - \bar{y}_{t_{-1}}^{PS}] - [\bar{y}_{t_0}^{WTS} - \bar{y}_{t_{-1}}^{WTS}]$$

where

$$\bar{y}_t^{PS} = \frac{\sum_{i \in PS} y_{i,t}}{\sum_{i \in PS} 1}, \qquad \bar{y}_t^{WTS} = \frac{\sum_{i \in TS} w_{i,t} y_{i,t}}{\sum_{i \in TS} w_{i,t}}$$

and $w_{i,t}$ is the number of individuals in the primary sample in time t with the same age and gender cell as individual i. This procedure produces the estimate given in Column 6 of Table 3.4. Note that, for the primary sample, spending is adjusted for anticipatory spending (which is irrelevant for the Truven sample individuals), using the procedure given in Appendix 5.

We also seek to provide an externally-valid estimate, that acknowledges that our sample has a demographic composition quite unlike that of the rest of the United States. To do so, we reverse the procedure above, instead reweighting our primary sample according to the demographics of the Truven sample. We use the same age and gender cells, and so our estimate represents, in this context, the average treatment effect on the untreated, as such:

$$ATOU = [\bar{y}_{t_0}^{WPS} - \bar{y}_{t_{-1}}^{WPS}] - [\bar{y}_{t_0}^{TS} - \bar{y}_{t_{-1}}^{TS}]$$

where

$$\bar{y}_t^{WPS} = \frac{\sum_{i \in PS} \tilde{w}_{i,t} y_{i,t}}{\sum_{i \in PS} \tilde{w}_{i,t}}, \qquad \bar{y}_t^{TS} = \frac{\sum_{i \in TS} y_{i,t}}{\sum_{i \in TS} 1}$$

and $\tilde{w}_{i,t}$ is the number of individuals in the Truven sample in time t with the same age and gender cell as individual i.

Because Truven samples in a way that is nationally representative, we consider this treatment effect estimate to roughly approximate that of the U.S. under-65 privately insured population. This result is given in Column 5 of Table 3.4.

Income-Linked Analysis. As noted, the firm we analyze employs workers at wages that well exceed national averages. Therefore, finding a matching control group based on income is quite challenging. For about half of the policyholders in the Truven data, we are able to link their claims to a secondary dataset from Experian that includes the policyholder's annual income. For this analysis (seen below), we further restrict the sample to only those making at least $75,000 per year. This is because the lowest income category in the dataset for the firm we study is "less than $75,000," and so we are unable to accurately match the incomes of anyone making below that amount in our firm.[6] These restrictions leave us with around 30,000 individuals in each year for our secondary Truven sample.

Figure A6 replicates the graph from the right side of Figure 4.2 with this secondary sample. Much like in that graph, the income-linked Truven sample has spending levels below that of our sample, although the gap is far smaller. This is likely due to the fact that we restrict this secondary sample to only high-income individuals, who have higher purchasing power and are likely also employed by firms with similarly-generous health benefits. However,

[6]Moreover, only 6% of employees at our firm are in this bracket, while the vast majority of employees within the Truven data are, so a very small portion of our data would have an excessively high weight.

the PPO plan offered by the firm we study was extraordinarily generous even among high-income employers, explaining why the gap still remains. To quantify the the relative spending reduction, we follow the same procedure given above to compute an estimate of the ATOT for our primary sample firm. However, in this analysis, we further subdivide the demographic cells by income, using eight brackets of $25,000, beginning with the bracket $75,000-$100,000 and ending with a bracket including all of those who make over $250,000. Column A2 in Table A4 gives our ATOT estimate, which we bound between -18% and -23.7%, which is slightly higher than our primary estimates but lower than the estimates from our other Truven-derived control group.

We also replicate the procedure generating an externally valid estimate on this new control group, given in Column A1 of Table A4. We get an estimate bounded between -2.1% and -6.7%, unusually lower than our other primary estimates. This likely comes from the fact that the most common income bracket in the Truven sample is the lowest one, whereas the same bracket comprises of only a miniscule amount of our primary sample, so a small number of individuals are given high weights. In our primary sample, as given in Table A5, the lowest income bracket is also the least responsive, thereby depressing our estimate.

Additional Analysis of Treatment Effect Heterogeneity

In this section, we present a number of figures and graphs that provide more detail on heterogeneity in spending trends across a variety of categories. First, we expand on Figure 4.1 in the text, by presenting Figures A7 and A8. In these figures, we break down the highest quartile of ACG score into four subgroups, and show that we can observe spending responses to the policy change broadly even across the top end of the sickness distribution. Figure A8 in particular shows that even in the 99th percentile of expected health risk, the median individual-level spending is reduced in the years following the change, despite the fact that individuals in this risk bracket should have no incentive to do so. Figure A9 first defines medical claims into categories based on the service location where medical care was received, and then plots spending in each of these categories over the entire timespan of our data. We see sharp reductions in office and emergency room visits, outpatient hospital care, and preventive care, with no real change in mental health spending or inpatient hospital care. Figure A10 breaks down spending cutbacks for prescription drugs, showing that cuts come from both branded and generic drugs.

The treatment effects estimated for these spending breakdowns, as well as others, is given in Table A5. We use the methods developed in Section 3 to estimate The table presents estimates comparing t_{-1} spending to t_0 spending for parsimony: t_{-1} to t_1 comparisons are similar and included in Table A8. We present three sets of treatment estimates: raw single-difference estimates (in Column 1), the same estimates adjusting for aging and health care CPI growth (Column 2), and our preferred estimates, which adjust for those factors as well as anticipatory spending (Column 3). The sickest quartile of individuals, who spend on average $12,335 in t_{-1}, reduce spending by between 18-22% under our preferred treatment effect measures. These treatment effects are slightly larger for the ex ante health status quartiles 1 (healthiest), 2, and 3 respectively, though off much lower pre-treatment spending bases.[7] [8] The table also presents these results for consumers categorized by number of documented chronic conditions entering a given calendar year, revealing surprisingly limited heterogeneity on this dimension.

Table A5 also documents heterogeneous treatment effects by (i) consumer demographics and (ii) type of location the medical service was performed at. One notable result is that spending reductions for dependents are limited (12%) and there are no anticipatory spending shifts for this group, suggesting that parents may be less willing to economize on care or shift care for their children. Spending reductions do not seem to vary much by age and, most surprisingly, income. In particular, the lowest spending bracket has the lowest respond, despite potentially having the large incentives to cut back conditional on health risk.

From our service location analysis, one notable result is that spending is reduced across all eight of these broad spending categories, and that the effects have a fairly narrow range of a 6% CPI adjusted reduction (mental health) to a 25% reduction (ER spending). This is somewhat surprising, since some categories seem more elective (e.g. physician office visits, 18% reduction) and others seem less elective (e.g. inpatient, 13% reduction). Notably, consumers reduce spending for both branded drugs (20%) and generic drugs (19%). In addition, spending on services that are classified as preventive is reduced by 10%. This is especially striking since (i) these services are all free to consumers under the HDHP (as mandated under the ACA) and (ii) these are services that may prevent higher spending and poor health in the future.

Table A6 displays our 'excess mass' calculations used to compute Column 3, constructed as described in Appendix 5. The first column shows the final excess mass calculation used in Table A5, while the second column gives the standard error for that calculation. The last three columns break down the excess mass for each month used in the data. We can see that most of the excess mass is driven by above-trend spending in December t_{-1}, as nearly every category of spending results in a positive excess mass calculation for that month.

Table A7 presents standard errors for the treatment effect estimates in Table A5. We construct standard error estimates by block bootstrap the same way we do for our primary

[7] The health status quartile treatment effect analysis fixes the quartiles based on predictive indices for t_{-1}, but allows consumers to switch between those quartiles from one year to the next. This means that the cross-sectional health status quartile populations change over time, but the definition of a quartile in terms of health status remains the same. This is why the % of consumers in each quartile is slightly different than 25%.

[8] We note that the average of these health status quartile treatment effects, weighted by total spending, is slightly larger than the treatment effect presented for the entire population in Table 3.3. In the raw spending and age/CPI-adjusted only treatment effects, this difference is because the quartiles have slightly different mixtures of health status *within the health status range for the quartile* over the years. For the anticipatory spending adjusted estimates, this difference could also come from the fact that anticipatory spending regressions /adjustments are done separately for each quartile. In Table A9, also in this section, we present some additional versions of this analysis, intended for robustness, where health status quartiles are defined as true quartiles on a year to year basis, though the ACG index boundaries of each quartile may change .

estimates, as described in Section 3. In our block bootstrap procedure for demographic subsamples (for example, the effect for spouses), we first take the relevant demographic subsample, and *then* perform our bootstrap procedure, so that the number of individuals in that subsample is held fixed.

Finally, Table A9 presents an alternate version of our ACG quartile analysis from Table A5. In the initial analysis, we allow ACG scores for a given individual to vary over time in order to measure the treatment effect. In this table, we instead fix an individual's ACG score at one point (using their score constructed using either t_{-2} or t_{-1} claims data), and calculate their treatment effect over time. This method can suffer from mean reversion, where consumers with high scores previously due to chance may look as though they decrease spending later, which is why we do not use it for our main analysis. Presented here, we can see some evidence of this mean reversion, although it is not very strong relative to our treatment effects.

Additional Analysis of Price Shopping

We do a number of robustness checks on our analysis of consumer price shopping. The first is that we verify that the rankings of prices across providers within a class of procedures is constant over time. To do so, for each procedure-year pair, we assign each provider in our restricted provider-procedure-year set a ranking according to their price for that procedure-year. We then calculate Spearman's rank correlation coefficient for each consecutive pair of years. The result from this exercise is given in Table A10. For nearly all pairs, the coefficient is very strong, over 0.92 for all year pairs. We view this as evidence supporting our modeling assumption that the rankings are approximately constant.

We additionally perform a version of our price shopping analysis on new employees. The key reason for doing so is because a lack of price shopping in the short run that we observe in our data may be driven by pre-existing relationships between consumers and providers. These relationships may make it difficult to switch to a new provider, even if the previous provider is more expensive. We do this by taking the claims of new employees in t_{-1} and t_0. We use claims from these employees only for the year in which they were a new employee, and we compare these two cross-sections in the same way we compared pairs of years in our main analysis. The results are given in Table A11. Again, we see no evidence for price shopping, instead finding slight increases in prices achieved. The primary driver of differences in spending for new employees, as in our main sample, is quantity reductions.

In Table A12, we show which of our decomposition pieces have positive value for the 30 top procedures by dollars spent. In Table A13, we disaggregate this, displaying the values for each of the 30 procedures. Due to space concerns, we present the decomposition only between t_{-1} and t_0. It is clear to see that very few procedures seem to exhibit meaningful consumer price shopping.

In Tables 3.7 and 3.8, we presented decompositions for high and low value care, for the year pairs t_{-1}-t_0 and t_{-3}-t_{-2}. Tables A15, A16, A17, and A18 present these same decompositions for all year pairs. Table A19 replicates the analysis performed in Table A5 for these precise definitions of high and low value care.

Additional Analysis of Reduction in Preventive Care

In Section 1.4 we investigated the nature of spending reductions for several different kinds of health care services. Preventive care services are of specific interest, because they are generally considered to be low cost, high value services that policymakers would like to encourage consumers to use: there is a range of past research, summarized nicely in Baicker et al. (2015) that shows ample evidence for underconsumption of preventive care by consumers. Under the Affordable Care Act, a baseline set of preventive services are required to be offered free of charge by insurers, with the intention of encouraging consumers to take up such services (see, e.g., Kaiser Family Foundation (2011)). Moreover, in this vein, there are many current supply-side policies (such as pay-for-performance bonus programs) that are implemented to incentivize medical providers to more effectively deliver preventive services.

Our analysis find that consumers reduce the quantity of general preventive services consumed by 7.5% in year t_0, the first year post-switch, and by a further 5.2% in the second year post-switch. We find that in the first year post-switch, the quantity of services that are considered to be preventive with a prior diagnosis is reduced by 12.2%. Given that these are considered to be cost-effective services, and that *these services are free* both before and after the switch to high-deductible health care, it is interesting to better understand exactly how consumers are reducing consumption of these services.

There are several potential reasons for why consumers reduce preventive services in our study:

1. Consumers (and potentially their providers) may not understand that preventive services are free, and instead think that their cost has gone up along with all other services. In this case, consumers would reduce preventive service consumption in response to a perceived increase in price.

2. Consumers may first choose whether to visit their provider to consume a bundle of services, some of which are free and some not, and then decide whether to consume preventive care. In this case, the extra cost of the bundled services may discourage consumers from going to their providers, and lead to lower consumption of preventative care. Relatedly, some consumers may not think about preventive care, and only consume such services during office visits when encouraged to by providers. In that scenario, a reduction in office visits would also lead to a reduction in preventive care consumption.

We investigate these potential explanations by studying whether reductions in preventive care occur on the extensive margin (fewer visits to primary care providers) or the intensive margin (fewer preventive services consumed conditional on a provider 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 feel that this decomposition provides useful evidence for distinguishing between these hypotheses, even though there are some subtleties in mapping the hypotheses above to this extensive-intensive margin decomposition,.

To conduct this analysis, we used four different methods for classifying a primary care provider office visit. In this appendix, we focus on what we view as the best two methods, and briefly describe the other two at the end of this section. Crucially, all methods yield similar results.

The first method looks in each month and determines whether an individual had a CPT code or ICD-9 code that specifically signified a primary care office visit.[9] For this method, we split the year up into months and for each individual develop an indicator of whether they had a primary care visit in a given month.

The second method defines primary care providers by specialty listed in the data (this field is populated for all providers). We consider providers listed under either (i) family medicine (ii) preventive medicine, general or (iii) internal medicine to be primary care providers. Then, any time a patient sees such a provider in a month, we classify the patient as having made a visit to a primary care provider in that month.

For each of these two methods, we study intensive margin preventive care use in two distinct ways. Our primary methods looks at dollars spent on preventive care per office visit. Our second method looks at a binary indicator variable of whether any preventive services were consumed during the office visit.

For each approach, we investigate both for our primary sample and for the price shopping decomposition sample (restricted to the main company location). Since the results are similar, here we only present the results for the price shopping sample, to be consistent with the tables in Section 1.4. Additionally, we run the analysis separately for services that are generally considered preventive, and also for services that are only considered preventive with a prior diagnosis.

Table A20 presents the results for our approach that measures the intensive margin of preventive care based on $ per provider visit spent on that care. When provider visits are measured according to ICD-9 and CPT codes, the number of provider visits (extensive margin) decreases by 12.1% from t_{-1} to t_0 (143,887 to 126,406). For general preventive care, spending per visit increases from $62.57 to $64.79 (3.5%) per visit. For preventive care with a prior diagnosis, spending per visit decreases from $117.18 to $114.59 (2.1%). Thus, for this approach, the reduction in preventive spending can be attributed almost entirely to the extensive margin, i.e. fewer office visits, rather than doing fewer preventive services per office visits.

[9]There are many such codes, and the list we used is available upon request.

When office visits are measured according to provider specialty, the overall results are quite similar. The number of office visits declines from 63,121 to 54,218 (14.1%) from t_{-1} to t_0. On the intensive margin, spending per visit for general preventive care increases by 3.3% while for preventive care with a prior diagnosis it decreases by 1.8%. A similar emphasis on the extensive margin holds for different classification methods we use for provider visits, suggesting that this result is robust to different definitions of the extensive margin.

Table A21 presents this decomposition when we measure preventive care use with a binary indicator of whether any preventive care was done in a given month (rather than $ per visit). The results are similar: almost all reductions come at the extensive margin. The extensive margin statistics are the same for this case as with the prior case, the only difference is with the measurement of the intensive margin. The change in the intensive margin spending for the ICD-9 / CPT office visit classification is 3.0% for general preventive care and -0.7% for preventive care with a prior diagnosis. Similar results hold for the second classification (provider specialty) showing that the extensive margin impact is similar regardless of the method chosen for measuring preventive care use. Figure A11 shows the entire density of intensive margin spending for general preventive care for the two definitions of an office visit. The density of $ per visit spent is very close together in each case, suggesting that almost all of the action is on the extensive margin in terms of reduced preventive quantities and spending.

Table A21 also investigates this decomposition for specific preventive services, including mammographies, colonoscopies, and urinalysis. These cases highlight the difference in extensive vs. intensive margin effects across services. For the physician specialty classification of office visits, mammographies per visit declines by 6.7%, urinalysis per visit declines by 0.4%, and colonoscopies per visit declines by 35.0%. Thus, the urinalysis effect is almost entirely related to the extensive margin, there is some intensive margin mammography effect, though not a large one, and the intensive margin effect for colonoscopies is substantial, indicating that that is one margin consumes / providers are clearly responding to in the treatment year. We caution that results for specific services like colonoscopies should be viewed in light of any changing guidelines that occur over time for how physicians should prescribe them, though there is no national downward trend in those services.

Taken together, these results suggest that consumers are reducing preventive care consumed primarily at the extensive margin: when they actually visit a provider they're doing almost the exact same level and type of preventive care as previously, but, they are doing less preventive care in general because they are reducing their doctor visits in general. This suggests that consumers are reducing preventive services consumed in large part because those services are bundled with other, costly, services for provider visits, and that patients are reducing the number of times they consume such bundles. The results work against the hypothesis that these spending reductions are coming primarily because consumers perceive higher prices for preventive services as well, and are purposefully reducing the consumption of such services when they go to the provider. Note that our results do not perfectly test these hypotheses, since consumers could be reducing visits because they perceive the prices of preventive care to be higher. However, the use of preventive care conditional on an office visit suggests that there is no substitution at all once a visit is scheduled, making it unlikely that consumers are reducing visits primarily because of perceived preventive care prices.

In analysis not reported here we perform a range of robustness checks. First, we investigate two additional methods for classifying office visits (i) medical events where service location is denoted Office and (ii) identifying providers as primary care providers based on claims that patients who visit them have (then applying this designation of primary care provider to all claims affiliated with that provider. We also test for preventive spending that doesn't fall under our different office visit definitions, and note that most spending classified as preventive falls under our office visit definitions. We also repeat the analysis for the primary sample in the main text (not restricted to the main company region). All of these contingencies point in the same direction: consumers are reducing office visits overall, but not reducing the amount of preventive care done per office visit.

Additional Analysis of Responses to Non-Linear Contract

We present versions of our descriptive analysis of employee responses to the non-linear structure of the HDHP, where we instead use single employees, or employees with only a single dependent, in Figures A13 and A14. These figures replicate the analysis shown in Figures 4.3, 4.4, and A12 in the text for those populations. Incremental spending for the next month and for the rest of the following year is given for employee-month combinations in a given tier of the HDHP in t_0. These figures provide results that are qualitatively similar in nature to those for employees with two or more dependents.

LASSO Results

To demonstrate further that variation in end of year price does not explain spending differences, we turn to a method originally employed by Backus (2014). We restructure our prior regression model (with all three prices) as a penalized linear model, specifically a LASSO model,[10] and estimate the model for different values for the coefficient constraint. As the LASSO coefficient size constraint binds more tightly, the solution algorithm will be forced to set some coefficients to zero. We use a stepwise regression model to focus on the set of constraint values that make the algorithm remove a variable from the model. It will begin with those variables that least explain variation in health spending. We think of this as a data-driven way to characterize the 'importance' of each of the price variables in explaining health spending choices. Furthermore, by estimating a penalized regression we can flexibly capture correlations between dependent variables, an advantage in our setting as different price measures are all based on a mapping from measures of health and spending over time.

Figure A18 presents the results of this exercise for the key price coefficient of interest: spot price, expected, end-of-the-year marginal price and last years end-of-the-year marginal price.

[10]LASSO is equivalent to OLS (a linear model minimizing squared residuals) with an additional constraint on the sum of the absolute values of the coefficients.

These results are based on t_0 and t_1 respectively. The coefficients at the far right represent the unconstrained OLS regression; the far left represents the completely constrained LASSO model (where all coefficients are set to zero), with points in between representing constraint levels between these two extremes.

As the constraint binds (moving from right to left), the coefficients on the expected end-of-year marginal price variables are the first set to zero, implying that they are relatively unimportant for explaining the variation. In t_0 and t_1 we see the most important factor, both in terms of effect size and the fact that it remains different from zero as the penalty function gets vary large (steps go to 0), is spot price of 1. In t_0 we see some impact of the 4th quartile of the E[EOY Marginal Price] though the magnitude is far smaller. A similar result occurs for last years marginal price of .1 in the t_0 plot. For t_1 the results are quite similar for spot price of 1: it is the most significant in terms of longevity as well as in magnitude. Together these results lend further evidence, using an alternate empirical approach that flexibly allows the price response to fit the data, that primary driver of the behavioral response is for those under the deductible.

Appendix A Tables and Figures

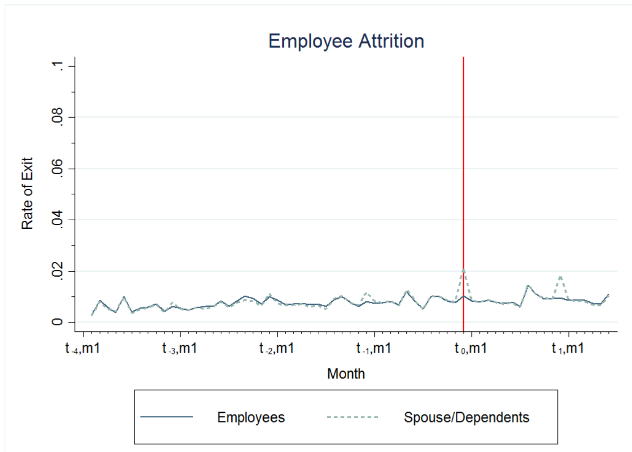


Figure A1: This figure plots the monthly hazard rate of exit from the firm's insurance coverage over time for employees and their dependents.

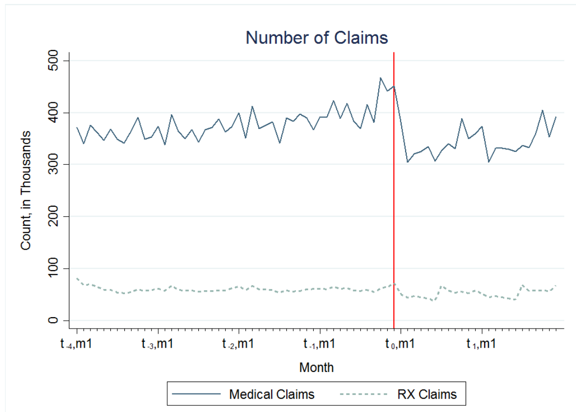


Figure A2: This figure plots total number of monthly claims, both for medical claims and prescription drug claims over time, for our primary sample.

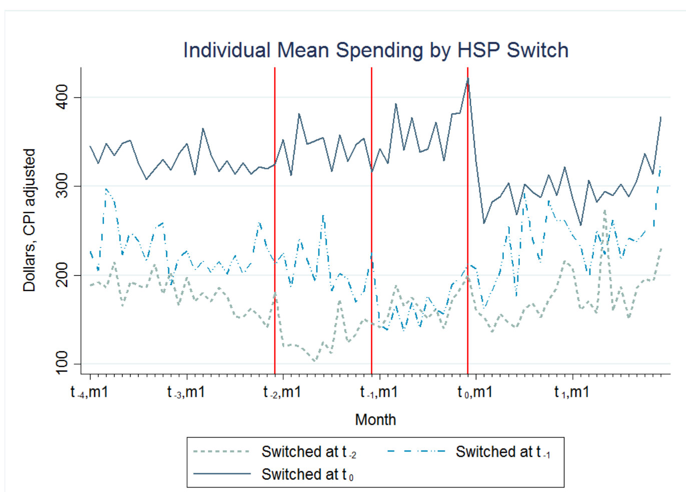


Figure A3: This figure plots mean monthly spending over time for consumers who (i) are in our primary sample (and thus were required to switch to the high-deductible plan in t_0) (ii) those who elected to switch early to the HDHP in t_{-1} and (iii) those who elected to switch early to the HDHP in t_{-2} (and stayed in that plan over time).

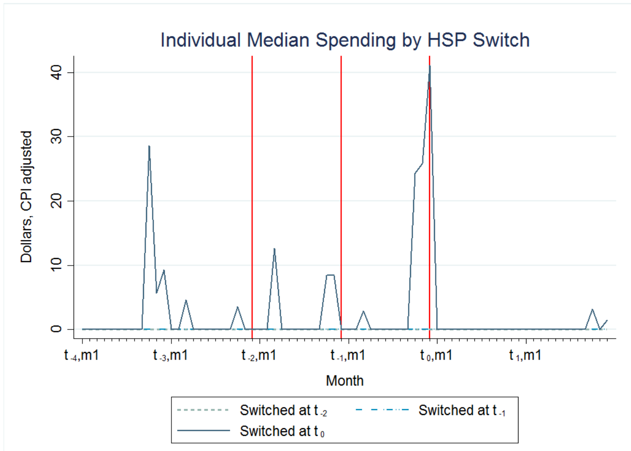


Figure A4: This figure plots median monthly spending over time for consumers who (i) are in our primary sample (and thus were required to switch to the high-deductible plan in t_0) (ii) those who elected to switch early to the HDHP in t_{-1} and (iii) those who elected to switch early to the HDHP in t_{-2} (and stayed in that plan over time).

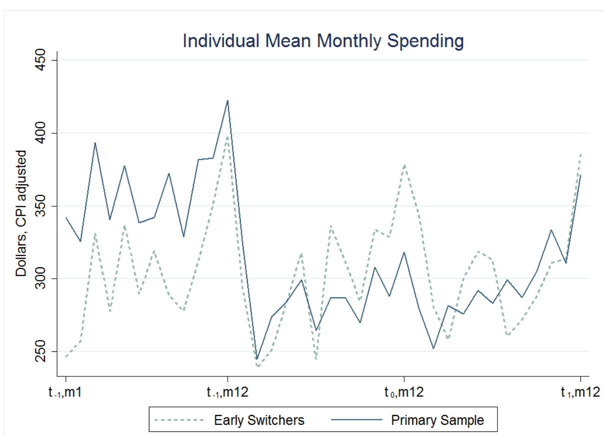


Figure A5: This figure plots mean monthly individual spending over time for consumers who (i) are in our pooled sample of early switchers and (ii) are in our weighted primarily sample through t_0, matched to the early switcher sample based on the health status distribution.

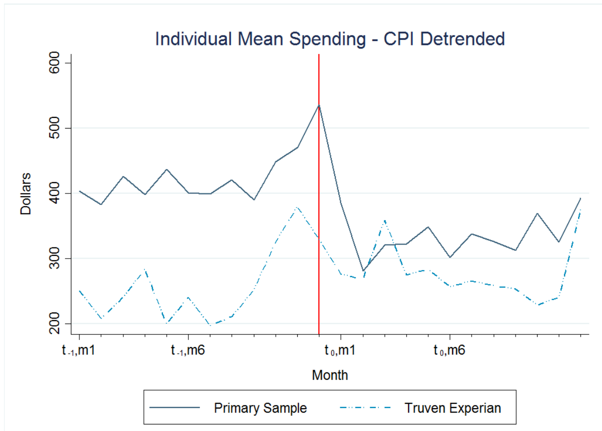


Figure A6: This graph presents spending for our primary sample alongside spending for the weighted control group formed from Truven MarketScan data with linked income fields.

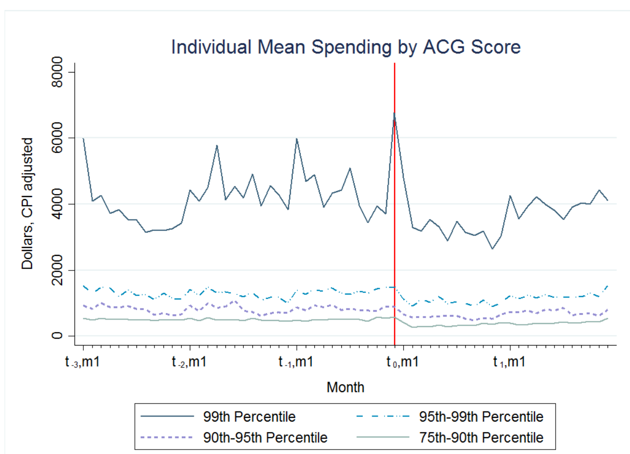


Figure A7: This figure plots adjusted mean spending for individuals in a given month, by ACG predictive health index bin (the index is calculated at the beginning of each calendar year). This graph divides individuals in the top quartile of the ACG index into smaller subgroups.

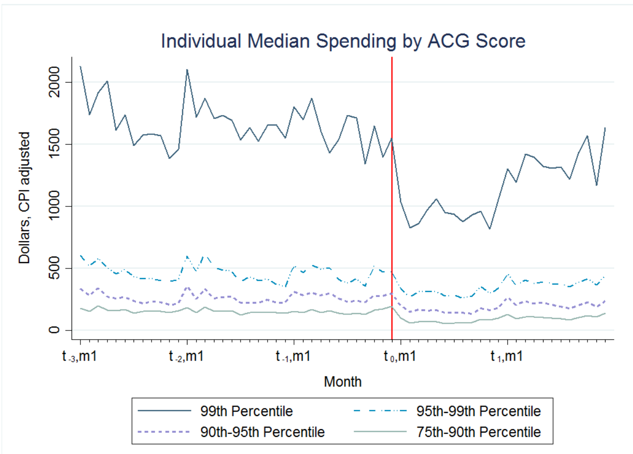


Figure A8: This figure plots adjusted median spending for individuals in a given month, by ACG predictive health index bin (the index is calculated at the beginning of each calendar year). This graph divides individuals in the top quartile of the ACG index into smaller subgroups.

Figure A9: This figure plots mean medical spending for individuals in a given month, by the type of care, both adjusted and unadjusted for age and price trends. These categories are mutually exclusive, except for Preventive.

Figure A10: This figure plots mean prescription drug spending for individuals in a given month, for brand and generic drugs, both adjusted and unadjusted for age and price trends.

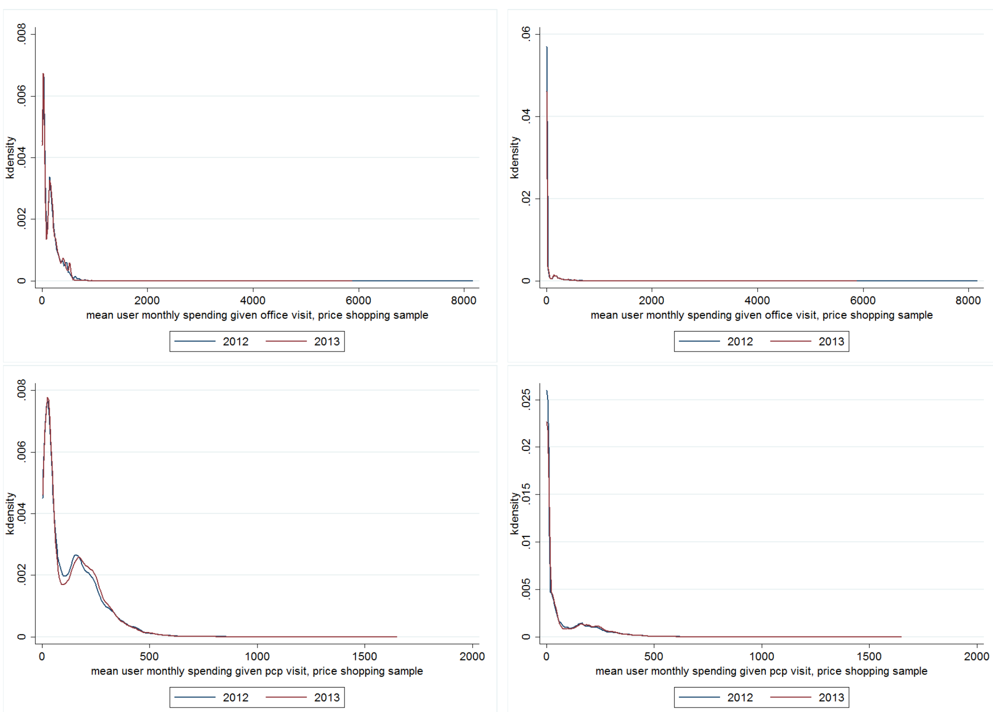


Figure A11: This figure presents per visit spending densities for general preventive care services, before and after the required switch to high deductible care. The top section of the figure presents these densities for our first definition of an office visit (defined by CPT codes) while the bottom presents these densities for our second definition, based on provider specialty. The left half of the figure presents densities conditional on spending great than 0 on preventive in a visit, the right half presents densities including zero spending on preventive care.

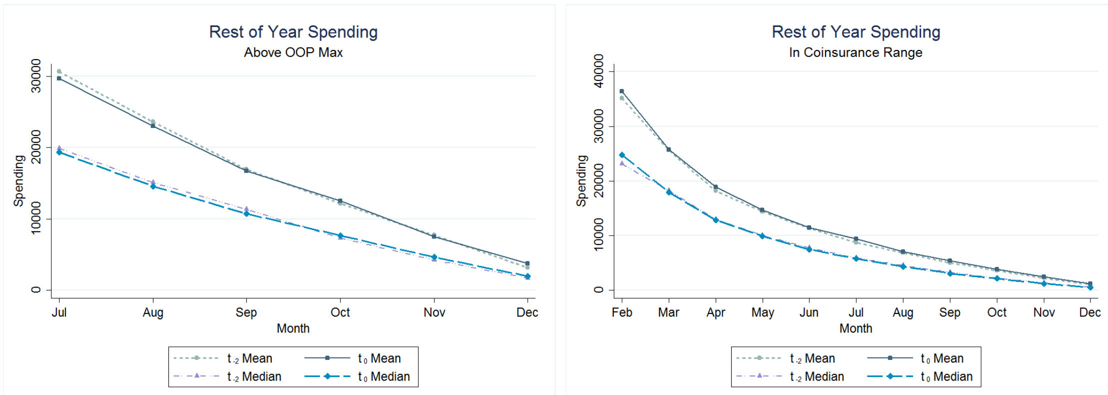


Figure A12: The left panel in this figure shows incremental spending for the rest of the year, for families who have passed the out-of-pocket maximum by the start of a given month in t_0, compared to t_{-2} incremental spending for equivalent quantiles of pre-period consumers. The right panel presents the analogous figure for families who start a given calendar year month in the coinsurance arm of the HDHP (and matched t_{-2} consumers).

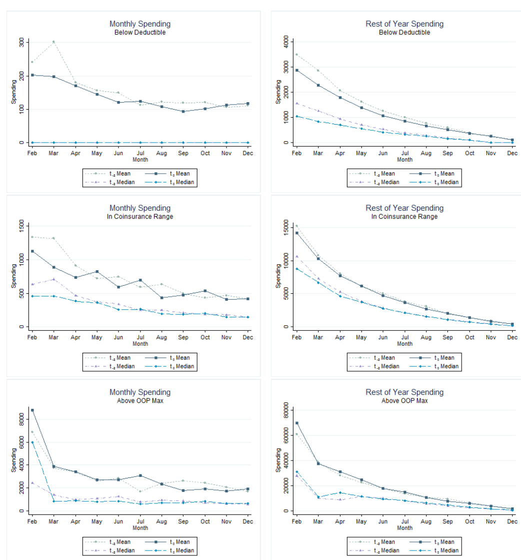


Figure A13: This figure shows incremental spending for employees who have passed the out-of-pocket maximum by the start of a given month in t_0, for single employees. The left side of the figure studies incremental spending for the next month, while the right side studies incremental spending for the rest of the year. This t_0 incremental spending is compared to t_{-2} incremental spending for the equivalent quantiles of consumers based on total yearly spending up to month m, M_m.

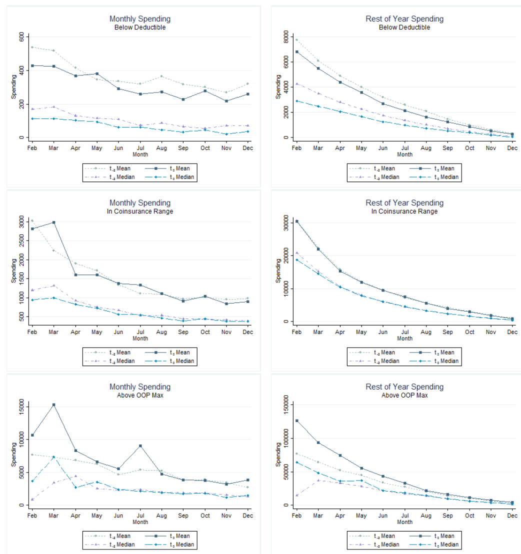


Figure A14: This figure shows incremental spending for employees who have passed the out-of-pocket maximum by the start of a given month in t_0, for employees with one dependent.

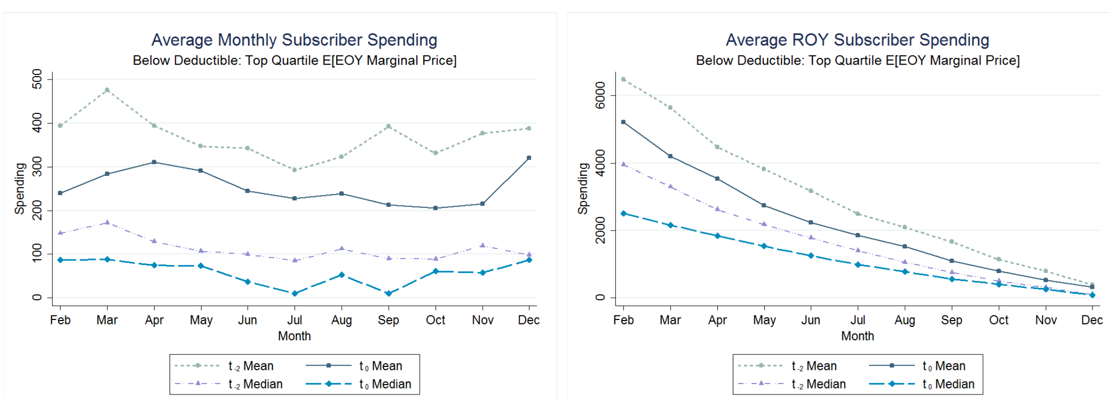


Figure A15: This figure shows incremental spending for employees who have passed the out-of-pocket maximum by the start of a given month in t_0, for families with the highest quartile of shadow price.

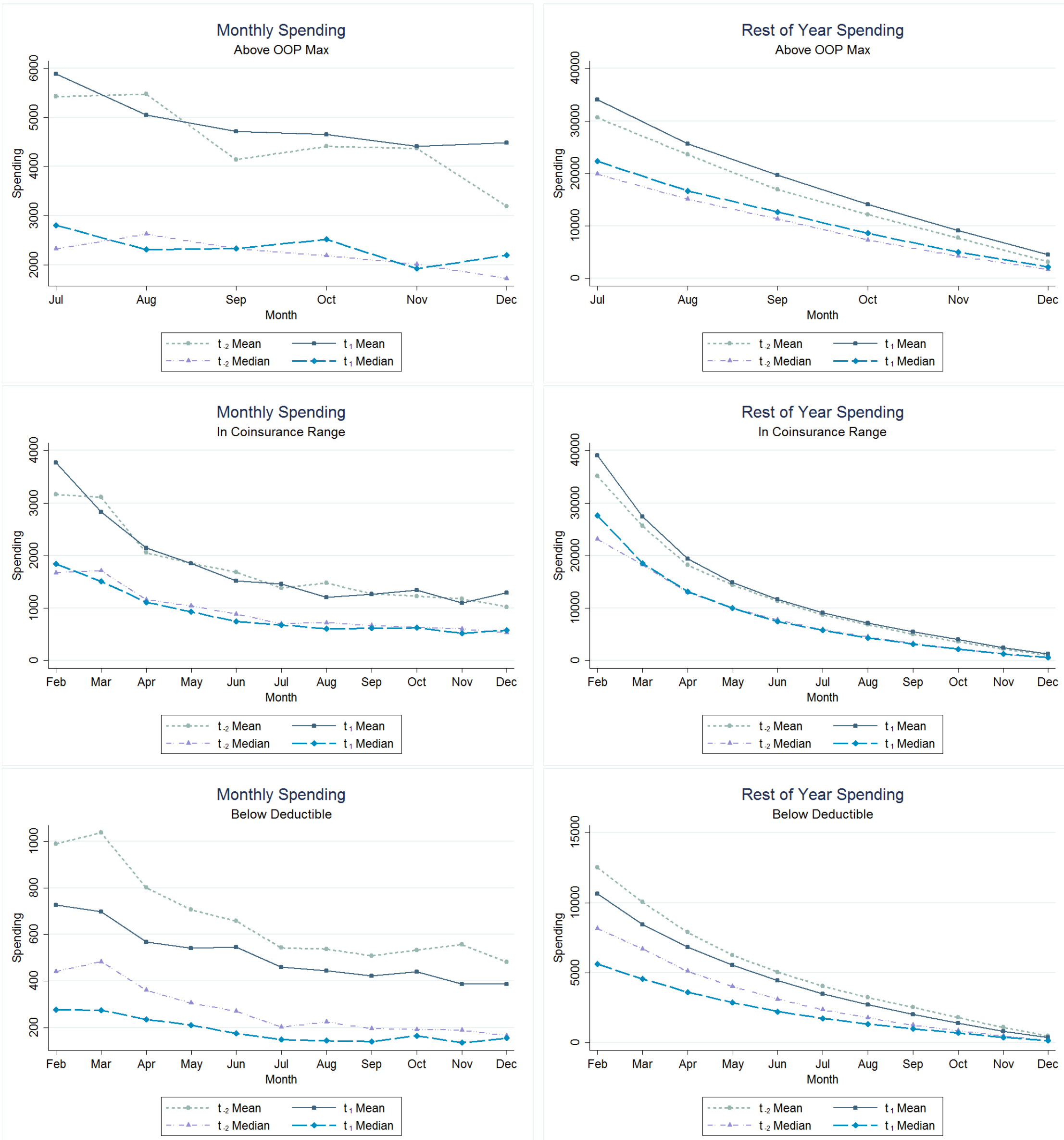


Figure A16: This figure presents descriptive results for t_1, comparing incremental spending in that year by plan arm to spending by equivalent quantiles of consumers in t_{-2}. These figures are directly analogous to those presented earlier in this section, describing how incremental spending in t_0 compares to that in t_{-2}. The left panels present incremental spending for the next month conditional on start of month plan arm, while the right panels present incremental spending for the rest of the year.

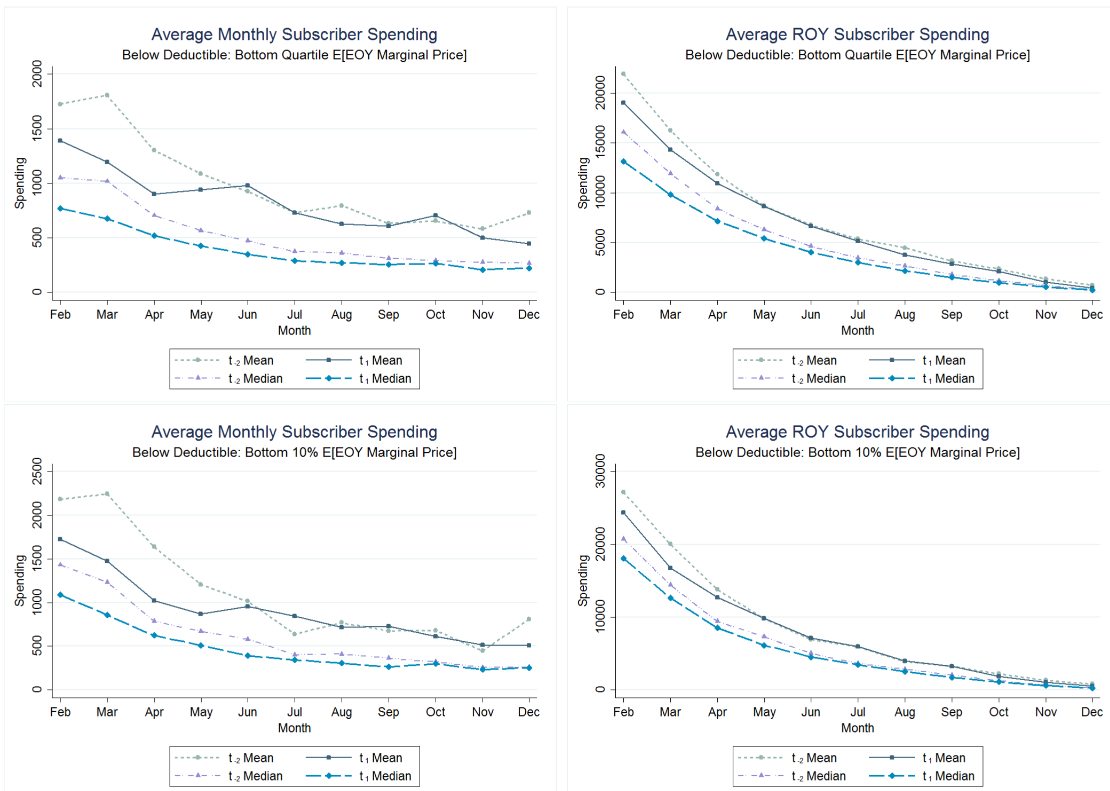


Figure A17: This figure presents descriptive results for t_1, and examines how predictably sick consumers under the deductible at the beginning of a month reduce incremental spending. These figures are directly analogous to those presented earlier in this section, describing how incremental spending in t_0 compares to that in t_{-2}. The left panels present incremental spending for the next month conditional on start of month plan arm, while the right panels present incremental spending for the rest of the year.

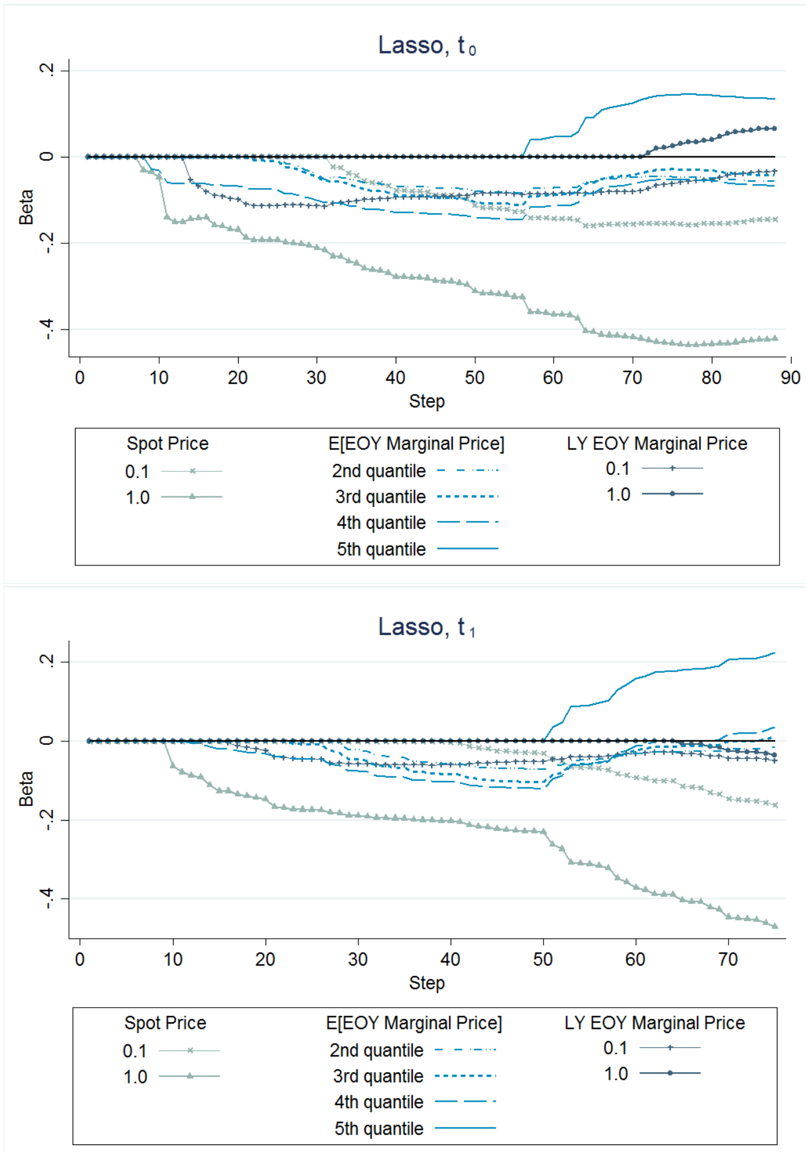


Figure A18: This figure presents our results from the LASSO procedure described in the text. Each step denotes the point where (moving from right to left) a variable is removed from the regression (i.e., its coefficient is set to zero).

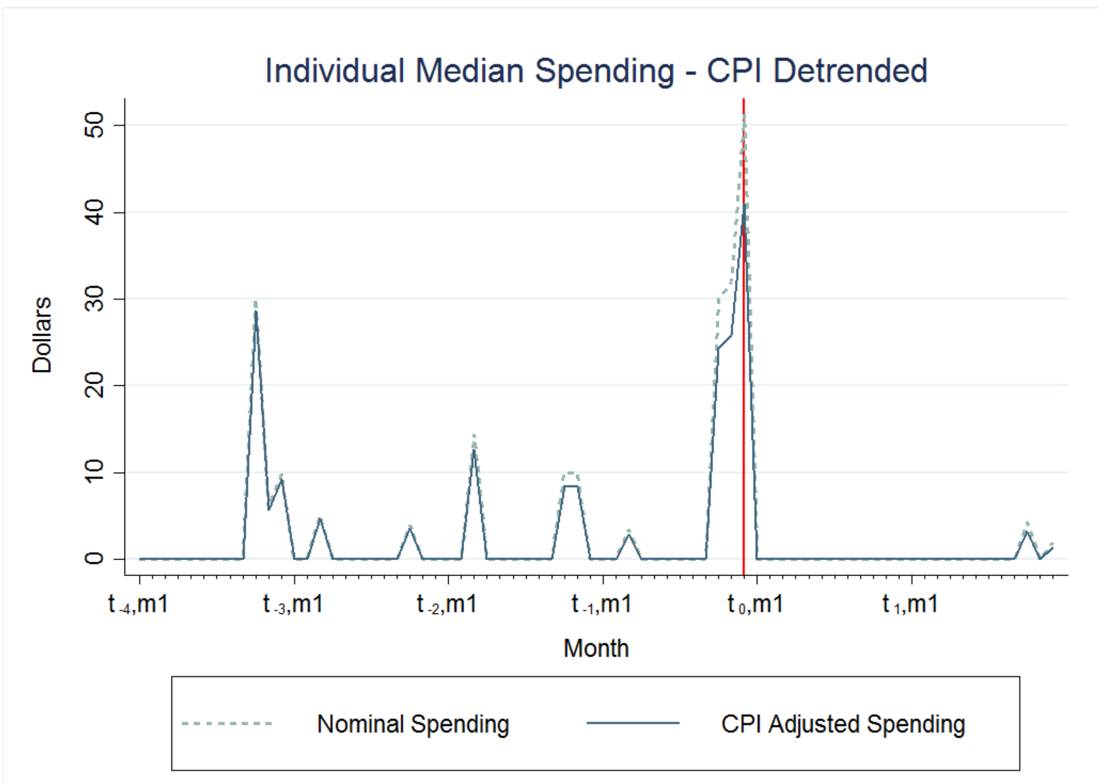


Figure A19: This figure plots median monthly spending for individuals in our primary sample from t_{-4}-t_1, both adjusted and unadjusted for age and price trends.

Table A1: This table presents summary demographic statistics for (i) our primary sample, which is restricted to employees present over the time horizon t_{-4}-t_1, and their dependents; and (ii) an alternate sample, which is only restricted to employees present over the time horizon t_{-2}-t_0. Statistics for the primary sample are presented for the year t_{-1}.

Sample Demographics	Primary Sample	Alternate Sample	Employees Exiting in t_{-1}
N - Employees	22,719	31,042	1,153
N - Emp. & Dep.	76,759	95,224	3,180
Enrollment in PPO in t_{-1}	100%	100%	100%
% Male- Emp. & Dep.	51.4%	48.8%	41.4%
Age, t_{-1} - Employees			
18-29	4.3%	7.0%	5.9%
30-54	91.4%	88.2%	77.0%
$\geq$ 55	4.3%	4.8%	6.4%
Age, t_{-1} - Emp. & Dep.			
$<$ 18	36.1%	33.2%	24.8%
18-29	8.8%	9.6%	10.9%
30-54	52.0%	48.9%	42.0%
$\geq$ 55	2.8%	2.9%	3.9%
Income, t_{-1}			
Tier 1 ($<$ $100K)	7.3%	7.6.8%	9.7%
Tier 2 ($100K-$150K)	64.7%	65.0%	59.0%
Tier 3 ($150K-$200K)	22.6%	20.1%	15.9%
Tier 4 ($>$ $200K)	4.7%	4.2%	2.6%
Family Size, t_{-1}			
1	16.1%	18.4%	15.2%
2	17.9%	18.7%	32.4%
3+	65.9%	62.9%	52.4%
Individual Spending, t_{-1}			
Mean	$5,223	$5,375	$5,921
25th Percentile	$631	$645	$533
Median	$1,795	$1,817	$1,796
75th Percentile	$4,827	$4,890	$5,151
99th Percentile	$52,360	$53,239	$59,481

Table A2: This table presents coefficients from the regression model used to measure excess mass.

Regression Results	
Variable	Coefficient
Months Since Jan. in Year t_{-4}	0.442
February	-32.37
March	15.28
April	-11.07
May	-11.90
June	-5.87
July	-32.34
August	-20.96
September	-31.93
October	-19.79
November	-22.54
December	-27.71

Table A3: This table presents the computed excess mass for each month in the second half of t_{-1}.

Excess Mass	
Month	Excess Mass
December	85.83
November	41.57
October	37.83
September	-2.15
August	20.91
July	12.21
January to June (average)	0.34

Table A4: This table details the treatment effect of the required HDHP switch under two specifications described in the accompanying text: (i) external validity difference-in-differences using weights derived from income-linked Truven MarketScan data and (ii) Income-linked Truven control group difference-in-differences.

HDHP Switch Differences-in-Differences Analysis

	Model	
	(A1) Ext. Validity Truven-Income Weighted DID	(A2) Truven-Income-Control DID
% Decrease, t_{-1}-t_0	[-2.1%, -6.7%]	[-18.0%, -23.7%]
Semi-Arc Elasticity	[-0.10,-0.34]	[-0.90,-1.18]

Table A5: This table summarizes our descriptive evidence for the heterogeneous treatment effects of the required HDHP switch. For parsimony, the tables presents the estimates from t_{-1}-t_0.

Heterogeneous HDHP Spending Impact	Group %	Spending %	t_{-1} Mean Spending	Treatment Effect (1) Nominal Spending	(2) CPI	(3) Anticipatory Spending
Age 0-17	36.26	24.29	3465.65	-0.07	-0.11	-0.11*
Age 18-29	8.81	7.59	4442.77	-0.15	-0.19	-0.19*
Age 30-54	51.99	62.08	6164.59	-0.19	-0.23	[-0.13,-0.18]
Age 55+	2.92	5.95	11051.14	-0.11	-0.15	[-0.05,-0.11]
Income $0-100K	6.30	6.91	5701.99	-0.03	-0.07	[-0.00, -0.04]
Income $100-150K	63.04	62.98	5209.86	-0.13	-0.17	[-0.08, -0.13]
Income $150-200K	24.93	24.20	5026.86	-0.15	-0.18	[-0.15, -0.17]
Income $200K+	5.73	5.91	5340.94	-0.12	-0.15	[-0.09,-0.12]
Employee	33.47	35.77	5532.76	-0.20	-0.23	[-0.12,-0.18]
Spouse	23.92	35.12	7495.02	-0.16	-0.20	[-0.10,-0.16]
Dependent	42.61	29.11	3570.33	-0.08	-0.12	-0.12*
ACG Quartile 1**	28.51	9.74	1643.56	-0.25	-0.28	-0.28*
ACG Quartile 2**	23.83	12.15	2824.78	-0.39	-0.41	[-0.39,-0.40]
ACG Quartile 3**	23.53	21.45	4564.50	-0.36	-0.38	[-0.33,-0.36]
ACG Quartile 4**	24.13	56.66	12335.85	-0.21	-0.25	[-0.18,-0.22]
ACG Top 1%**	0.79	9.33	66606.47	-0.25	-0.28	-0.28*
0 Chronic Cond.	62.78	38.34	3202.64	-0.15	-0.19	[-0.16,-0.18]
1-2 Chronic Cond.	33.13	47.38	7240.37	-0.18	-0.22	[-0.18, -0.20]
3+ Chronic Cond.	4.19	14.18	19093.34	-0.13	-0.17	[-0.05,-0.12]
Inpatient Hosp.		16.53	863.48	-0.09	-0.13	[-0.07,-0.11]
Outpatient Hosp.		18.07	944.15	-0.13	-0.17	[-0.06,-0.12]
ER		3.11	162.40	-0.21	-0.25	-0.25*
Office Visit		7.61	397.86	-0.15	-0.18	[-0.13,-0.16]
RX		16.91	883.62	-0.16	-0.19	[-0.15,-0.17]
RX - Brand		12.23	638.82	-0.16	-0.20	[-0.16,-0.18]
RX - Generic		4.05	211.62	-0.15	-0.19	[-0.19,-0.19]
Mental Health		9.45	493.86	-0.02	-0.06	-0.06*
Preventive		9.50	496.28	-0.06	-0.10	[-0.05,-0.08]
Other		22.94	1198.07	-0.26	-0.29	[-0.17,-0.24]

*Anticipatory spending estimate itself is negative or not significant from 0

Table A6: This table gives the excess mass calculations (with their associated standard error) for each category of individual spending, calculated as detailed in Appendix 5. These excess mass calculations are used in the construction of the final column of Table A5.

Excess Mass Calculation			Individual Month Calculations		
	Total Excess Mass	Standard Error	October	November	December
Age 0-17	-85.51	12.09	-26.65	-43.50	-15.37
Age 18-29	-33.24	38.13	-20.89	-2.70	-9.65
Age 30-54	253.49	8.65	42.24	61.23	150.01
Age 55+	525.20	78.48	110.05	68.57	346.58
Income 0-100K	201.84	29.77	99.47	28.29	74.08
Income 100-150K	190.07	15.36	43.67	52.99	93.41
Income 150-200K	71.60	21.73	0.20	19.47	51.93
Income 200K+	126.37	23.98	51.14	28.09	47.14
Employee	243.51	9.75	46.09	46.36	151.06
Spouse	308.67	19.70	53.90	89.33	165.44
Dependent	-91.79	13.15	-32.01	-41.88	-17.90
ACG Quartile 1	0.12	7.72	-3.15	2.18	1.09
ACG Quartile 2	42.49	11.94	-9.33	18.68	33.14
ACG Quartile 3	101.35	11.69	29.46	-13.83	85.72
ACG Quartile 4	446.90	26.67	77.45	107.11	262.34
ACG Top 1%	139.48	664.99	-945.06	-1068.03	2152.57
0 Chronic Conditions	56.33	9.10	9.13	14.57	32.63
1-2 Chronic Conditions	118.64	16.04	10.94	5.75	101.94
3+ Chronic Conditions	985.15	65.44	102.65	165.03	717.47
Inpatient Hosp.	25.89	8.79	9.80	1.81	14.27
Outpatient Hosp.	48.37	3.70	8.05	15.95	24.38
ER	-1.40	0.69	-1.64	-1.20	1.44
Office Visit	12.48	1.02	2.56	4.04	5.88
RX	18.87	1.47	0.94	5.54	12.39
RX - Brand	11.93	1.05	-0.39	3.50	8.83
RX - Generic	1.82	0.58	0.06	0.35	1.42
Mental Health	-5.58	1.96	2.30	-4.63	-3.25
Preventive	11.52	1.15	1.96	3.58	5.99
Other	61.34	2.44	14.58	18.56	28.20

Table A7: This table presents the standard errors for the treatment effects given in Table A5.

	(1) Nominal Spending	(2) CPI	(3) Anticipatory Spending
Age 0-17	4.0	3.8	[4.3, 0.8]
Age 18-29	3.3	3.1	[3.6, 1]
Age 30-54	1.2	1.2	[1.3, 0.3]
Age 55+	9.4	9.1	[10, 1.6]
Income 0-100K	17.4	16.8	[17.1, 1.3]
Income 100-150K	2.6	2.5	[2.9, 0.4]
Income 150-200K	1.9	1.9	[2.1, 0.4]
Income 200K+	4.0	3.8	[4.3, 0.8]
Employee	2.3	2.2	[2.4, 0.4]
Spouse	1.8	1.7	[2, 0.5]
Dependent	3.3	3.2	[3.6, 0.6]
ACG Quartile 1	3.8	3.7	[4.2, 0.5]
ACG Quartile 2	2.7	2.6	[2.8, 0.4]
ACG Quartile 3	2.8	2.7	[3, 0.3]
ACG Quartile 4	1.6	1.6	[1.7, 0.3]
ACG Top 1%	5.8	5.6	[6.5, 1.1]
0 Chronic Conditions	2.1	2.0	[2.3, 0.4]
1-2 Chronic Conditions	2.1	2.0	[2.3, 0.3]
3+ Chronic Conditions	3.9	3.7	[4, 0.7]
Inpatient	7.6	7.3	[7.8, 1.4]
Outpatient Hosp.	2.2	2.1	[2.5, 0.5]
ER	1.3	1.3	[1.4, 0.4]
Office Visit	0.4	0.4	[0.4, 0.1]
RX	1.1	1.1	[1.2, 0.2]
RX - Brand	1.5	1.5	[1.6, 0.3]
RX - Generic	0.7	0.6	[0.7, 0.2]
Mental Health	2.4	2.3	[2.5, 0.5]
Preventive	0.9	0.8	[0.9, 0.2]
Other	1.1	1.1	[1.2, 0.3]

Table A8: This table summarizes our descriptive evidence for the heterogeneous treatment effects of the required HDHP switch, for estimates giving the effect between t_{-1} and t_1.

Heterogeneous HDHP Spending Impact				Treatment Effect		
				(1)	(2)	(3)
	Group %	Spending %	t_{-1} Mean Spending	Nominal Spending	CPI	Anticipatory Spending
Age 0-17	34.41	22.83	3465.65	-0.03	-0.11	-0.11*
Age 18-29	8.39	7.13	4442.77	-0.07	-0.15	-0.15*
Age 30-54	49.45	58.37	6164.59	-0.12	-0.19	[-0.09,-0.14]
Age 55+	2.65	5.60	11051.14	-0.07	-0.15	[-0.04,-0.09]
Income 0-100K	6.09	6.64	5701.99	-0.02	-0.10	[-0.01,-0.06]
Income 100-150K	61.34	61.19	5209.86	-0.09	-0.17	[-0.08,-0.12]
Income 150-200K	24.50	23.58	5026.86	-0.07	-0.14	[-0.11,-0.13]
Income 200K+	5.31	5.43	5340.94	-0.08	-0.16	[-0.10,-0.13]
Employee	31.66	33.54	5532.77	-0.07	-0.15	[-0.04,-0.09]
Spouse	22.85	32.79	7495.02	-0.12	-0.20	[-0.10,-0.15]
Dependent	40.38	27.61	3570.33	-0.02	-0.11	-0.11*
ACG Quartile 1	27.21	8.56	1643.56	-0.09	-0.17	-0.17*
ACG Quartile 2	22.63	12.24	2824.79	-0.29	-0.35	[-0.31,-0.33]
ACG Quartile 3	22.36	19.54	4564.51	-0.26	-0.32	[-0.27,-0.29]
ACG Quartile 4	22.69	53.59	12335.85	-0.02	-0.10	[-0.01,-0.06]
ACG Top 1%	0.69	8.80	66606.47	-0.05	-0.13	-0.13*
0 Chronic Conditions	59.76	36.65	3202.64	-0.07	-0.14	[-0.10,-0.12]
1-2 Chronic Conditions	31.34	43.46	7240.37	-0.04	-0.13	[-0.09,-0.11]
3+ Chronic Conditions	3.78	13.83	19093.35	0.02	-0.07	[0.06,0]
Inpatient		16.53	863.48	-0.13	-0.20	[-0.13,-0.16]
Outpatient Hosp.		18.08	944.16	-0.08	-0.15	[-0.03,-0.09]
ER		3.11	162.41	0.12	0.03	0.03*
Office Visit		7.62	397.86	-0.10	-0.18	[-0.10,-0.14]
RX		16.92	883.62	-0.01	-0.09	[-0.04,-0.07]
RX - Brand		12.23	638.83	-0.08	-0.16	[-0.11,-0.14]
RX - Generic		4.05	211.62	-0.17	-0.24	[-0.22,-0.23]
Mental Health		9.46	493.87	0.07	-0.02	-0.02*
Preventive		9.50	496.29	0.01	-0.07	[-0.02,-0.05]
Other		22.94	1198.08	-0.21	-0.27	[-0.15,-0.21]

Table A9: This table measures heterogeneous treatment effects by ACG quartile in two alternative ways.

Heterogeneous HDHP Spending Impact				Treatment Effect		
				(1)	(2)	(3)
	Group %	Spending %	t_{-1} Mean Spending	Nominal Spending	CPI	Anticipatory Spending
t_{-2} Quartile 1	23.86	7.59	1636.85	-0.26	-0.29	[-0.28,-0.28]
t_{-2} Quartile 2	23.64	11.53	2592.70	-0.33	-0.36	[-0.33,-0.35]
t_{-2} Quartile 3	23.60	20.03	4412.69	-0.37	-0.39	[-0.35,-0.37]
t_{-2} Quartile 4	23.74	54.78	12051.12	-0.22	-0.25	[-0.16,-0.21]
t_{-1} Quartile 1	32.29	10.99	1752.40	-0.24	-0.27	[-0.26,-0.27]
t_{-1} Quartile 2	24.49	14.74	3209.34	-0.38	-0.40	[-0.34,-0.37]
t_{-1} Quartile 3	19.07	19.15	5174.46	-0.36	-0.39	[-0.32,-0.35]
t_{-1} Quartile 4	18.99	49.05	13617.06	-0.20	-0.24	[-0.15,-0.20]

Table A10: This table gives Spearman's rank correlation coefficient for provider rankings in prices for a given procedure across year pairs in our data.

Years	Rank Correlation
t_{-4}-t_{-3}	0.9363
t_{-3}-t_{-2}	0.9370
t_{-2}-t_{-1}	0.9275
t_{-1}-t_0	0.9321
t_0-t_1	0.9371

Table A11: This table analyzes price shopping behavior, comparing new employees at the firm in t_{-1} to new employees in t_0.

	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$	$\Delta PS_{t+1,t}$	$\Delta Q_{t+1,t}$
All Claims	-10.4%	1.3%	1.6%	-16.5%
Preventive w/ Diagnosis	-7.5%	1.8%	0.7%	-10.2%
Preventive Always	3.3%	6.8%	0.6%	-6.5%
Imaging	-22.2%	-0.1%	4.5%	-22.4%

Table A12: This table presents the proportion of positive % changes for each part of the spending decomposition, for all 30 of the medical procedures that the firm and its employees spent the most money on, for year pairs from t_{-3} to t_1. The decomposition for the spending change from $t_{-1} - t_0$ is presented for each of these 30 procedures in Table A13 in Appendix 5.

Total Spending Change
Decomposition
High Spend Procedures

	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$	$\Delta PS_{t+1,t}$	$\Delta QE_{t+1,t}$*
% of top 30 w/ Positive Value				
t_{-3}-t_{-2}	80%	63%	43%	73%
t_{-2}-t_{-1}	80%	70%	63%	80%
t_{-1}-t_0	13%	53%	60%	17%
t_0-t_1	77%	37%	57%	80%

*We only present $\Delta QE_{t+1,t}$, given that these results are for one procedure at a time.

Table A13: This table presents the results for our decomposition of the total reduction in medical spending between t_{-1} and t_0, for the top 30 procedures by firm-wide spending.

	% Total Spend	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$	$\Delta PS_{t+1,t}$	$\Delta QE_{t+1,t}$
Routine Vaginal Birth (59400)	2.7%	-13.6%	-15.4%	1.4%	0.4%
Infliximab, 10mg (J1745)	2.6%	24.1%	10.2%	-2.6%	16.6%
MRI, Brain (70553)	2.0%	-6.1%	4.7%	-1.8%	-9.0%
Surgical Pathology, Skin (88305)	2.0%	-9.1%	-1.7%	-2.9%	-4.5%
Routine Cesarean Section Birth (59510)	1.9%	-19.1%	-16.8%	-0.1%	-2.2%
CT Scan, Abdomen and Pelvis (74177)	1.9%	-35.1%	-11.2%	-3.5%	-20.5%
Mammography Screening (G0202)	1.5%	-7.6%	0.3%	1.1%	-8.9%
Anesthesia for Vaginal Birth (01967)	1.3%	-15.4%	-1.0%	1.0%	-15.4%
Colonoscopy, with Biopsy (45380)	1.3%	-28.3%	2.6%	0.6%	-31.6%
MRI, Hip/Knee/Ankle (73721)	1.3%	-24.8%	1.2%	2.3%	-28.4%
Upper Gastrointestinal Endoscopy (43239)	1.2%	-24.2%	2.6%	1.1%	-27.9%
Colonoscopy, Diagnostic (45378)	1.1%	-28.5%	0.5%	2.2%	-31.2%
Wart Removal (17110)	1.1%	-24.9%	2.9%	0.7%	-28.4%
Foot, Molded Insert (L3000)	1.1%	-60.3%	2.0%	1.4%	-63.7%
Transvaginal Echography (76830)	1.0%	-21.5%	2.2%	-0.3%	-23.4%
Globulin, 500mg (J1561)	1.0%	49.7%	99.7%	0.0%	-50.0%
Pegfilgrastim, 6mg (J2505)	0.9%	28.0%	-1.2%	7.7%	21.4%
Fetal Non-Stress Test (59025)	0.8%	-11.5%	-4.7%	-8.5%	1.7%
Trastuzumab, 10mg (J9355)	0.8%	16.5%	-19.1%	0.2%	35.4%
Disposable Contact Lens (S0500)	0.7%	-5.9%	3.1%	4.7%	-13.7%
Laparoscopic Cholecystectomy (47563)	0.7%	-27.2%	4.3%	-3.4%	-28.1%
Ultrasound (76817)	0.7%	-17.8%	-5.7%	1.7%	-13.8%
Blood Count Test (85025)	0.7%	-5.0%	-1.7%	5.0%	-8.4%
Ultrasound (76811)	0.7%	-24.4%	-2.2%	1.2%	-23.3%
Echography of Pregnant Uterus (76805)	0.7%	-23.5%	-3.2%	-1.0%	-19.3%
Chest X-Ray (71020)	0.6%	-24.3%	5.7%	0.0%	-30.0%
Ultrasound (76801)	0.6%	-23.1%	0.4%	-0.6%	-22.9%
CT Scan, Abdomen and Pelvis (74176)	0.6%	-34.0%	-26.5%	13.1%	-20.6%
Thyroid Stimulating Hormone (84443)	0.6%	-8.3%	-2.3%	1.5%	-7.5%
MRI, Lumbar (72148)	0.6%	-26.6%	10.6%	-5.4%	-31.8%

Table A14: This table presents the results for our spending reduction decomposition, applied to prescription drugs. The numbers in parenthesis in the first column indicate the percentage of drugs used that are brand vs. generic.

Prescription Drug Spending Change Decomposition	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t} + \Delta PS_{t+1,t}$	$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
t_{-4}-t_{-3}	10.1%	6.4%	3.6%	0.1%
— Brand (38.8%)	10.5%	14.0%	-3.0%	-0.5%
— Generic (61/2%)	16.3%	5.2%	10.5%	0.6%
t_{-3}-t_{-2}	6.6%	5.3%	1.2%	0.1%
— Brand (35.3%)	7.5%	13.1%	-4.9%	0.7%
— Generic (64.7%)	8.3%	1.1%	7.1%	0.1%
t_{-2}-t_{-1}	4.2%	-0.2%	4.5%	-0.1%
— Brand (32.9%)	7.1%	6.7%	0.3%	0.1%
— Generic (67.1%)	-4.1%	-10.4%	6.9%	-0.6%
t_{-1}-t_{0}	-21.3%	-4.3%	-17.8%	0.8%
— Brand (28.7%)	-20.7%	13.6%	-30.3%	-4.0%
— Generic (71.3%)	-22.4%	-12.0%	-11.8%	1.4%
t_{0}-t_{1}	13.9	5.3%	8.1%	0.5%
— Brand (25.1%)	19.1%	17.5%	1.3%	0.3%
— Generic (74.9%)	-2.7%	-10.2%	8.3%	-0.8%

Table A15: This table presents our spending change decomposition for types of health care that are potentially of high value to consumers, for medical care, for all year pairs in our data. This table extends Table 3.7.

Medical Care	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$	$\Delta PS_{t+1,t}$	$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Preventive Care, General					
t_{-4}-t_{-3}	4.0%	3.9%	-2.1%	-5.7%	7.9%
t_{-3}-t_{-2}	4.1%	-1.6%	9.2%	-0.4%	-3.1%
t_{-2}-t_{-1}	1.3%	-6.5%	-0.5%	6.3%	2.0%
t_{-1}-t_{0}	-0.3%	6.4%	2.1%	-7.5%	-1.3%
t_{0}-t_{1}	13.0%	12.6%	4.8%	-5.2%	0.8%
Preventive Care w/ Prior Diagnosis					
t_{-4}-t_{-3}	1.5%	3.0%	-0.8%	-0.4%	-0.3%
t_{-3}-t_{-2}	3.0%	2.4%	-0.7%	0.1%	1.2%
t_{-2}-t_{-1}	13.0%	3.6%	0.8%	7.3%	1.3%
t_{-1}-t_{0}	-10.6%	2.0%	1.0%	-12.2%	-1.4%
t_{0}-t_{1}	10.3%	5.8%	-0.2%	3.8%	0.9%
Preventive Care, Diabetics					
t_{-4}-t_{-3}	11.9%	3.3%	-0.9%	9.5%	0.0%
t_{-3}-t_{-2}	15.9%	-1.9%	2.9%	12.5%	2.4%
t_{-2}-t_{-1}	15.5%	1.2%	4.4%	9.7%	0.2%
t_{-1}-t_{0}	-1.5%	-2.0%	-0.5%	-1.6%	2.7%
t_{0}-t_{1}	0.1%	-16.5%	3.6%	16.4%	-3.3%
Mental Health					
t_{-4}-t_{-3}	15.5%	-0.3%	-0.1%	13.9%	2.0%
t_{-3}-t_{-2}	16.2%	-1.3%	0.0%	14.8%	2.8%
t_{-2}-t_{-1}	7.1%	-5.1%	-0.2%	11.5%	1.0%
t_{-1}-t_{0}	-2.9%	-1.0%	0.2%	-5.4%	3.4%
t_{0}-t_{1}	2.9%	-2.7%	1.3%	1.9%	2.4%
Physical Therapy					
t_{-4}-t_{-3}	16.6%	-0.2%	-0.1%	16.1%	0.7%
t_{-3}-t_{-2}	13.5%	0.9%	3.1%	8.5%	0.9%
t_{-2}-t_{-1}	9.2%	-1.1%	-1.7%	10.5%	1.4%
t_{-1}-t_{0}	-23.9%	0.3%	7.2%	-29.7%	-1.7%
t_{0}-t_{1}	1.4%	1.2%	-1.7%	0.5%	1.4%

Table A16: This table presents our spending change decomposition for types of health care that are potentially of high value to consumers, for drugs, for all year pairs in our data. This table extends Table 3.7.

Drugs	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t} + \Delta PS_{t+1,t}$	$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Diabetes Drugs				
t_{-4}-t_{-3}	12.0%	12.1%	0.0%	0.0%
t_{-3}-t_{-2}	29.2%	14.8%	12.6%	1.9%
t_{-2}-t_{-1}	20.1%	4.9%	14.5%	0.7%
t_{-1}-t_0	-44.5%	6.7%	-48.0%	-3.2%
t_0-t_1	175.7%	27.9%	115.6%	32.2%
Statins (for cholesterol)				
t_{-4}-t_{-3}	18.2%	15.1%	2.6%	0.4%
t_{-3}-t_{-2}	14.6%	16.8%	-1.9%	-0.3%
t_{-2}-t_{-1}	-31.3%	-8.8%	-24.7%	2.2%
t_{-1}-t_0	-47.1%	-34.3%	-19.6%	6.7%
t_0-t_1	59.6%	23.1%	29.6%	6.8%
Antidepressants				
t_{-4}-t_{-3}	-13.1%	-5.4%	-8.1%	0.4%
t_{-3}-t_{-2}	12.0%	0.4%	11.6%	0.0%
t_{-2}-t_{-1}	-17.7%	-19.7%	2.6%	-0.5%
t_{-1}-t_0	-48.7%	-37.4%	-18.0%	6.8%
t_0-t_1	4.2%	-30.3%	49.5%	-15.0%
Hypertension Drugs				
t_{-4}-t_{-3}	8.6%	1.5%	7.0%	0.1%
t_{-3}-t_{-2}	16.3%	3.2%	12.7%	0.4%
t_{-2}-t_{-1}	-9.0%	-10.9%	2.2%	-0.2%
t_{-1}-t_0	-27.9%	-4.9%	-24.2%	1.2%
t_0-t_1	14.8%	-22.1%	47.3%	-10.4%

Table A17: This table presents our spending change decomposition for types of medical care that are potentially of low value to consumers, for all year pairs in our data. This table extends Table 3.8.

Medical Care	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t}$	$\Delta PS_{t+1,t}$	$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Imaging					
t_{-4}-t_{-3}	7.5%	5.6%	0.1%	3.1%	-1.3%
t_{-3}-t_{-2}	5.5%	2.7%	-1.9%	6.3%	-1.6%
t_{-2}-t_{-1}	12.4%	0.4%	0.2%	13.5%	-1.7%
t_{-1}-t_{0}	-19.5%	-0.4%	0.6%	-17.7%	-2.0%
t_{0}-t_{1}	-2.3%	-2.3%	3.7%	1.1%	-4.8%
CT Scan for Sinuses w/ Acute Sinusitis					
t_{-4}-t_{-3}	-17.3%	4.7%	-0.3%	-22.0%	0.3%
t_{-3}-t_{-2}	11.3%	0.4%	3.9%	5.2%	1.7%
t_{-2}-t_{-1}	30.9%	-8.1%	-3.3%	43.5%	-1.2%
t_{-1}-t_{0}	-24.8%	0.6%	1.1%	-26.0%	-0.4%
t_{0}-t_{1}	-10.7%	-4.6%	1.8%	-8.0%	0.2%
Back Imaging for Non-Specific Low Back Pain					
t_{-4}-t_{-3}	-7.0%	-0.4%	-4.9%	5.5%	-7.2%
t_{-3}-t_{-2}	22.3%	4.3%	-7.6%	14.5%	11.1%
t_{-2}-t_{-1}	9.9%	-1.6%	-5.8%	17.4%	-0.1%
t_{-1}-t_{0}	-26.1%	6.9%	-6.8%	-21.3%	-4.9%
t_{0}-t_{1}	22.1%	-7.4%	5.0%	11.8%	12.7%
Head Imaging for Uncomplicated Headache					
t_{-4}-t_{-3}	-5.0%	3.6%	5.6%	-12.6%	-1.5%
t_{-3}-t_{-2}	18.0%	0.4%	-1.8%	17.9%	1.4%
t_{-2}-t_{-1}	4.2%	1.3%	-12.6%	15.0%	0.5%
t_{-1}-t_{0}	-23.9%	-1.0%	6.6%	-30.7%	1.2%
t_{0}-t_{1}	-10.3%	-11.0%	0.0%	-0.7%	1.4%
Colorectal Cancer Screening for Patients Under 50					
t_{-4}-t_{-3}	-2.0%	5.2%	-1.1%	-6.0%	-0.2%
t_{-3}-t_{-2}	7.6%	1.3%	5.2%	-3.4%	4.5%
t_{-2}-t_{-1}	47.5%	-1.0%	4.4%	25.1%	19.0%
t_{-1}-t_{0}	-32.3%	0.7%	-0.8%	-26.2%	-6.0%
t_{0}-t_{1}	12.2%	0.1%	2.9%	3.6%	5.6%

Table A18: This table presents our spending change decomposition for types of drugs that are potentially of low value to consumers, for all year pairs in our data. This table extends Table 3.8.

Drugs	$\Delta TS_{t+1,t}$	$\Delta PPI_{t+1,t} + \Delta PS_{t+1,t}$	$\Delta Q_{t+1,t}$	$\Delta QS_{t+1,t}$
Antibiotics for Acute Respiratory Inspection				
t_{-4}-t_{-3}	13.5%	4.2%	8.9%	0.4%
t_{-3}-t_{-2}	-4.8%	-5.3%	0.5%	0.0%
t_{-2}-t_{-1}	-34.2%	-18.9%	-18.8%	3.6%
t_{-1}-t_{0}	-47.8%	-6.2%	-44.4%	2.8%
t_{0}-t_{1}	4.4%	-29.1%	47.3%	-13.8%

Table A19: This table replicates the analysis from Table A5, applied to the categories of care defined as high value and low value in Section 4.

			Treatment Effect		
			(1)	(2)	(3)
	Spending %	t_{-1} Mean Spending	Nominal Spending	CPI	Anticipatory Spending
Low Value Care					
CT Scan for Sinuses w/ Acute Sinusitis	0.05	2.50	-28.34%	-31.19%	[-16.31%,-25.13%]
Head Imaging for Uncomp. Headache	0.14	7.10	-26.22%	-29.42%	[-27.34%,-28.56%]
Back Imaging for Non-Specific Low Back Pain	0.22	11.62	-21.13%	-24.52%	[-12.25%,-19.24%]
Colorectal Cancer Scrng. for Patients Under 50	0.37	19.45	-30.98%	-33.72%	[-9.52%,-24.07%]
Antibiotics for Acute Respiratory Infection	0.11	5.52	-45.93%	-48.01%	[-67.66%,-54.73%]
All Low Value Care	0.98	51.30	-27.72%	-30.71%	[-20.41%,-26.5%]
High Value Care					
Preventive Care, Diabetes	0.02	0.99	-3.69%	-7.73%	[-5.23%,-6.53%]
Physical Therapy	6.45	337.10	-24.99%	-28.08%	[-24.7%,-26.66%]
Diabetes Drugs	0.54	28.43	-46.66%	-48.28%	[-38.83%,-45.06%]
Antidepressants	0.93	48.82	-46.85%	-48.85%	[-52.48%,-50.08%]
Statins	0.38	19.97	-10.82%	-14.69%	-14.69%*
Hypertension Drugs	0.30	15.55	-33.04%	-35.52%	[-37.41%,-36.26%]
All High Value Care	16.96	885.51	-14.59%	-18.13%	[-18.63%,-18.35%]

*Anticipatory spending estimate negative or not significant from 0

Table A20: This table analyzes consumer reductions in preventive care shifting from the pre-period t_{-1} to the treatment year t_0. It decomposes the reduction in preventive care into the extensive margin, measured by monthly office visits, and an intensive margin, measured by \$ per office visit spent on preventive care. The tables shows how office visits changed, according to two definitions of office visits described in the text, and documents spending on such care per visit.

Preventive Care Decomposition
\$ per Primary Care Visit

	Visits t_{-1}	Visits t_0	\$ per visit t_{-1}	\$ per visit t_0	Ext. Δ%	Int. Δ%
General Preventive Services						
ICD-9 / CPT	143,887	126,406	\$62.57	\$64.79	-12.1%	3.5%
Specialty	63,121	54,218	\$65.46	\$67.68	-14.1%	3.3%
Preventive w/ Prior Diagnosis						
ICD-9 / CPT	143,887	126,406	\$117.18	\$114.59	-12.1%	-2.1%
Specialty	63,121	54,218	\$136.46	\$133.98	-14.1%	-1.8%

Table A21: This table analyzes consumer reductions in preventive care shifting from the pre-period t_{-1} to the treatment year t_0. It decomposes the reduction in preventive care into the extensive margin, measured by monthly office visits, and an intensive margin, measured by the proportion of office visits where any preventive care services are consumed. The tables shows how office visits changed, according to two definitions of office visits described in the text, and documents concordent changes in care per visit. We also present this decomposition for three specific preventive services.

Preventive Care Decomposition
Binary Care Indicator

	Visits t_{-1}	Visits t_0	Prev. Vis. / Vis. t_{-1}	Prev. Vis. / Vis. t_0	Ext. Δ%	Int. Δ%
General Preventive Services						
ICD-9 / CPT	143,887	126,406	0.396	0.409	-12.1%	3.0%
Specialty	63,121	54,218	0.496	0.502	-14.1%	1.3%
Preventive w/ Prior Diagnosis						
ICD-9 / CPT	143,887	126,406	0.768	0.763	-12.1%	-0.7%
Specialty	63,121	54,218	0.792	0.793	-14.1%	-2.4%
Mammography						
ICD-9 / CPT	143,887	126,406	0.0086	0.0074	-12.1%	-14.8%
Specialty	63,121	54,218	0.0170	0.0159	-14.1%	-6.7%
Urinalysis						
ICD-9 / CPT	143,887	126,406	0.0716	0.0694	-12.1%	-3.0%
Specialty	63,121	54,218	0.1234	0.1229	-14.1%	-0.4%
Colonoscopy						
ICD-9 / CPT	143,887	126,406	0.0032	0.0024	-12.1%	-23.4%
Specialty	63,121	54,218	0.0047	0.0031	-14.1%	-35.0%

Table A22: This table presents statistics for our primary sample describing the average and marginal price changes resulting from the required HDHP switch. We present the average % of total spending paid, as well as the likelihood of reaching each arm of the non-linear HDHP contract.

Policy Change: Price Impact					
t_{-1} **Total Spending**					
Coverage Tier	Avg. HDHP Price	% Under Deductible	% Over Ded., Under OOP Max.	% Over OOP Max.	Actuarial Value
0 Dependents	0.428	37.92%	49.16%	12.92%	78.31%
1 Dependent	0.293	23.22%	61.08%	15.70%	76.59%
2+ Dependents	0.201	13.30%	68.40%	18.30%	78.24%
All Tiers	0.249	18.42%	64.46%	17.12%	78.05%

Table A23: This table presents the coefficients on shadow price ventiles for our non-linear contract price regressions.

Ventile Regression Coefficients

	Coefficient	
Ventile	Treatment	Treatment X t_1
2	-0.0516	0.0428
	(0.0454)	(0.0440)
3	-0.0409	0.00463
	(0.0475)	(0.0466)
4	-0.148***	0.0346
	(0.0486)	(0.0474)
5	-0.140***	0.0399
	(0.0489)	(0.0476)
6	-0.164***	0.0915*
	(0.0495)	(0.0482)
7	-0.121**	0.0429
	(0.0494)	(0.0482)
8	-0.0780	0.0835*
	(0.0494)	(0.0483)
9	-0.150***	0.0913*
	(0.0502)	(0.0492)
10	-0.0376	0.0119
	(0.0529)	(0.0522)
11	-0.0891*	0.114**
	(0.0536)	(0.0527)
12	-0.100*	0.0760
	(0.0542)	(0.0531)
13	-0.145***	0.187***
	(0.0545)	(0.0534)
14	-0.171***	0.135**
	(0.0552)	(0.0537)
15	-0.000201	0.0884
	(0.0555)	(0.0539)
16	-0.0212	0.0719
	(0.0557)	(0.0542)
17	0.0403	0.129**
	(0.0562)	(0.0543)
18	0.113**	0.0911*
	(0.0564)	(0.0547)
19	0.185***	0.0933*
	(0.0565)	(0.0550)
20	0.151***	0.120**
	(0.0568)	(0.0551)

Table A24: This table gives mean spending by individuals for a set of months in our data.

Mean Individual Spending By Month		
Month	Mean Spending	Mean Spending, Detrended
t_{-4}, March	352.15	347.91
t_{-4}, June	360.89	351.71
t_{-4}, September	333.98	319.80
t_{-4}, December	358.07	337.26
t_{-3}, March	397.97	365.47
t_{-3}, June	362.47	328.91
t_{-3}, September	351.97	313.95
t_{-3}, December	368.23	324.94
t_{-2}, March	436.87	381.86
t_{-2}, June	412.69	355.13
t_{-2}, September	385.52	327.83
t_{-2}, December	376.79	316.01
t_{-1}, March	471.71	393.43
t_{-1}, June	414.34	338.62
t_{-1}, September	404.86	329.01
t_{-1}, December	526.96	422.53
t_0, March	355.94	282.28
t_0, June	338.97	268.07
t_0, September	372.86	287.69
t_0, December	417.47	322.12
t_1, March	405.21	306.96
t_1, June	386.42	290.04
t_1, September	412.19	307.42
t_1, December	512.89	378.54

Table A25: This table shows the share of families who begin a month in t_0 in a given arm of the non-linear HDHP, as well as total spending by month and plan arm across these families for that month.

Family Shares and Total Spend by HDHP Plan Arm

	February	April	June	August	October	December
Family Shares						
t_0 Deductible Arm	93.1%	77.4%	61.6%	50.2%	40.5%	33.1%
t_0 Coinsurance Arm	6.5%	21.1%	34.9%	44.1%	51.6%	56.0%
t_0 OOP Maximum Arm	0.4%	1.5%	3.4%	5.6%	7.9%	10.9%
Total Spend ($ million)						
t_0 Deductible Arm	10.44	7.93	4.45	3.37	2.54	1.86
t_0 Coinsurance Arm	3.86	6.84	7.59	8.74	9.76	10.24
t_0 OOP Maximum Arm	0.72	2.02	3.13	4.76	5.59	6.25

Table A26: This table shows mean t_0 family shadow prices, i.e. true expected end-of-year marginal prices, as a function of (i) their spot price at the start of a month and (ii) where they fall in the distribution of family expected-of-year price, conditional on their spot price. Quartile 1 is the sickest quartile.

Shadow Prices by Plan Arm and Health Status

	Sickest 10%	Quartile 1	Quartile 2	Quartile 3	Quartile 4
t_0 **Deductible Arm**					
February	0.06	0.08	0.15	0.31	0.58
April	0.09	0.10	0.17	0.40	0.70
June	0.10	0.10	0.22	0.52	0.80
August	0.10	0.11	0.31	0.67	0.88
October	0.10	0.14	0.51	0.83	0.95
December	0.10	0.19	0.75	0.96	0.99
t_0 **Coinsurance Arm**					
February	–	0.01	0.04	0.06	0.10
April	–	0.03	0.06	0.08	0.10
June	–	0.04	0.08	0.09	0.10
August	–	0.05	0.09	0.10	0.10
October	–	0.07	0.09	0.10	0.10
December	–	0.08	0.10	0.10	0.10

Table A27: This table shows the correlation in different non-linear contract prices that we consider in our primary regressions, for months pooled over the treatment years t_0-t_1.

Price Correlations by Month, t_0-t_1

	Spot-Shadow	Spot-Prior End	Shadow-Prior End
February	0.285	0.131	0.627
April	0.489	0.229	0.564
July	0.668	0.315	0.513
October	0.798	0.363	0.460
December	0.857	0.381	0.437

Chapter 6

Appendix For "Intermediation and Vertical Integration in the Market for Surgeons"

Data Construction

Orthopedic Surgery Procedure Codes

In Section 2.4, we decompose the variation across orthopedists into variation from performing surgery, against incurring costs through some other means. To do this, we need a measure of whether an orthopedic surgery occurred. We identify this through the presence of certain Current Procedure Terminology (CPT) codes in each patient's claims. These codes describe what procedure was billed for in the claim. We consulted with a number of billing resources and compiled a list of codes that 1) described a surgical procedure; and 2) were for an obviously orthopedic medical issue. This generated 255 codes, 236 of which were performed on patients in our data. In Table B1, we list the 25 most common surgery codes, and the share of patients who were billed for that code.[1] Note that a patient can receive multiple surgeries, so the computed shares are not exclusive. We do not double-count surgeries for patients who received the same surgery twice. We can see that the most common surgeries are arthroscopic knee surgery to treat meniscus tears (29881 and 29880), total knee (27447) and hip (27130) joint replacements, and arthroscopy shoulder surgeries to treat torn rotator cuffs (29827). The fact that the top 25 surgery codes are all obvious joint surgeries is reassuring that our orthopedist specialty restrictions successfully only captured joint specialists.

[1] A full list of all of the codes can be downloaded at `https://sites.google.com/site/zarekcb/files/surgerycodes.txt`.

Elixhauser Comorbidity Index

To measure patient health status, we turn to the Elixhauser Comorbidity Index. The Index was introduced by Elixhauser, Steiner, Harris and Coffey (1998) as a measure of patient health, for use in risk-adjusting measures of patient spending and health outcomes in in-patient settings. In our setting, it allows us to control for patient health status when we estimate our regression model of 1-year costs. The process of constructing it involves coding patient binary indicators for the presence 30 chronic conditions from ICD-9 diagnosis codes. The Index is simply the sum of the indicators.

For our primary sample, we implement this by using a 12-month look-back from the first orthopedist visit. For each patient, we collect all medical claims incurred in the 12 months prior to the orthopedist visit, and check the diagnosis codes of those claims. For each of the 30 conditions, we define a set of procedure codes whose presence would indicate that the patient suffered from that condition. We define the patient as having a given condition if any diagnosis contained within that condition is billed in this 12 month period.

In Table B2, we define the 30 conditions, and report the share of our patients who suffer from each. The most common chronic conditions are hypertension, depression, and obesity. In Figure B1, we plot the distribution of comorbidity counts. The average patient in our data has 1.8 chronic conditions, but this masks a great deal of heterogeneity, with many patients having zero or one condition, and a handful having over ten.

Empirical Bayes Shrinkage

In Section 2.4, we estimate the impact of a given orthopedist on log 1-year patient total medical expenditures Y_i. In this appendix, we describe the empirical Bayes procedure we use to adjust our estimates of orthopedist costliness in the face of potential measurement error. We describe this in minimal detail. For an extended discussion, see the Appendix of Chandra et al. (2016).

To understand why this procedure is necessary, we must first note that our estimates of orthopedist costliness are estimated with noise. We can depict this as

$$\widehat{\gamma}_k = \gamma_k + e_k$$

where γ_k is the 'true' orthopedist effect on costs, and e_k is measurement error, assumed to be independent of γ_k. The purpose of constructing the estimators $\widehat{\gamma}_k$ is to eventually use as a dependent variable in our regressions in Section 2.5 and as an input into our structural model in Section 2.6. Since the variance of e_k is nonzero, the variance of $\widehat{\gamma}_k$ will be greater than the variance of γ_k. Therefore, without an adjustment, using $\widehat{\gamma}_k$ as a dependent variable in a regression will attenuate our estimates of the effect of γ_k. The empirical Bayes shrinkage procedure helps to correct for this bias.[2]

[2] We are not able to compute what bias might be generated when this measurement error enters a nonlinear model like the one we employ in Section 2.6. In practice we estimate that our measurement error is relatively small, so we do not explore this more deeply.

We assume that $e_k \sim N(0, \pi_k^2)$ independently, so that

$$\widehat{\gamma}_k | \gamma_k, \pi_k^2 \sim N(\gamma_k, \pi_k^2) \text{ independently}$$

where π_k^2 is the variance of the measurement error of $\widehat{\gamma}_k$.

We use a Bayesian prior distribution of γ_k of

$$\gamma_k | \sigma^2 \sim N(0, \sigma^2)$$

where σ^2 is constant across all k. Using Bayes rule, this combined with knowledge of π_k^2 and an estimator $\widehat{\gamma}_k$ produces the following Bayesian posterior distribution of γ_k:

$$\gamma_k | \sigma^2, \pi_k^2, \widehat{\gamma}_k \sim N(\theta_k \widehat{\gamma}_k, \theta_k \pi_k^2)$$

where $\theta_k = \frac{\sigma^2}{\pi_k^2 + \sigma^2}$. $\theta_k \widehat{\gamma}_k$ serves as our empirical Bayes-adjusted estimator of γ_k.

An obvious problem with this procedure is that we observe neither π_k^2, nor σ^2. Instead, we estimate both. We begin by constructing an estimator $\widehat{\pi}_k^2$ of the variance of measurement error for a given orthopedist k. For this, we simply use the standard error of $\widehat{\gamma}_k$.

To estimate σ^2, we first note that our assumptions imply that our assumptions about the prior distribution of γ_k, and the distribution of the estimator $\widehat{\gamma}_k$ imply that

$$\widehat{\gamma}_k | \widehat{\pi}_k^2, \sigma^2 \sim N(0, \widehat{\pi}_k^2 + \sigma^2)$$

Since these assumptions generate distributional assumptions for $\widehat{\gamma}_k$, we can use maximum likelihood methods to recover the final unknown parameter σ^2. Specifically, our estimator is

$$\widehat{\sigma}^2 = \arg\max_{\sigma^2} \phi \left(\frac{\widehat{\gamma}_k}{\sqrt{\widehat{\pi}_k^2 + \sigma^2}} \right)$$

where ϕ is the standard normal probability density function.

Table B3 presents the output from this procedure. We can see that our estimates of $\widehat{\theta}_k$ are very close to 1, and indeed the distribution of $\widehat{\theta}_k \widehat{\gamma}_k$ is not far from the distribution of $\widehat{\gamma}_k$. Only 5 orthopedists have an estimated $\widehat{\theta}_k$ below 0.99, although for two of those the estimated $\widehat{\theta}_k$ is very low, around 0.2. These adjusted estimates are the ones we use in Sections 2.5 and 2.6.

Table B1: A list of the 25 most commonly-billed orthopedic surgery CPT codes and the share of patients who receive them.

Code	Description	Patient Share
29881	Knee arthroscopy with medial or lateral meniscectomy including debridement	4.52%
27447	Total knee arthroplasty (knee replacement with prosthetic)	2.93%
27130	Total hip arthroplasty (hip replacement with prosthetic)	2.73%
29826	Shoulder arthroscopy w decompression of subacromial space with partial acromioplasty (add-on code)	2.37%
29880	Knee arthroscopy with medial and lateral meniscectomy including debridement	1.64%
29827	Shoulder arthroscopy with rotator cuff repair	1.51%
29823	Shoulder arthroscopy with extensive debridement	1.25%
29877	Knee arthroscopy with chrondroplasty	0.87%
29822	Shoulder arthroscopy with limited debridement	0.81%
29824	Shoulder arthroscopy with distal claviculectomy	0.81%
29875	Knee arthroscopy with limited synovectomy	0.79%
29888	Arthroscopically aided anterior cruciate ligament repair or reconstruction	0.77%
64721	Neuroplasty, median nerve at carpal tunnel	0.68%
29876	Knee arthroscopy with major synovectomy	0.58%
29879	Knee arthroscopy with abrasion arthroplasty or multiple drilling or microfracture	0.40%
29828	Shoulder arthroscopy with biceps tenodesis	0.37%
29806	Shoulder arthroscopy with capsulorrhaphy	0.30%
29807	Shoulder arthroscopy with repair of SLAP lesion	0.29%
63030	Lumbar laminotomy with decompression of nerve root(s) and/or excision of herniated disc	0.27%
29862	Hip arthroscopy with chondroplasty, abrasion arthroplasty, and/or resection of labrum	0.24%
29914	Hip arthroscopy with femoroplasty	0.22%
23412	Open repair of chronic ruptured rotator cuff	0.20%
29874	Knee arthroscopy for removal of loose or foreign body	0.19%
63047	Lumbar laminectomy. facetectomy, and foraminotomy, single vertebral segment	0.19%
64635	Destruction of cervical or thoracic paravertibral facet joint nerve(s) by neurolytic agent, with imaging guidance	0.18%

Table B2: A list of the 30 conditions in the Elixhauser Comorbidity Index.

Condition Description	Patient Share
Hypertension	36.2%
Depression	21.3%
Obesity	15.2%
Chronic pulmonary disease	15.0%
Diabetes mellitus	11.9%
Hypothyroidism	11.2%
Arrhythmias	8.8%
Solid tumor without metastasis	5.6%
Fluid and electrolyte disorders	4.8%
Liver disease	4.5%
Rheumatoid arthritis	4.4%
Diabetes mellitus with complications	4.3%
Valvular disease	3.8%
Peripheral vascular disease	3.5%
Renal failure	3.4%
Deficiency anemias	3.4%
Alcohol abuse	3.2%
Drug abuse	2.9%
Other neurological disorders	2.9%
Congestive heart failure	2.6%
Weight loss	2.1%
Coagulopathy	1.6%
Psychoses	1.6%
Disease of pulmonary circulation	1.0%
Metastatic cancer	0.6%
Peptic ulcer disease	0.6%
Lymphoma	0.6%
Chronic blood loss anemia	0.5%
Paralysis	0.5%
AIDS	0.2%

Table B3: Estimate output from our empirical Bayes procedure.

Estimate	γ_k	γ_k^{surg}	γ_k^{other}
Standard Deviation of $\widehat{\gamma}_k$	0.304	0.104	0.233
Average $\widehat{\pi}_k^2$	0.013	0.001	0.010
$\widehat{\sigma}^2$	0.121	0.012	0.060
Average $\widehat{\theta}_k$	0.994	0.998	0.994
Standard Deviation of $\widehat{\theta}_k\widehat{\gamma}_k$	0.294	0.103	0.206

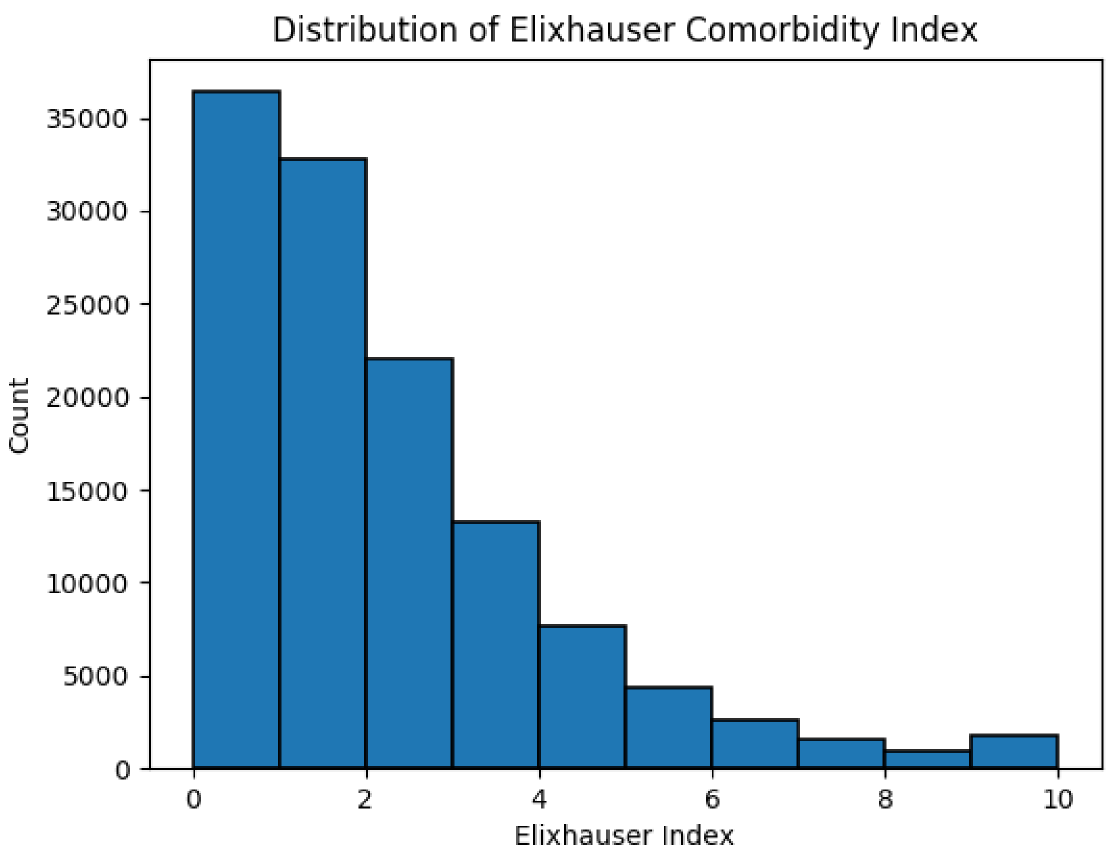


Figure B1: Counts of patient Elixhauser Comorbidity Index values in our data. The final bin includes patients with 10 or more comorbidities.