Intermediation and Vertical Integration in the Market for Surgeons[*]

Zarek C. Brot-Goldberg and Mathijs de Vaan

April 11, 2019

Abstract

Health care markets are increasingly dominated by large, highly-integrated systems. Understanding the impact of this integration on market outcomes is central to designing good antitrust and regulatory policy. Integration presents a central efficiency trade-off: It may improve the productive efficiency of care through coordination between primary care providers (PCPs) and specialists, but it may also reduce allocative efficiency by allowing systems to distort PCPs' care recommendations through encouragement to steer patient referrals towards affiliated specialists. We study how these forces shape referrals to orthopedic joint surgeons in Massachusetts. We find that both are present, but that internal referrals are primarily driven by anticompetitive steering rather than efficiencies, and removing steering incentives would reduce internal referrals by over half. Counterintuitively, dis-integrating health systems would increase expected costs by 5%. We find that this is explained by the absence of referral cost-sensitivity. In the status quo, the only forces that improve allocative efficiency are the steering efforts of low-cost systems. We study insurers' attempts to enhance competition through the use of "global budget" capitation contracts, which force PCPs to bear a share of the cost of their referrals. We find that the introduction of capitation contracts do introduce cost-sensitivity and shift referrals towards orthopedists who are 3-6% less expensive. Nonetheless, these incentives would need to have 42% stronger effects to offset the efficiency losses of dis-integration.

[*] Zarek Brot-Goldberg is deeply grateful to his dissertation committee—Ben Handel, Kei Kawai, and Jon Kolstad—for their invaluable guidance and encouragement. We additionally thank Natalie Bachas, Giovanni Compiani, Keith Ericson, Joe Farrell, Kim Geissler, Jen Gong, Sean Higgins, Jonathan Holmes, Abigail Jacobs, Jennifer Kwok, Jon Schellenberg, Yotam Shem-Tov, Mark Shepard, Avner Shlain, Katalin Springel, Toby Stuart, Ashley Swanson, Tiffany Tsai, Boris Vabson, and seminar participants at UC Berkeley (IO and Haas EAP) for helpful comments. All errors are our own.

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 et al., 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 et al., 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 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

[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.

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 et al. (2016), Cooper et al. (forthcoming), and Chernew et al. (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 et al. (2017) and Sood et al. (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 et al. (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 et al. (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 et al. (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

[3] See Cuellar and Gertler (2006) and Baker et al. (2014) for hospitals and Baker et al. (2017) for specialist physicians.

[4] See Capps et al. (2018).

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) 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 et al., 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, we describe the institutions we study and the data we use. In Section 3, we write down a model of orthopedist referral choice that depends on incentives and potential cost outcomes. Section 4 presents our estimates of the extent of cost dispersion across orthopedist and vertical cost efficiencies. In Section 5 we present reduced-form evidence for how referral patterns depend on integration and global budgets. In Section 6, we estimate a structural model of referrals, while in Section 7 we present the results from counterfactual policy simulations based on these estimates. Section 8 concludes.

[5] E.g. Swanson (2013), Baker et al. (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 et al. (2007), Nakamura (2010), and Walden (2016).

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

2 Setting & Data

2.1 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.

2.2 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

[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.

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.

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 5 on.

2.3 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 1, 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 mechanisms by private insurers, as well. In 2018, after eight years, nearly 33 million lives were

[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.

covered by such an arrangement (Muhlestein et al., 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 et al., 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.

2.4 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 et al., 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 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 net-

[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.

[12] A medical group, in the MPD, nests practices. MHQP defines it as "A "parent" provider organization that may include multiple

work(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 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.

2.5 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

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.

[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.

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 et al. (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 2 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 code. The list of these codes is given in Appendix A1. 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 2. 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.

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 5.4. 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

[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.

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]

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

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'.

[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.

3.1 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 6.1, 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 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

$$C_{ij} = \sum_{k} s_{ijk} Y_{ijk}$$

3.2 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).

3.2.1 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:

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

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

[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

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, 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.

3.2.2 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

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

the sum of market shares is always 1.

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, 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.

3.2.3 Interaction

Finally, we can ask how the effects we have described above change when they interact. In Section 5.3, 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.

3.2.4 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 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 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 4. We can then use these to estimate how PCP incentives change referral patterns, which we do in Sections 5 and 6.

4 Orthopedist Costs

As described in Section 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.

4.1 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 4.3.

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 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 5.We see that our estimates of the distribution of γ_k change by a relatively small amount as controls are added.

[22] In Table 9, 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.

[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 A2.

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 et al. (2016)) as well as the education quality literature (Kane and Staiger (2001), Rose et al. (2018)). We describe this procedure in more detail in Appendix B. The resulting variation from these, our preferred estimates, is described in the fourth column of Table 5. The shrinkage procedure reduces estimated variation by a small amount.

Table 5 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 2. The upper plot is a histogram of the $\gamma_{k(i)}$ faced by each patient 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 7, 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 3, 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 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.

4.2 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 A1. 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 6. 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. 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 8. We can see that the surgery decision alone explains around 30% of the variation in γ_k.

4.3 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 et al. (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 4.1. Figure 5 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 6. 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.

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.

5.1 PCPs and Referral Patterns

We begin by documenting how PCPs play a role in referral patterns. In Table 9, 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 10 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.

5.2 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 3. Ideally, we would be able to observe whether i was covered under a global budget contract exactly. However, as described in Section 2.4, 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 5.4.

[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 $GB_i = 1$.

Our initial estimate is presented in the first column of Table 11. 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 of 1-year spending in our sample ($12,218), this represents a modest reduction in spending of $745 per patient.

5.3 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 11, 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 11. 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 et al., 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 11. Our results are presented in Table 12. 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 10, respectively. We use estimates of $\frac{\partial s_{in}}{\partial GB}$ from Table 12. 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 14. 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 3.2, 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.

5.4 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 15. 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 5.2. To address this, in Figure 7 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 4.1. 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 4.3 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.

5.5 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 4.2, 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 11. 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 13. 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.

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 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.

6.1 Functional Form and Estimation

Recall that in our model of choice from Section 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]

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

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

[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 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

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

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

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

6.2 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

[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.

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.

6.3 Results

We present parameter estimates in Table 16. 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]

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.

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

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

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.

7 Counterfactuals

The parameter estimates presented in Section 6.3 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.

7.1 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 17. 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.

7.2 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 18 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 19 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 19 to the first column of Table 18. Still, even without efficiencies on the table, we still see that removing vertical ties (comparing the first and third columns of Table 19) 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.

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 disin-

[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.

tegrating 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.

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 et al. (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.

References

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.

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.

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 _ , "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.

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.

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.

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.

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.

Chernew, Michael, 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.

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

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

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.

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.

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.

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

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.

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.

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.

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

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.

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

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.

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.

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.

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

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

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.

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.

Tables and Figures

	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 1: 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 4. 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 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 2: Characteristics of our samples of patients. The first column cotnains the full set of patients we enumerate in Section 2.5. 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 A1.

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

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

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 5: 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 B. The units of the rows are in log costs, which can be approximately interpreted as percent changes.

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 6: Estimated parameters from risk-adjusted two-outcomes model. This table presents the same parameters as Table 5, 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 5, but add an additional control for whether or not the patient received a surgery. θ is the estimated coefficient for that control variable.

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 7: 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.

	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 8: 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 $\frac{\text{Var}(\theta\gamma_k^{surg})}{\text{Var}(\gamma_k)}$. The second presents $\frac{\text{Var}(\gamma_k^{other})}{\text{Var}(\gamma_k)}$, and the third presents $\frac{2\cdot\text{Cov}(\theta\gamma_k^{surg},\gamma_k^{other})}{\text{Var}(\gamma_k)}$.

	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 9: 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.

		Avg. $\gamma_k + \eta V_{jk}$ of Referred Orthopedist		
System	Internal Ref. Rate	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 10: Referral patterns of primary care physicians in our data.

	$\gamma_{k(i)} + \eta V_{j(i)k(i)}$		$\gamma^{surg}_{k(i)} + \eta^{surg} V_{j(i)k(i)}$		$\gamma^{other}_{k(i)} + \eta^{other} V_{j(i)k(i)}$	
	(1)	(2)	(1)	(2)	(1)	(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 11: Regressions of patient allocation to surgeons (based on cost and surgery propensity) on global budget contract utilization. β^0 measures the change in orthopedist costliness (or propensity to do surgery, for the third and fourth columns) that would occur through reallocation if patients were covered by a global budget contract. β^M measures the additional effect for a PCP who is affiliated with system M. ζ^M measures the relative difference in average costliness (or propensity to do surgery) of orthopedist referred by PCPs who are affiliated with M.

	$V_{j(i)k(i)}$	
	(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 12: Regressions of PCP tendency to refer to integrated surgeons on global budget contract utilization. Inclusion in regression is conditional on a patient's PCP being integrated with at least one surgeon. β measures the impact of global budget contracts on the propensity to refer internally. ζ measures the difference in underlying propensity to refer internally at different systems.

	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 13: 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 11. Estimates for system-specific rows come from fourth and sixth) regressions in Table 11.

		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 14: 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.

	$Referred_i$	
	(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 15: 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.

	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 16: 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.

	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 17: 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 ($V_{jk} = 0$).

	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 18: 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	0.131	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)

Table 19: Recreation of Table 18, however, in all simulations we have removed efficiencies ($\eta = 0$).

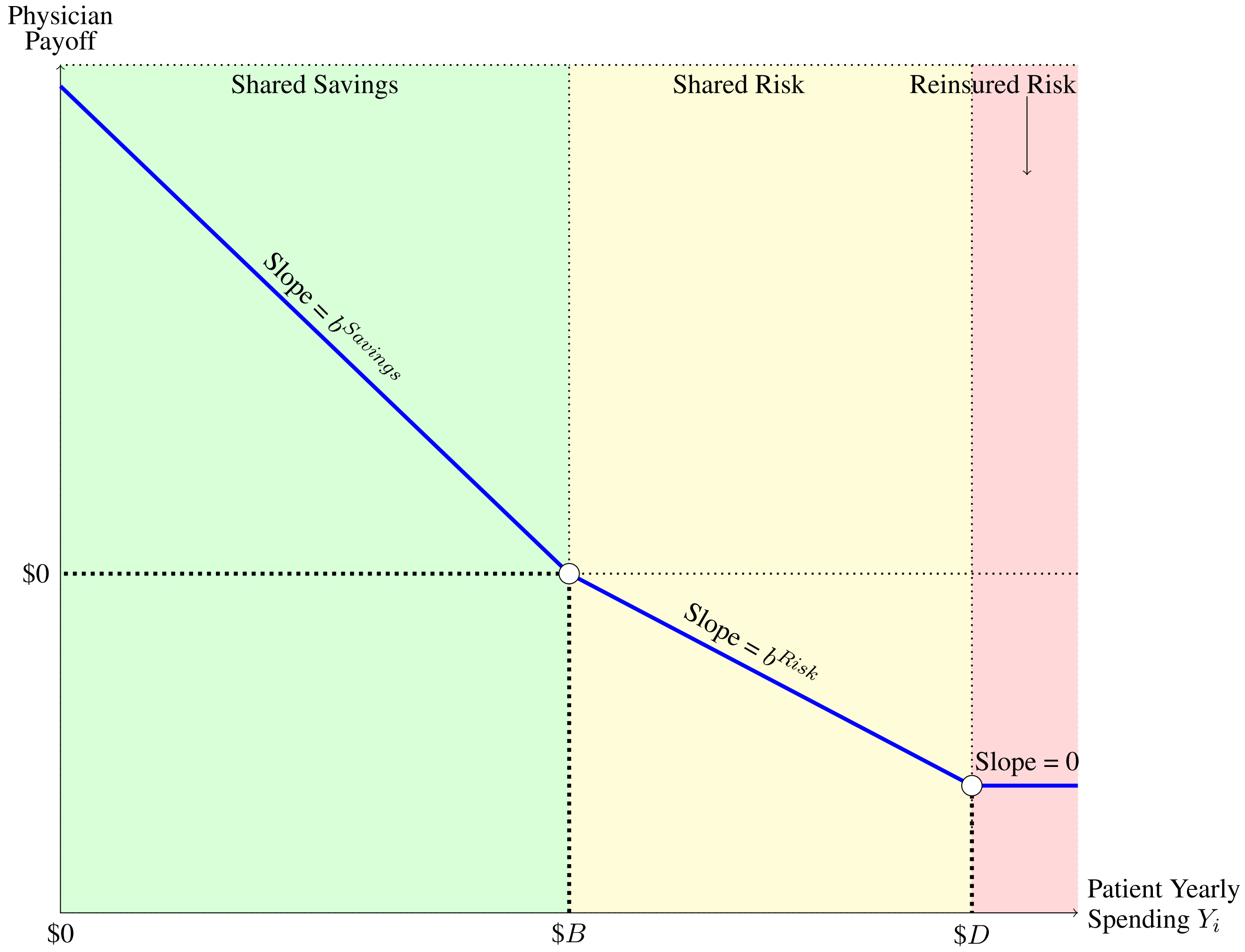


Figure 1: The blue line in this figure depicts a hypothetical global budget contract. The x-axis is the amount of total medical expenditures a given patient incurs during a year, whereas the y-axis is the total dollar amount transferred from the insurer to the patient's primary care provider. The contract 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.

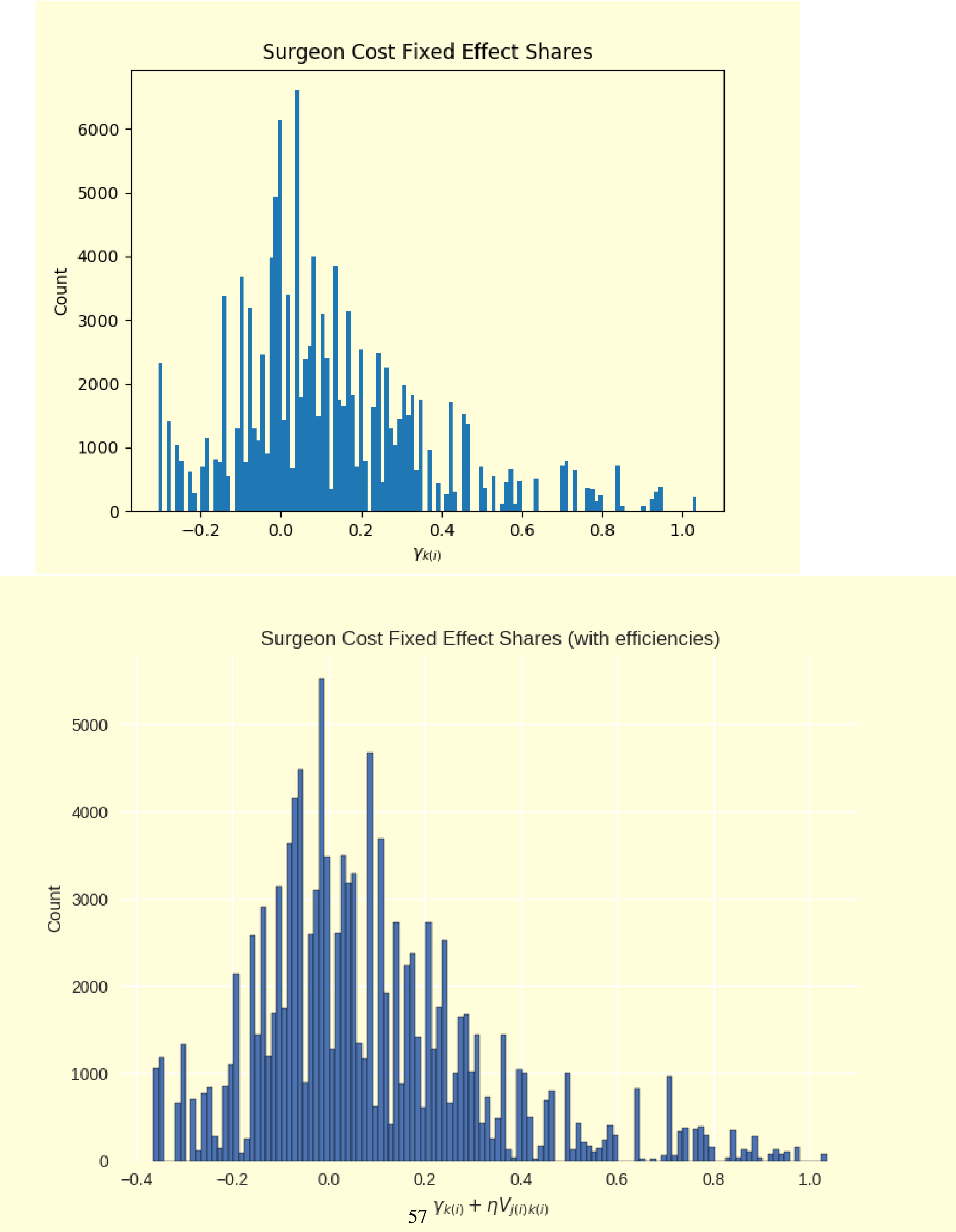


Figure 2: Distribution of surgeon fixed effects $\gamma_{k(i)}$ and expected surgeon costs $\gamma_{k(i)} + \eta V_{j(i)k(i)}$, by patient.

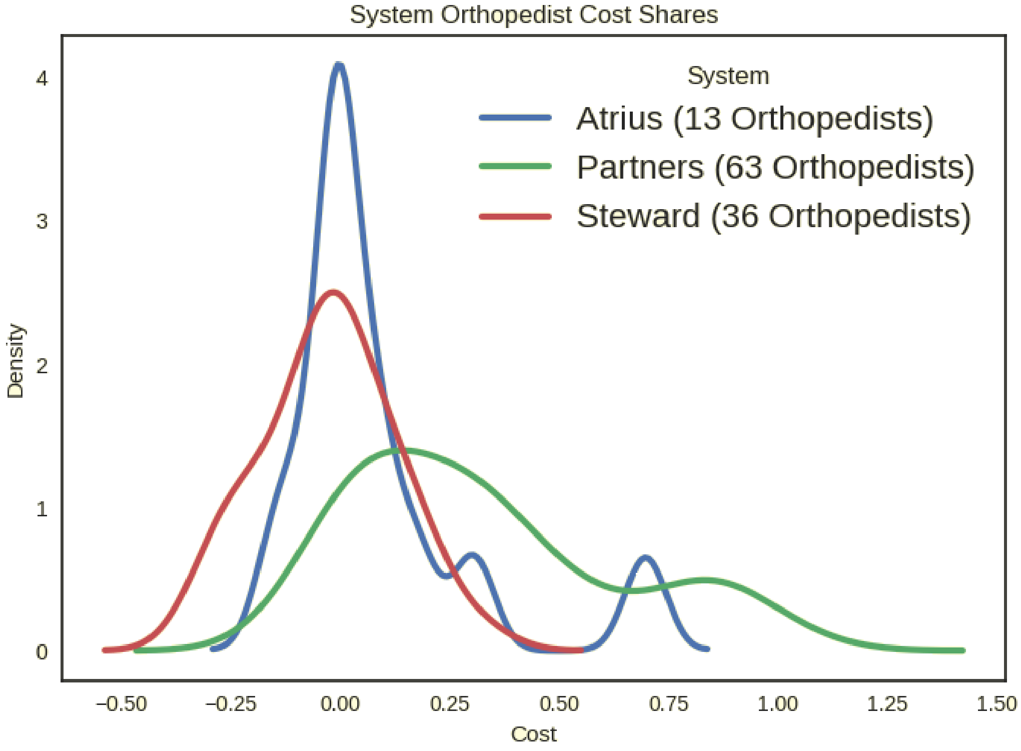


Figure 3: Distribution of surgeon fixed effects $\gamma_{k(i)}$ for orthopedists in three major health systems: Atrius, Partners, and Steward..

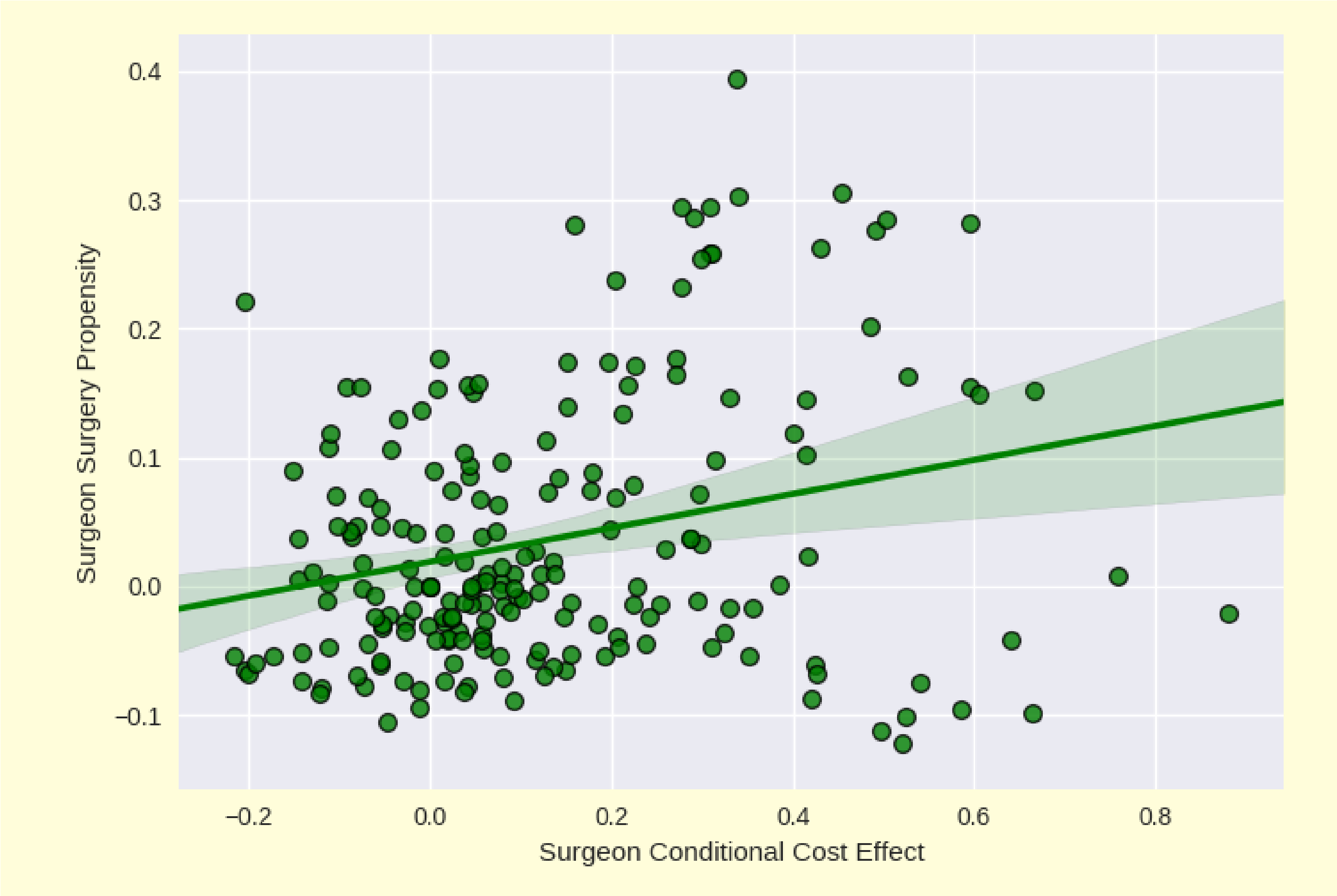
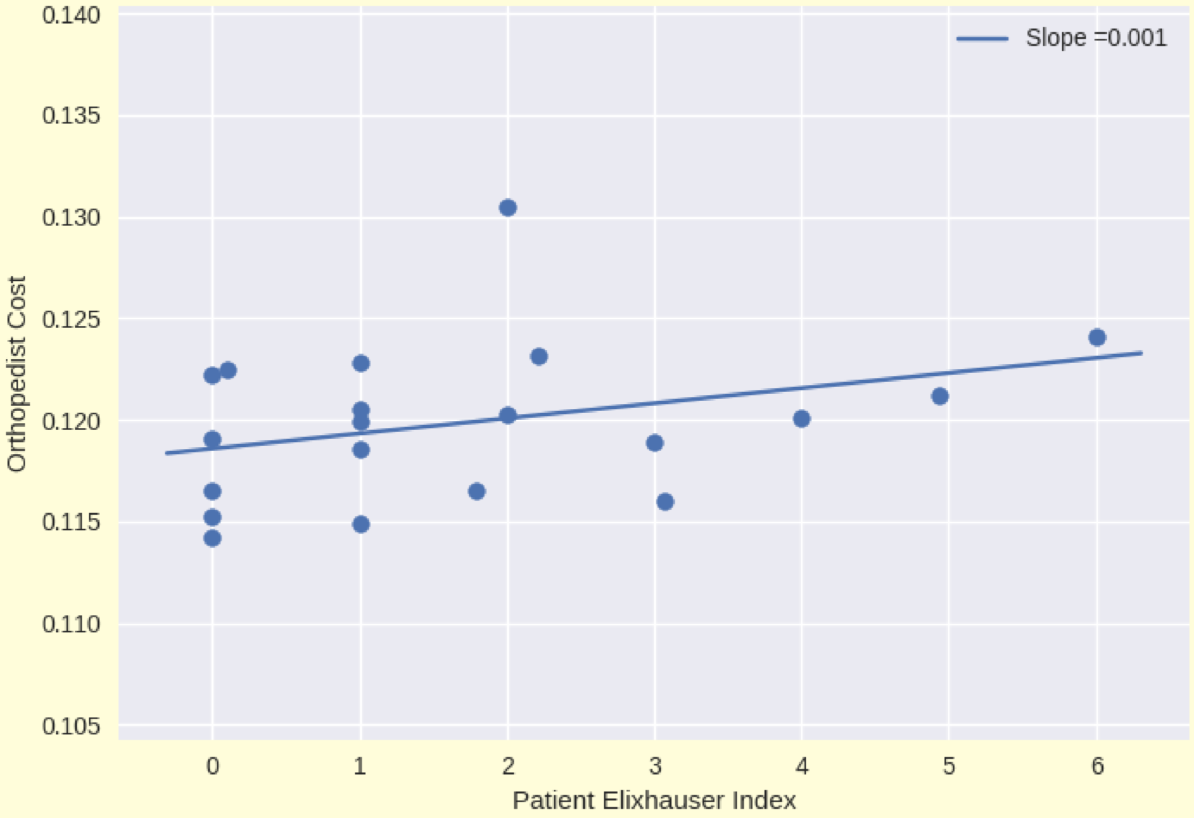
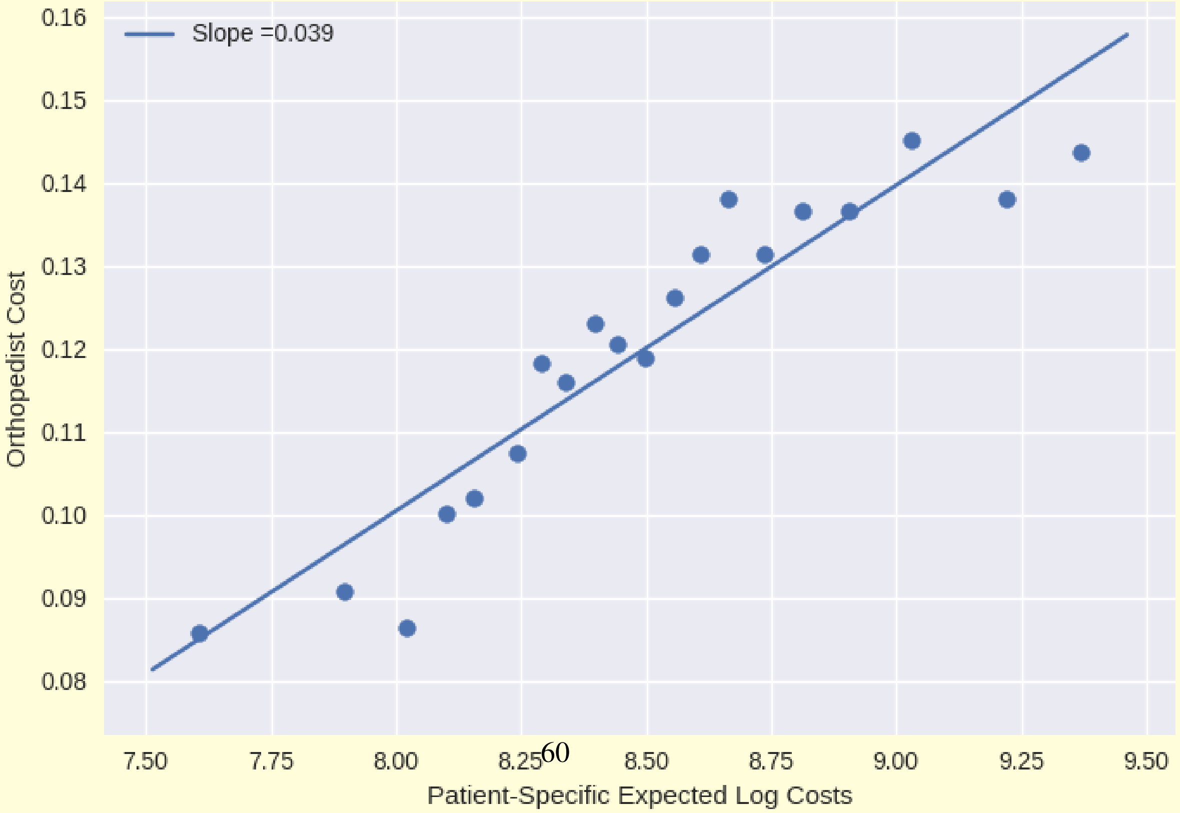
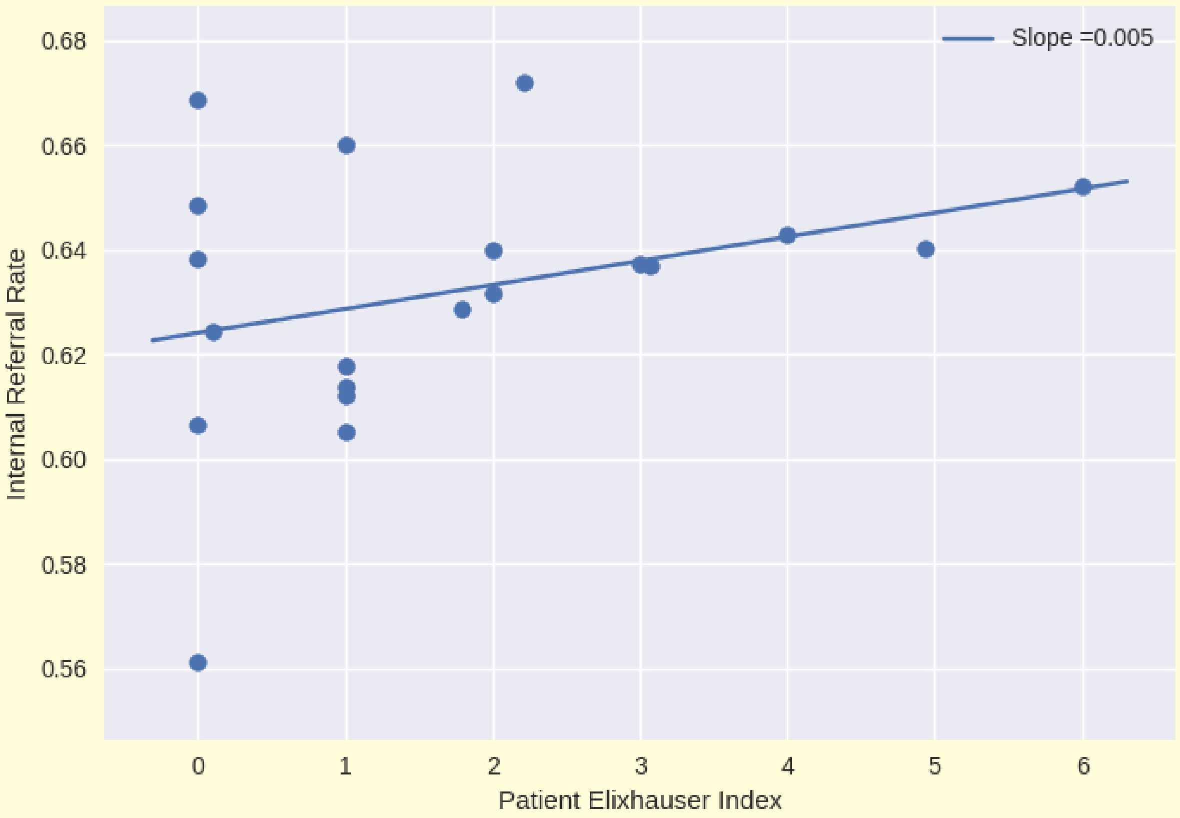
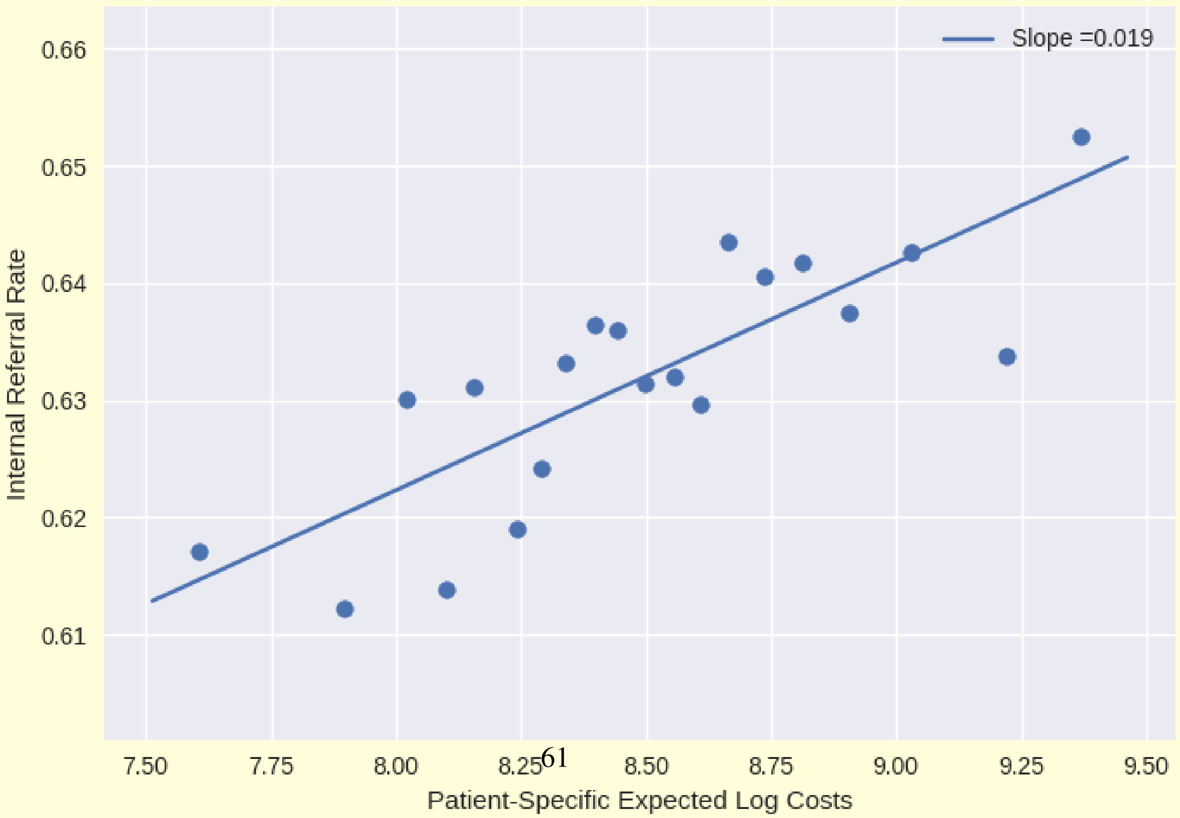


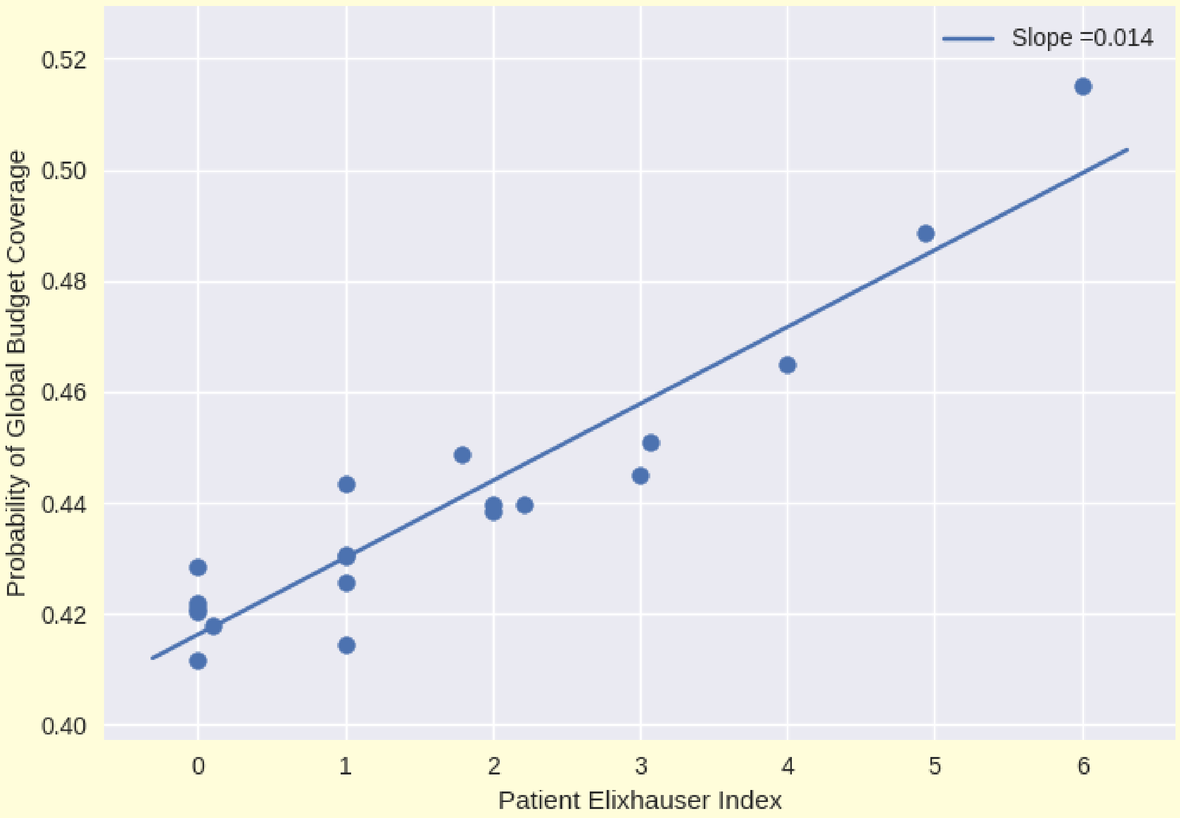
Figure 4: Scatterplot depicting the joint distribution of γ_k^{surg}, an orthopedist k's propensity to do surgery, against γ_k^{other}, his propensity to incur costs conditional on a surgery decision.

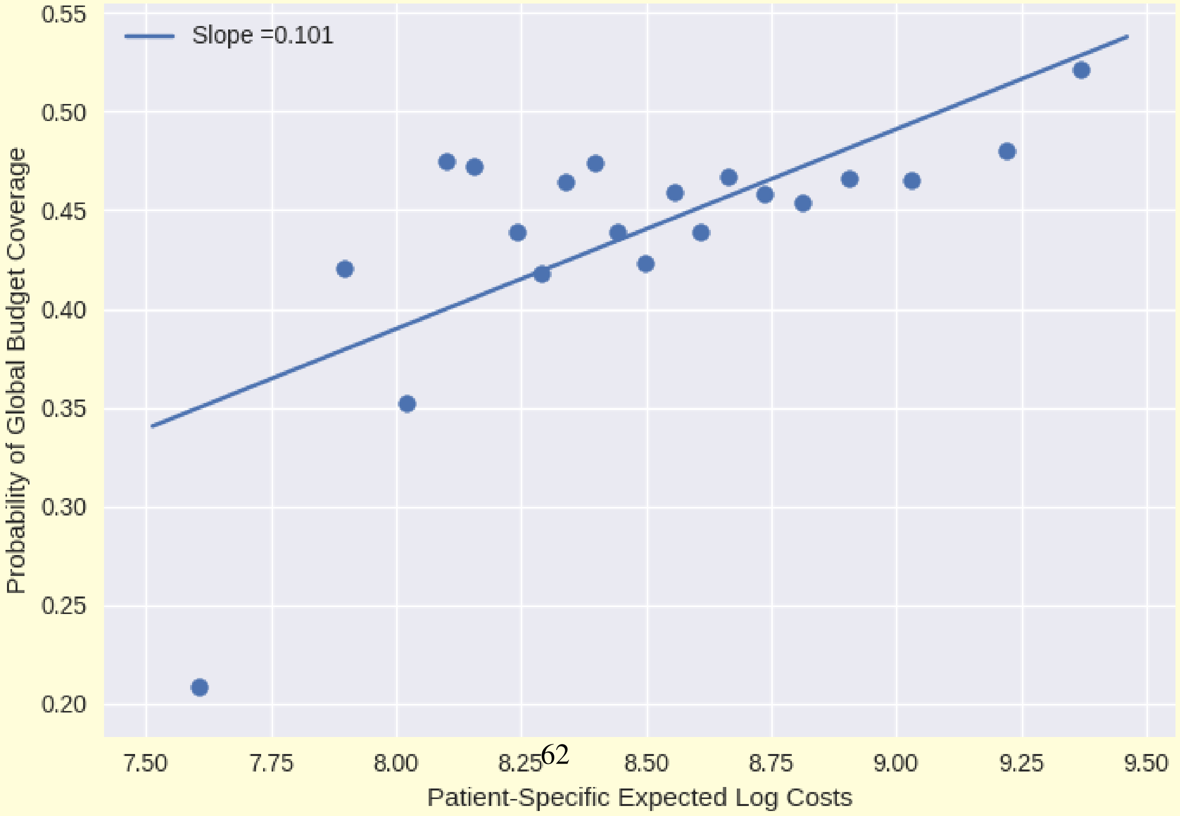

Slope =0.001
Orthopedist Cost
0.140
0.135
0.130
0.125
0.120
0.115
0.110
0.105
0
1
2
3
4
5
6
Patient Elixhauser Index


Slope =0.039
Orthopedist Cost
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
7.50
7.75
8.00
8.25
8.50
8.75
9.00
9.25
9.50
Patient-Specific Expected Log Costs


Slope =0.005
Internal Referral Rate
0.68
0.66
0.64
0.62
0.60
0.58
0.56
0
1
2
3
4
5
6
Patient Elixhauser Index


Slope =0.019
Internal Referral Rate
0.66
0.65
0.64
0.63
0.62
0.61
7.50
7.75
8.00
8.25
8.50
8.75
9.00
9.25
9.50
Patient-Specific Expected Log Costs


Slope =0.014
0.52
0.50
0.48
0.46
0.44
0.42
0.40
Probability of Global Budget Coverage
0
1
2
3
4
5
6
Patient Elixhauser Index


Slope =0.101
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
Probability of Global Budget Coverage
7.50
7.75
8.00
8.25
8.50
8.75
9.00
9.25
9.50
Patient-Specific Expected Log Costs

A Data Construction

A1 Orthopedic Surgery Procedure Codes

In Section 4.2, 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 A1, we list the 25 most common surgery codes, and the share of patients who were billed for that code.[33] 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.

A2 Elixhauser Comorbidity Index

To measure patient health status, we turn to the Elixhauser Comorbidity Index. The Index was introduced by Elixhauser et al. (1998) as a measure of patient health, for use in risk-adjusting measures of patient spending and health outcomes in inpatient 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 A2, 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 A1, 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.

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

B Empirical Bayes Shrinkage

In Section 4.1, 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 5 and as an input into our structural model in Section 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.[34]

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

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

[34] 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 6. In practice we estimate that our measurement error is relatively small, so we do not explore this more deeply.

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 A3 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 5 and 6.

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, typically billed with other surgery)	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 laminectomoy. 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 A1: A list of the 25 most commonly-billed orthopedic surgery CPT codes and the share of patients who receive them.

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 A2: A list of the 30 conditions in the Elixhauser Comorbidity Index.

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

Table A3: Estimate output from our empirical Bayes procedure.

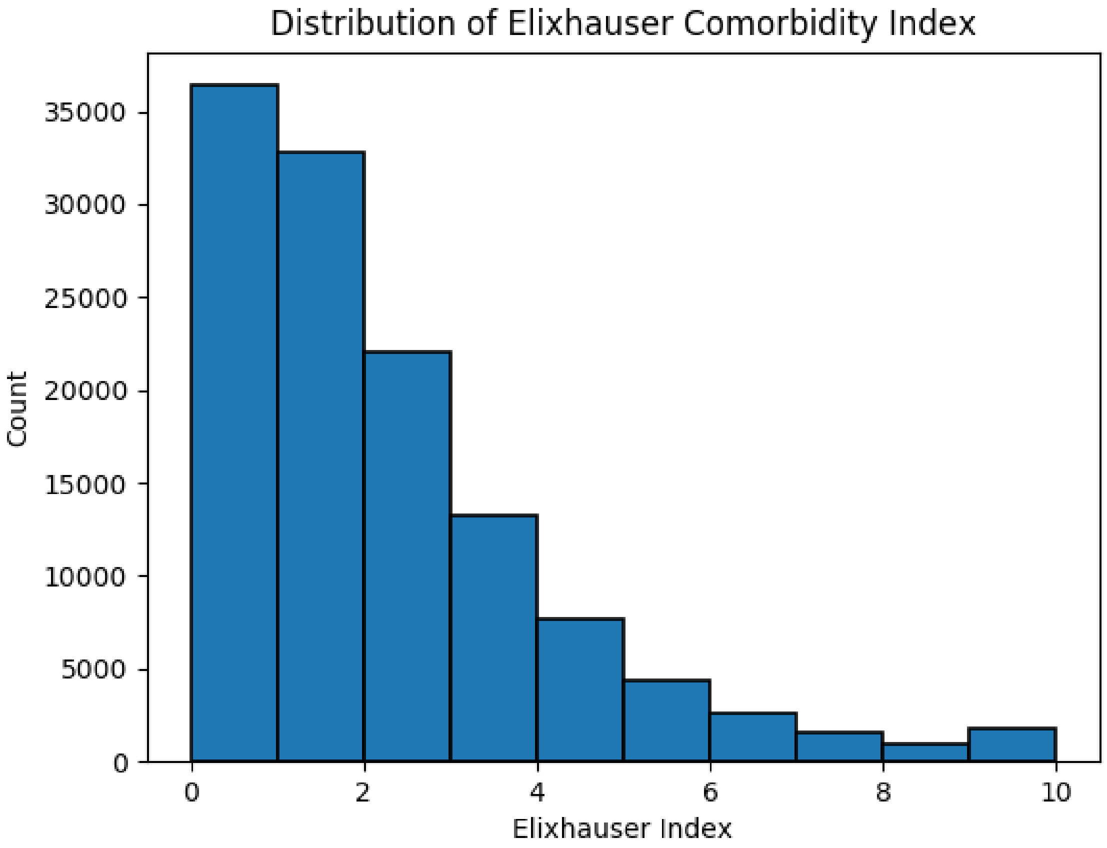


Figure A1: Counts of patient Elixhauser Comorbidity Index values in our data. The final bin includes patients with 10 or more comorbidities.