NBER WORKING PAPER SERIES

INDUCING PARTICIPATION AND PERFORMANCE USING FINANCIAL INCENTIVES: EVIDENCE FROM MEDICARE

Atul Gupta
Joseph R. Martinez
Amol S. Navathe

Working Paper 31256
http://www.nber.org/papers/w31256

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2023, revised October 2024

Previously circulated as 'Selection and Causal Effects in Voluntary Programs: Bundled Payments in Medicare.' We are grateful to Rachel Werner and the LDI analytics center for help with access to the Medicare claims data. We thank Diane Alexander, Yunan Ji, Meredith Rosenthal, Seth Richards-Shubik, Zarek Brot-Goldberg and seminar participants at the University of Pennsylvania, Harvard T. H. Chan School of Public Health, ASHEcon, Notre Dame, and the University of Chicago. Sam Cohen provided outstanding research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Inducing Participation and Performance Using Financial Incentives: Evidence from Medicare
Atul Gupta, Joseph R. Martinez, and Amol S. Navathe
NBER Working Paper No. 31256
May 2023, revised October 2024
JEL No. H00,I10

ABSTRACT

Voluntary participation is a central feature of payment reforms that are being tested in US healthcare. These programs frequently incentivize provider participation with financial rewards linked to performance, although there are questions about their efficacy in attracting participants or on the treatment effects for participants drawn by such incentives. We inform this debate by studying the introduction of a voluntary Medicare payment reform to reduce spending on hospital surgeries. We show that hospitals with a higher financial gain predicted by the payment formula are more likely to participate. We leverage quasi-experimental variation in the potential financial gain to instrument for participation and quantify the treatment effect for "complier" hospitals: those induced to participate by the incentive. Although participating hospitals are disproportionately nonprofit overall, financial compliers are disproportionately of for-profit ownership. Compliers generate greater savings than the average participant by decreasing medical services more aggressively. Selection on favorable historical trends or changes in patient mix do not explain the savings. However, we detect some worsening of care quality in complier hospitals. Overall, the results suggest that the use of financial incentives increases both the participation in and the treatment effect of this voluntary program.

Atul Gupta
Wharton Health Care Management
3641 Locust Walk, CPC 306
Philadelphia, PA 19104
and NBER
atulgup@wharton.upenn.edu

Joseph R. Martinez
The Wharton School of the
University of Pennsylvania
Philadelphia, PA 19104
joseph.ri.martinez@gmail.com

Amol S. Navathe
University of Pennsylvania
1108 Blockley Hall
423 Guardian Dr.
Philadelphia, PA 19104
amol@wharton.upenn.edu

1 Introduction

Linking provider payments to their quality and efficiency has become the cornerstone of federal healthcare policy since the passage of the Affordable Care Act (ACA). During the last decade, the Centers for Medicare and Medicaid Services (CMS) launched 54 alternate payment models, affecting contracts for more than a million providers serving 26 million patients concentrated in Medicare, the US public insurance program for the elderly (Smith 2021). Introducing performance pay helps align the incentives of providers with those of payers and patients, but can expose providers to financial risk, which has generated some resistance to these reforms. Perhaps as a result, 50 of the 54 aforementioned contracts allowed providers to opt in. This pattern is not limited to Medicare; Commercial insurers also allow voluntary participation in alternative contracts as the predominant approach (Milad et al. 2022).

Voluntary payment contracts pose several design challenges for health insurers, which are often large government programs funded by taxpayers, such as Medicare and Medicaid. Chief among them is designing the contract to attract the participation of providers who will improve performance relative to the default "fee-for-service" (FFS) contract. The consensus solution appears to be to align incentives by offering providers a share of the savings they generate under the new contract. In theory, financial incentives should attract participants who expect greater treatment gains, and therefore magnify the program's benefits. This is also the conventional wisdom among payers since it is often the primary mechanism to attract participation; it also aligns with greater feasibility than mandated participation. Even if incentives are designed well, providers may not accurately anticipate their ability to succeed under an alternative payment system, in part because success requires changing established clinical practice patterns. This challenge may nullify the theoretical advantage in attracting participants with the greatest ability to generate savings for the payer. This has led to an active policy debate on the use of financial incentives in payment reform, with opponents criticizing them as ineffective or wasteful (McWilliams 2022; Ryan et al. 2023). Surprisingly, there is limited evidence on how financial incentives drive participation in and affect the performance of voluntary programs.

This paper attempts to fill this gap by studying the Bundled Payments for Care Improvement program (BPCI), which was the largest voluntary payment program in US healthcare when it was introduced. Beginning in the fourth quarter of 2013, hospitals could enroll in BPCI for care episodes initiated by inpatient stays. We study the most prominent bundle, which was for lower extremity joint replacement surgery (LEJR), selected by more than 70% of the hospitals participating in BPCI. At the end of our sample, December 2016, approximately 300 of the 2,500 eligible hospitals had enrolled. The participating hospitals and their partners agreed to receive a prospectively set amount to cover all types of care received by the patient during LEJR surgery and the following 90 days (the "episode"). In FFS, each provider received a separate payment for each service rendered, regardless of the total amount spent. Hence, participants assumed financial risk under the new contract, with the promise of payments above FFS levels if they could reduce episode spending below the target set by CMS.

To financially succeed, hospitals would have to judiciously reduce the use of post-acute care (PAC), such as skilled nursing, and rehospitalizations, which they had no reason to restrict previously. This would involve implementing new care protocols and ensuring nurses and physicians (many of whom are not employed by the hospital) change their behavior. Some hospitals may have been well informed about their ability to implement these changes and reap financial rewards under the new contract, but this may not have been uniform.

Previous studies have typically examined the gross savings of such voluntary payment contracts using a difference-in-differences (DD) design, which compares trends between participants and nonparticipants before and after program participation.[1] Under the parallel trends assumption, the DD estimator recovers the average treatment effect on the treated (ATT). In the case of BPCI, the prior literature has estimated a 3–5% reduction in spending (Agarwal et al. 2020; Chopra et al. 2023). We also estimate a 3% decrease in spending using a DD design, with event study plots indicating no evidence of differential pre-trends for the key outcomes. The results are also robust against a battery of specification checks. This estimate offers a valuable benchmark for the treatment effect for hospitals induced to join by the financial incentive, whom we refer to as financial compliers.

Our IV approach leverages variation in the *potential* for financial gain from participation due to the interaction of the contract design of BPCI, set exogenously by CMS, and predetermined observed hospital attributes. We hypothesize that potential financial gain varies along two dimensions. First, the new contract allowed hospitals with historically higher episode spending levels to have a higher target price. Hospitals and industry consultants widely believed that facilities with higher historical spending levels also had more low-value care, or "fat", to easily cut (Navathe et al. 2018; Berlin et al. 2021). Hence, these hospitals may have believed that they had more financial upside under BPCI. Second, as we will discuss, adapting to BPCI largely involved incurring fixed costs. Hence, hospitals with a higher historical (and thus future expected) volume of LEJR surgeries enjoyed lower per-episode participation costs and higher aggregate payoffs.

To formalize this variation, we define the instrument as an indicator for hospitals with both above-median baseline mean episode spending and above-median baseline aggregate revenue within Medicare orthopedic procedures prior to program initiation. We interpret the instrument as quasi-randomly assigning hospitals to low or high potential for financial reward. Across a range of specifications, the instrument strongly predicts participation in BPCI, confirming that financial incentives drive participation in the program. However, hospitals likely also participated for other reasons beyond financial gain (e.g., learning) during this period, since Medicare had signaled large shifts to performance pay. For example, several hospital administrators professed an intrinsic motivation to improve performance as their key motivation to join BPCI (Lewin Group 2016).

Identification relies on the exclusion restriction that hospitals designated as having

1. Net savings also incorporate the incentive payments made by the payer to the provider.

high potential for financial gain would have progressed on similar trends in the absence of BPCI as hospitals designated as low potential. This is untestable, but we present extensive evidence consistent with this assumption. We report reduced form dynamic effects not only for all outcomes, but also for relevant correlates to assess differential pretrends (Goldsmith-Pinkham, Sorkin, and Swift 2020). We further support the validity of the instrument using a falsification test in the spirit of Angrist, Lavy, and Schlosser (2010). The instrument is intended to specifically predict financial attractiveness of an orthopedic service bundle and should not predict participation in non-orthopedic bundles, barring spurious correlations through hospital-level factors (e.g., contemporaneous changes in management quality). Reassuringly, the instrument does not predict participation or changes in spending in the five largest non-orthopedic bundles offered through BPCI.

The IV estimate implies an 11% reduction in episode spending, which is more than three times the corresponding DD coefficient. This implies a gross saving of approximately $750 for the average participant versus $2,750 for the average complier. Both the DD and IV estimates imply that the main savings mechanism is a reduction in the probability of using PAC. As discussed earlier, the new contract incentivized participants to avoid care outside the hospital as a source of savings. This channel can explain about two thirds of the decline in episode spending in both the average participant and the average complier. The key difference is that compliers reduce PAC much more aggressively. For example, DD implies a decrease in any use of PAC of 2% compared to the baseline, while IV estimates indicate a decrease of 15%. We also detect a novel savings mechanism, an increase in the proportion of short stays for the index surgery, which are reimbursed at lower rates.[2] Again, financial compliers pursue this approach more aggressively than the average participant.

The large decrease in care services for patients of financial compliers raises concerns about whether they changed the composition of their patients. We extensively test for differential changes in the patient mix based on demographics and the history of comorbidities and healthcare use. We also test for changes in unobserved dimensions by studying changes in the distance of the patient to the hospital. However, both DD and IV designs estimate small and statistically insignificant changes in observed and unobserved risk measures. This pattern is consistent with the lack of patient selection documented in previous studies of voluntary payment reforms (Finkelstein et al. 2018; McWilliams et al. 2020).

The large decrease in spending achieved by financial compliers also raises the question of whether they provide a lower quality of care. We examine this issue closely and find mixed evidence. We estimate small and statistically insignificant effects on the quality metrics monitored by CMS as part of the program's governance (e.g., short-term mortality and readmissions). However, we also precisely estimate a relatively large increase in the probability of revision surgery, an unambiguously bad outcome that was not monitored by CMS. We calculate that the average savings at financial compliers remain large after

2. Hospitals had the option to discharge patients early under FFS too, but would lose revenue if they did so. They were also unable to discharge patients to nursing homes until after a minimum 3 day stay. Under BPCI, they would no longer lose this revenue as they recoup it as episode savings, and CMS waived the rule regarding the minimum of 3 days for discharge to nursing homes for BPCI participants.

adjusting for the expected increase in revision surgeries, although this does not incorporate non-financial costs to the patient.

This paper contributes to different strands of the literature. Our main contribution is to the debate on the use of financial incentives to align the performance of healthcare providers with social goals. Previous studies have shown that sophisticated providers accurately assess the expected effects of such contracts and respond accordingly (Norton et al. 2018; Gupta 2021; McWilliams et al. 2020). However, inefficient contract design, gaming by providers, and the potential for unintended consequences have led to questions about the wisdom of using financial incentives (Alexander 2020; Einav et al. 2022; Darden, Mccarthy, and Barrette 2023). Overall, the results in this paper are reassuring. We show that participating hospitals are not disproportionately predicted to receive a payoff under BPCI just by maintaining historical performance. This is intuitive since the BPCI contract linked payoffs to changes in spending relative to a hospital's baseline spending level rather than to a regional or national benchmark. Hence, while inefficient selection is an important concern with voluntary programs, it can be mitigated by appropriate contract design. We confirm that a nontrivial fraction of participants are induced by financial incentives and show that financial compliers achieve much greater gross savings than the average participant, without accompanying patient selection. Hence, financial incentives enhance the effect of the program, consistent with sorting on treatment gains, an argument in favor of voluntary programs. In contrast, our results also imply that providers whose participation is induced by other factors (e.g., learning) generate very little savings. More work is needed to understand how payers can support these providers in making appropriate participation decisions or motivate them to improve their performance.

Our novel IV approach offers a template to studying other voluntary programs. Previous studies have acknowledged concerns due to endogenous provider participation, but have relied on a DD design, with or without the use of matching and provider fixed effects to mitigate selection (Werner et al. 2011; Dummit et al. 2016; Navathe et al. 2017; McWilliams et al. 2020). Our IV design leverages the variation generated by the payment formula as a source of identification. Such idiosyncrasies are common in contracts, and therefore our approach can be applied to other voluntary programs to recover the treatment effects for a highly policy-relevant subgroup of participants.

Finally, we extend the literature on bundled payment contracts in healthcare. Previous studies have typically focused on average treatment effects and found modest savings (Agarwal et al. 2020). We show that financial compliers achieve greater savings by decreasing care more aggressively and identify a previously undescribed mechanism of savings by shortening the index hospital stay. To our knowledge, previous studies of bundled payment contracts have not detected an effect on quality of care (Finkelstein et al. 2018; Navathe et al. 2020). Although we reach a similar conclusion for the average participant, our results for compliers suggest that hospitals trade off patient health to achieve a high level of savings. This suggests a ceiling on the savings potential of such contracts.

The remainder of the paper is organized as follows. Section 2 provides the necessary background information on the new contract, hospital incentives, and evidence from previous studies. Section 3 describes the data sources and presents descriptive evidence. Section 4 describes the empirical strategy, including the IV approach. Section 5 presents the main results, including falsification and sensitivity checks. Section 6 concludes.

2 Background

2.1 Alternative payment contracts

The ACA was enacted in 2010 and introduced payment reforms to better align provider incentives with Medicare than under the prevailing FFS system. This included mandatory performance pay programs that incentivized hospitals to improve quality of care (Norton et al. 2018). Recognizing the need to test the effects of different policy designs before broad implementation, the ACA also tasked CMS with launching an innovation center to evaluate multiple alternative payment contracts and develop policy recommendations. The resulting Centers for Medicare and Medicaid Innovation (CMMI) launched more than 50 different initiatives offering alternate payment contracts to providers for services provided to Medicare patients (MedPAC 2021). This paper focuses on Bundled Payments for Care Improvement (BPCI), one of the largest alternate contract initiatives.

2.2 BPCI

In 2013, CMMI launched the voluntary BPCI program, its first national bundled payment contract. Between October 2013 and September 2018, Medicare tested bundled payments for medical and surgical episodes through BPCI in several different types of episodes. Lower extremity joint replacement surgery (LEJR), which includes major hip and knee replacement surgeries, was by far the most frequently selected clinical episode in terms of the number of participating hospitals. About 300 hospitals participated in the LEJR bundle. The next largest was congestive heart failure, with about 170 hospitals.[3] Therefore, in this article we focus on LEJR surgeries. The program bundles payment for all associated medical care regardless of site of care, starting with the index hospitalization and including up to 90 days of post-discharge care. CMS allowed staggered entry into BPCI and after a gradual start in 2014, there was rapid entry in 2015. Similarly, hospitals could also exit the program at their discretion. Figure 1 Panel (a) shows that 18% of all LEJR episodes in 2016 were performed by hospitals participating in BPCI.

From the perspective of the payer, Medicare, the goal of a bundled payment is to encourage hospitals to better coordinate with other providers and eliminate low-value care to generate savings. It is conceptually similar to the introduction of prospective payments

3. The difference was even greater in terms of patients – LEJR patient episodes were more than three times the corresponding number for heart failure (Lewin Group 2018a). CMS allowed hospitals to choose one of four different contract types, which they called "models." We study Model 2, which represented more than 85% of all patient episodes in BPCI (Lewin Group 2017). For completeness, we retain the 26 hospitals that participated continuously in Models 1 and 4 in our comparison group. They contribute about 1% of hospitals in the comparison group.

for hospitals in 1983 which bundled payment for all inpatient services provided in a single hospital stay (Acemoglu and Finkelstein 2008). However, BPCI is more ambitious since it assumes that hospitals can coordinate with providers from other firms involved in patient care and optimize care costs over a *much* longer duration. The average LEJR surgery involves a stay of 3.8 days, while BPCI episodes also include the next 90 days.

CMS changed some regulations to facilitate spending reductions, altering established clinical protocols. We discuss only two here in the interest of brevity. Hospitals were allowed to transmit incentives to physicians and post-acute care (PAC) providers by sharing savings with them. 66% of the participants submitted plans to exercise this "gainsharing" option and about $144 million were distributed in gainsharing through 2016 across all bundles.[4] Medicare also waived the requirement of a minimum stay of 3 days to discharge patients to a skilled nursing facility (SNF). The goal was to allow hospitals to discharge patients earlier (and therefore bill a lower amount) if they determined that it was medically appropriate to do so (Lewin Group 2018b).

Payment formula

The target episode spending level or "price" $\hat{p}_{ht}$ for a participant h in any year t is calculated in three steps. First, CMS computes the mean LEJR episode spending in 2009–2012 for the hospital. An episode includes all care spending during index surgery and the 90 days following discharge. Then it adjusts this amount by the national trend in mean LEJR spending for Medicare patients between 2012 and t. For example, the national mean episode spending level decreased by 2.5% between 2011-12 and 2012-13. This rate would be applied to calculate the predicted spending level in 2013, and so on for each subsequent year. Finally, to ensure that CMS captures a fixed share of the savings, it deducts 2% of the trend-adjusted amount to arrive at the final target price. This last step implies that a hospital must perform better than the national trend on average to limit spending below the target price. An illustrative example may help to fix ideas. Consider a hospital with a mean episode spending level in June 2012 of $100. Assuming a constant national annual spending trend of -2.5% through 2016, the desired spending level for 2016 would be $90.4. Applying the additional 2% discount further leads to $88.6. To achieve this target price, the hospital must reduce annual spending by an average of 3% or more between 2012 and 2016. Note that the program's target price is not determined using a comparison group of non-participating hospitals, as would be done in a quasi-experimental research design.

The payoff for episode i in year t at a participating hospital h, $gain_{iht}$, is simply:

$$gain_{iht} = \hat{p}_{ht} - b_{iht}, \tag{1}$$

4. Hospitals had flexibility in structuring the gainsharing payments. In interviews with the Lewin group, hospitals indicated that they used gainsharing payments to focus physician effort on reducing readmissions or changing discharge destination. They also offered higher volumes of referral to SNFs to influence their behavior, particularly duration of stay. Hospitals could also provide incentives to their patients in the form of in-kind items or free additional services. The most common patient incentive was free transportation to outpatient therapy or rehabilitation and medication management tools.

where b_{iht} is the amount billed for patient i during the 90-day episode.

Implications

The formula indicates that a hospital will be financially better off under BPCI than under FFS if it can limit mean episode spending, $\bar{b}_{ht}$, below the target price, $\hat{p}_{ht}$. The objective of the program is for hospitals to change their care protocols and change the trajectory of their spending trend. Hospitals have the incentive to earn as revenue the amount saved from what was expected to be billed by other providers, such as inpatient rehabilitation, skilled nursing, and home health care. In the pre-BPCI period, inpatient care accounted for only about half of total episode spending (Dummit et al. 2016). Hence, hospitals can reap substantial incremental revenue under BPCI if they reduce post-discharge care.

One may be concerned that BPCI attracts hospitals that expect to be financially rewarded without changing their behavior. This is socially inefficient and represents the "selection on levels" problem documented in the case of another voluntary program by Einav et al. (2022). In the case of BPCI, since the target price is linked to a hospital's own past performance, it should not attract hospitals with low baseline spending levels. Instead, inefficient selection would apply to hospitals that already enjoyed a greater decreasing trend in spending than the national average. Since BPCI sets the target price using the national, rather than hospital-specific, mean spending trend, it creates an arbitrage opportunity. Continuing the illustrative example above, consider a hospital that had already demonstrated an annual spending decrease of, say, 3.5% in the years leading up to BPCI. Such hospitals may be tempted to join the program under the assumption that they could receive a financial reward by continuing to do "business as usual." If maintaining a better than average trend decline is, in fact, continuously possible without making additional efforts, these hospitals would be financially rewarded for no effort. We formally investigate this possibility in Section 3.3 and find little evidence of selection on trends in BPCI. Hospitals with worse than average historic spending trends were just as likely to participate in BPCI as those with better historic trends.

2.3 Perspectives of administrators

CMS contractors conducted detailed interviews with hospital administrators, providing some insight into their decision making. The most important stated reason for participation was that BPCI provided hospitals an opportunity to experiment with new payment contracts. This seems to be driven partly by long-term financial considerations, since they believed these contracts would eventually dominate the payment landscape. There was also some intrinsic motivation to improve. Some participants mentioned that this was in line with their status as "leaders of innovation." In their report, Lewin Group (2016) note, "More than half of the respondents we interviewed claimed that the learning opportunities were the main reason for their interest in BPCI."

The next factor in terms of importance appears to be the direct financial incentive. The participants were confident in their ability to engage hospital staff and physicians to change practice patterns and succeed. This was particularly true for LEJR surgeries, which they viewed as a relatively predictable procedure with lower financial risk. Hospitals often partnered with consultants who advised them on the entry decision and sometimes provided ongoing analytic support after entry. This process suggests that hospitals expecting greater financial rewards were more likely to participate.[5]

Participating hospitals indicated three main cost savings strategies in the interviews. First, they would reduce the use of institutional PAC by changing their discharge guidelines, engaging physicians, and collaborating with PAC partners. Second, they would reduce readmissions through improved analytics and care coordination. Third, they would reduce the procurement costs for implants used in surgeries by standardizing and consolidating vendors. This mechanism has subsequently been empirically validated (Navathe et al. 2017), but notably does not change the spending for the Medicare program. They claimed that these strategies entailed substantial new fixed costs and administrative overhead (Lewin Group 2016). For example, they hired new analysts, nurses, and case managers for patients in BPCI episodes. They dedicated considerable time to training physicians and staff for new protocols, patient education, and discharge planning. Larger hospitals could absorb these additional costs more easily.

2.4 Prior evidence

Given the scale and importance of this program in determining future Medicare payment models, several studies have attempted to estimate the causal effects of participation in BPCI on episode spending and patient health. Previous studies have typically used matching methods to identify suitable comparison hospitals and have estimated DD models to quantify the causal effects on spending and patient health (Dummit et al. 2016; Lewin Group 2018b; Agarwal et al. 2020; Navathe et al. 2020). These studies reported 3–5% episode savings without decrements in quality. To our knowledge, the endogenous participation decision has not been explicitly addressed in previous studies. Hence, while the DD estimate can credibly recover the average treatment effect across all participants (ATT) assuming parallel trends hold, there is no guidance on the treatment effect for hospitals that were induced to join by the financial incentive.

CMS has also piloted other bundled payment contracts. The most well-known among these is Comprehensive Care for Joint Replacement (CJR), a mandatory program in which 67 markets were randomized into bundled payments, while 104 markets were maintained as randomized controls. Early evaluations of CJR also report a causal spending reduction of 2–3%, comparable to that reported for BPCI (Barnett et al. 2019; Einav et al. 2022). However, due to differences in the payment formula, the incentives for participation and improvement in CJR were completely different from those in BPCI. Specifically, the target price in CJR was largely determined by the average episode spending level in the market.[6]

5. We interpret these details as hospital claims, not facts. For example, it is possible that hospital executives were reluctant to admit that financial incentive was the key draw.

6. CMS stipulated a phase-in period to set the target price. Initially, it was set based on a combination of the hospital's past performance and the market average. In steady state, the market average would serve as

Hence, hospitals with baseline spending below that of their market average could receive payments without any further reductions, an inefficiency noted by Einav et al. (2022). They show that when, after 2 years, hospitals in some markets were allowed to exit the program, hospitals with low baseline spending levels were more likely to continue. Although the payment formula in BPCI does not encourage selection on spending levels, one may be concerned about selection on historic trends, as discussed previously.

3 Data and descriptive evidence

3.1 Data sources and sample construction

Our primary data source is the universe of Medicare FFS facility claims accessed through the Medicare Provider Analysis and Review (MedPAR) files over 2006–16. For all hospital and PAC stays involving FFS patients (short-term and long-term acute care, inpatient rehabilitation, skilled nursing, hospice, and home health), we observe demographic information about the patient: the reason for the care, start and end dates, a vector of diagnosis and procedure codes, and the discharge destination. We link the MedPAR files with the Master Beneficiary Summary File (MBSF) using the beneficiary's unique identifier to observe enrollment information, additional demographics, and death. Thus, we comprehensively record institutional health care use for the 90-day LEJR episode. A limitation of our data is that we do not observe the use of outpatient care and durable medical equipment for these patients. However, this is only a small component of total episode spending. Based on evidence from previous studies, these dimensions of care account for approximately 11% of total expenditure on 90-day episodes.[7] Another limitation is that the MedPAR does not record the physician ID performing the procedure(s) during the stay, so we cannot test whether participation in BPCI leads to changes in the physicians performing LEJR.

We supplement the administrative claims data with information on hospitals (e.g., number of beds, total inpatient admissions, etc.) from the American Hospital Association (AHA) annual survey files and market-level information from the Dartmouth Atlas and the US Census. Finally, we use data from CMS Hospital Compare on hospital quality and penalties under various performance pay programs.

We identify LEJR index cases (MS-DRGs 469 and 470) following the same approach used by CMS when it calculates episode spending and target prices for hospitals under BPCI. Routine cases are billed under MS-DRG 470, while patients with complications or comorbidities are billed under MS-DRG 469. About 95% of all cases are billed under 470.[8]

the benchmark.

7. For MS-DRG 470, Navathe et al. (2018) Exhibit 4 reports that BPCI hospitals spent approximately $2,400 on physician fees, outpatient facilities, and durable medical equipment not observable in our data. The authors report $22,100 in total spending for these episodes. Hence, these components account for 11% of the total expenditure.

8. Our estimates remain undisturbed if we limit the analysis to 470 episodes only. Before 2008, CMS used a different version of DRG classification in which all LEJR cases were billed under a single DRG (544). Index cases are defined such that the patient did not have LEJR surgery in the previous 90 days. To ensure we have a 90-day follow-up period for all index cases, we limit index cases to those discharged by September 30, 2016.

We exclude about 600 hospitals with fewer than 25 index LEJR cases during the baseline period 2006–2007 from the analysis. These account for less than 5% of all LEJR cases. Our final sample has about 2,500 hospitals, of which about 300 participated in BPCI during this period. Section A.1 provides more details on our sample construction approach.

3.2 Descriptive evidence

Table 1 presents descriptive statistics from our analysis sample. Panels A,B, and C, respectively, present market-, hospital-, and episode-level statistics for 2009–13, before hospitals enrolled in the program. We present statistics separately for nonparticipant and participant hospitals in Columns 1 and 2, respectively. All spending values are inflation adjusted and expressed in 2016 dollars. Panel A shows that participating and nonparticipant hospitals are located in markets with similar Medicare Advantage (MA) penetration. We compute this value for each hospital-year as the weighted average MA share across the hospital service areas (HSA) that contribute patients to the hospital. The weights are the HSA's share of the hospital's total LEJR patients in 2007, held constant.[9] However, participating hospitals are located in more expensive markets, evidenced by a 3% higher geographic adjustment factor (GAF) and a nearly 10% higher average hourly wage. CMS uses both as inputs to adjust the prospective payment rate for LEJR surgery across hospitals.

We note two patterns about participants based on the information reported in Panel B. First, participating hospitals are substantially larger (40% more beds) and obtained nearly $1,200 more Medicare revenue per bed from orthopedic procedures. These differences may reflect the fixed cost of participation that makes entry more favorable for hospitals with a larger orthopedic practice. Second, participants are 40% more likely to be teaching hospitals, suggesting that administrators at academic facilities were more enthusiastic about adopting the new contract.

Table 1 Panel C shows that the sample contains approximately 2.5 million episodes throughout the period, with approximately 18% performed in hospitals that participated in BPCI. Participating hospitals had a higher intensity of treatment prior to the program – an episode cost Medicare about $2,000 more (9%) on average in participating hospitals ($25,300 vs. $23,300). Hence, participants were not selected on a favorable baseline spending level. This is an encouraging sign from the perspective of program design, since a goal of such payment reforms is to attract potentially inefficient providers and motivate them to improve their performance (McWilliams et al. 2020).

We examine the difference in episode spending between participants and nonparticipants more granularly to assess whether inefficiencies could play a major role. About 40% of the gap ($800) is explained by the higher billing for the LEJR surgery at participating hospitals. The higher LEJR payment is mostly explained by differences in market-level cost adjustment factors indicated in Panel A and additional payments for medical educa-

9. The Dartmouth Atlas group defined HSAs as contiguous zip codes whose residents receive most of their inpatient care from hospitals in the same area. There are approximately 3,400 HSAs in the US, slightly more than the number of counties.

tion and capital costs (not reported here). Both groups of hospitals had a similar share of LEJR patients billed under the more intense DRG; therefore, differential patient complexity is not an explanatory factor.

However, the remaining difference in episode spending is largely explained by the higher readmission rates and use of PAC by patients at participating hospitals. For example, participating hospitals had a 1.2 percentage point (11%) higher readmission rate and a 5 percentage point (6%) higher use of PAC at baseline. Patients at participating hospitals were more likely to use the three forms of PAC we consider: inpatient rehabilitation (IRF), skilled nursing (SNF), and home health care (HH). This difference in post-discharge utilization cannot be explained by differences in patient complexity. Participating and non-participating hospitals appear to have treated patients at a similar risk of PAC use, which we proxy by constructing an index of propensity for SNF use.[10] In results not reported here, we find similar patterns using indexes constructed based on other forms of PAC. Overall, these patterns suggest that a large fraction of the higher spending of participating hospitals at baseline was due to inefficient care use.

Aggregate unadjusted trends in medical spending and utilization of care by Medicare LEJR patients suggest that BPCI may have affected spending. Figure 1 Panel (b) presents the trend in mean LEJR episode spending for Medicare FFS patients during 2009–2016, expressed in 2016 dollars. In 2013, the mean expenditure on episodes was around $23,000 and declined to about $21,000 by 2016. Similar observations can be made on the length of stay for the index surgery (Panel c) and PAC use (Panel d). However, with the exception of the trend for PAC use, which shows a noticeable acceleration in decline after 2014, these measures were already declining in a similar fashion prior to 2014. Hence, aggregate trends alone are not conclusive about the causal effect of BPCI.

3.3 Historic spending trend and participation

A concern with voluntary programs is that they attract and financially reward participants without any change in their behavior. As described in Section 2.2, BPCI set the target price for hospitals in a way that left open the possibility of such arbitrage. In this section, we investigate to what extent such selection on trends occurred in BPCI. Specifically, we test whether hospitals with a predicted episode spending value lower than the corresponding BPCI target price are more likely to participate. We predict a hospital-specific episode spending value and BPCI target price for each hospital in 2016. In the interest of brevity, we briefly provide intuition here and leave the details for Section A.2. Each hospital's predicted episode spending value in 2016, $\bar{b}_{h,2016}$, is calculated by projecting forward its mean episode spending in 2013 using its observed historic spending trend between 2010 and 2013. In contrast, the target price for 2016, $\hat{p}_{h,2016}$, is calculated by projecting its mean episode spending in 2013 using the national average spending trend between 2010 and 2013 along with the 2% discount imposed by BPCI. We calculate the expected gain per

10. We use 2008 data to estimate a linear model predicting the probability of having a SNF stay in the 90 days following discharge on a rich vector of patient risk covariates. We apply the coefficients of this model to the data from 2009–2013 to generate predicted SNF probabilities for each patient.

episode for hospital h as

$$\bar{gain}_{h,2016} = \hat{p}_{h,2016} - \bar{b}_{h,2016}.$$

Hospitals with positive values of expected gain have the advantage that they can theoretically expect to receive a payoff under BPCI if they maintain their historic levels of spending reduction. We test whether the expected gain is positively correlated with participation. Figure A.1 Panel (a) presents a bar graph of the participation share by quartile of expected gain. The dashed red line indicates the average share of participation in the sample, approximately 13%. We also present the mean expected gain in each quartile, which ranges from $ -6,116 in the bottom quartile to $2,747 in the top quartile. The plot shows that the participation rate ranges between 12% and 14% in all quartiles of expected gain and is not greater at higher values of expected gain. We formally test the correlation by regressing quartile indicators on an indicator of participation in BPCI. These coefficients are individually and jointly statistically insignificant. The p-value from the joint test of significance is 0.80. We get qualitatively similar results when we test the correlation between expected gain and participation in 2015 instead of 2016. Hence, the evidence does not support the selection on past trends in BPCI.

This result may appear to contradict the selection on spending levels observed in CJR when hospitals were given the choice to continue or leave the program in 2018. We believe that the difference in selection patterns is largely explained by the differences in contract design between the two programs, which made inefficient selection less likely in BPCI. In CJR, hospitals with mean episode spending below the average spending level in their market were, to a first order, likely to receive a positive payoff without additional effort. However, in BPCI, since payments are not linked to a market-level benchmark, arbitrage on spending levels is not a relevant margin. The selection on trends requires hospitals to maintain their historic trend of reducing spending under the new contract. Since spending levels are more highly correlated over time than spending trends, it is easier to predict future spending levels and sort on them (McWilliams et al. 2020). In addition, it seems unlikely that hospitals can reduce spending at the same rate every year off a lower base without making additional efforts. In fact, such ratcheting of targets is a common criticism of compensation schemes that rely on past performance to set future targets (Weitzman 1980).

4 Empirical strategy

4.1 Average treatment effect on the treated

To obtain estimates of the ATT, we follow previous studies and implement a difference-in-differences (DD) research design. We compare the trends of hospitals participating in BPCI with those of nonparticipants. Although hospitals joined BPCI in a staggered fashion, we ignore the temporal variation in participation status and estimate an average effect assuming treatment starts in 2014 for all participants. Once a hospital enrolls in BPCI, we assume that it participates through the end of our sample, ignoring early exit. The DD estimate should therefore be interpreted as the average effect of ever participating in BPCI. This simplification allows us to abstract away from staggered treatment and will tend to bias downward our estimated treatment effects, thus producing more conservative estimates.[11]

We estimate the following model:

$$Y_{ihmt} = \alpha_h + \alpha_{mt} + X'_{iht}\,\delta + \beta_{dd}\,d_h \cdot \mathbf{1}(t \geq 2014)_t + \epsilon_{ihmt}, \tag{2}$$

where i, h, m and t denote patient, hospital, market, and year, respectively. The model includes fixed effects by hospital and market-year. The use of market-year fixed effects allows markets to trend differently and is more flexible than conventional two-way fixed effect models. We use hospital referral region (HRR) to define hospital markets. This is a standard market definition developed by the Dartmouth Atlas group to delineate contiguous zipcodes where residents receive most of their hospital care within the market. There are approximately 300 HRRs in the US and, in our sample, each market contains about 8 hospitals on average.

The model also includes a vector of covariates, X_{iht}, which adjusts for differences in patient mix in our preferred model. Patient controls are demographics (sex, race, age in 5 year bands), Elixhauser comorbidities, and indicators for hip fracture and MS-DRG 469. In some specifications, we also include time-varying hospital controls: beds, total admissions, Medicare and Medicaid proportions of admissions, FTEs, Medicare orthopedic share of admissions, ACO participation, and hospital share of total admissions in the HRR. d_h is a static indicator of participation in BPCI. The coefficient of interest, β_{dd}, is on the interaction of d_h and the indicator for the post-BPCI period. ϵ_{ihmt} denotes unobserved factors.

Under the parallel trends assumption, β_{dd} recovers the ATT of ever participating in BPCI. This assumption is untestable, but this design has previously been validated and used by other studies (e.g., Navathe et al. 2017; Navathe et al. 2020). We present event study plots for key outcomes of interest to help assess the plausibility of the assumption in our sample. The event studies plot the dynamic coefficients, γ_s, obtained by estimating Equation 3 below.

$$Y_{ihmt} = \alpha_h + \alpha_{mt} + X'_{iht}\,\delta + \sum_{s \neq 2013}^{2016} \gamma_s\,d_h \cdot \mathbf{1}(t = s)_t + \epsilon_{ihmt}. \tag{3}$$

Reassuringly, the event studies indicate that participants and nonparticipants trended in parallel prior to BPCI. Although we estimate the DD coefficient for all outcomes of interest, they primarily serve to provide a benchmark for the IV estimates, our main empirical

11. The patient-weighted average duration of participation across hospitals is 1.6 years out of a maximum possible 2.75 years in the sample. 58/302 hospitals (nearly 20%) left BPCI before the end of our sample period. However, only 5% of the "treated" patient episodes in the post-treatment period are at participating hospitals after they exit. Since all our estimates are patient-weighted, early exit is not a major empirical concern.

object.

4.2 Treatment effect for financial compliers

The focus of this study is to quantify the causal effect of BPCI participation in hospitals that were induced to join by the financial incentives offered. To do so, we use an instrumental variable approach, described below.

4.2.1 Instrumental variable

The design exploits the idiosyncratic choices made by CMS in setting the target price, which generates plausibly exogenous variation across hospitals in the ex-ante potential for financial gain from participation. We hypothesize that the potential for gain varies along two dimensions.

The conventional wisdom in the hospital industry was that reducing episode spending required investments in additional nurses, case managers, and data analysts, improving the functionality of electronic health records to better coordinate with external providers, redesigning clinical protocols, and training staff on the new protocols (Lewin Group 2016). Since these actions, which we interpret as the "technology" of spending reduction, largely involve incurring fixed costs, scale economies come into play, making it easier for larger hospitals or those with greater orthopedic volume to justify incurring these. Hospitals with a larger LEJR practice at baseline also expect a higher aggregate benefit from adopting bundled payments. We therefore leverage the variation in Medicare FFS LEJR revenue across hospitals, measured in a baseline period prior to BPCI.

CMS chose a contract design such that each hospital's target price was determined by its episode spending level just prior to BPCI. Several alternative target setting options were available. For example, in the CJR program, CMS set the target price based on market level, not the hospital's own, historic mean episode spending. In other reforms, such as the Hospital Readmissions Reduction Program, CMS has set the benchmark equal to the national average. Depending on the chosen benchmark, different groups of hospitals would find it easier or more difficult to succeed. Higher mean episode spending in a hospital is typically driven by high rates of PAC use and readmissions among its patients. Assuming that the spending reduction technology described above has diminishing returns, hospitals with higher pre-BPCI spending levels had the potential to obtain larger reductions per unit cost. This suggests that we leverage the variation in the baseline mean post-discharge spending for LEJR episodes across hospitals.

The aggregate potential for financial gain increases in both dimensions, individually and in their interaction. We hypothesize that hospitals with higher values in both dimensions expect greater financial rewards and are more likely to participate in BPCI.

We take several steps to reduce the measurement error and the influence of transient noise when we construct the baseline variables. We use data from 2007 to compute the baseline values. This was much before the passage of the ACA in 2010, let alone the launch of BPCI. This avoids predicting participation on the basis of values that happened to be high just as BPCI was launched. By allowing a 7 year interim period, we hope to use the persistent component of these attributes when predicting participation. Second, we pool data across all orthopedic procedures rather than limiting ourselves to LEJR surgeries alone.[12] This reduces noise in the measure, particularly for smaller hospitals. Third, to further expunge omitted factors, we use predicted and not observed values of orthopedic revenue and mean episode spending. Section A.3 describes the prediction step.

Figure 2 presents binned scatter plots showing the variation in mean participation rate in decile bins of baseline Medicare orthopedic revenue (Panel a) and post-discharge spending (Panel b). The plots also present the slope coefficient obtained using a linear regression on the underlying hospital-level data. Panel (a) implies that a hospital with \$1.5 million higher orthopedic revenues at baseline (approximately 20%) has a 1.5 percentage point greater probability of participation, approximately 13% of the mean participation rate. Similarly, Panel (b) implies that a hospital with \$2,000 higher post-discharge spending at baseline (20%) has a 4.6 percentage point higher probability of participation, approximately 38% of the mean. Hence, participation increases in both dimensions. Panel (c) presents the mean participation rate in percentile bins obtained by interacting the decile bins of both dimensions. The figure shows that participation is disproportionately higher among hospitals that have above-median values on *both* dimensions. We incorporate this pattern in our empirical approach. Our instrument is an indicator, z_h, for hospitals with higher than median values of both variables. For ease of exposition, we refer to hospitals with $z_h = 1$ as having greater potential financial gain. Note that this approach does not impose in any way that hospitals with greater potential must also obtain higher savings than hospitals assigned low potential.

Using an indicator circumvents the need for parametric assumptions to convert the baseline attributes into dollars of expected gain. Since the instrument collapses continuous variation in the two variables into a binary indicator, it is not sensitive to using predicted or observed values of the underlying variables. In sensitivity checks, we confirm that using observed instead of predicted baseline values to define the indicator or using expected savings as the instrument by making additional parametric assumptions produces qualitatively similar results.

The cross-sectional nature of the instrument implies that we do not have identifying variation to predict differential time of entry. We therefore abstract away from staggered entry and assume that treatment begins in 2014 regardless of when hospitals joined BPCI. Hence, both the DD and IV analyses recover the average effect of ever participating in BPCI. Our results will therefore be conservative estimates of the effects of BPCI during the period in which hospitals participated.

Equation 4 presents the first stage equation and is equivalent to the DD model presented earlier as Equation 2. The outcome of interest is BPCI participation interacted with an indicator for the period beginning in 2014. We also interact the instrument z_h with the same indicator. Equation 5 presents the outcome model with β_{iv} as the coefficient of

12. To develop these measures we identify all index episodes with a MS-DRG ranging from 453 through 563 and 956. We then compute post-discharge episode spending over the 90 days following these stays.

interest. We estimate both models simultaneously using two stage least squares (2SLS).

$$d_h \cdot \mathbf{1}(t \geq 2014)_t = \alpha_h + \alpha_{mt} + X'_{iht}\,\delta_1 + \pi\, z_h \cdot \mathbf{1}(t \geq 2014)_t + \eta_{ihmt}, \tag{4}$$

$$Y_{ihmt} = \alpha_h + \alpha_{mt} + X'_{iht}\,\delta_2 + \beta_{iv}\, d_h \cdot \mathbf{1}(t \geq 2014)_t + \epsilon_{ihmt}. \tag{5}$$

The coefficient β_{iv} estimates the local average treatment effect (LATE) of BPCI participation on "complier" hospitals, i.e., those that comply with the instrument. Compliers are induced to join BPCI due to the greater potential financial reward, represented by the indicator z_h. The treatment effect for this set of compliers is of great policy relevance, since most alternative payment contracts attract participants using financial incentives.

4.2.2 Identification assumptions

A causal interpretation of the IV estimate β_{iv} is based on two untestable but standard identification assumptions.

Conditional independence: This assumption is standard in IV designs and implies that the instrument is as good as randomly assigned after conditioning on the covariates and fixed effects. Since this is an example of a shift-share or Bartik instrument, the key assumption here is that baseline hospital attributes are uncorrelated with future *changes* in the outcomes of interest. This assumption subsumes the exclusion restriction, i.e., having a high potential for financial gain in BPCI affects future patient outcomes only through the hospital's decision to participate in bundled payments. An implication of the exclusion restriction is that groups of hospitals facing low and high potential financial gain, as captured by the binary instrument, should evolve along parallel trends in the absence of BPCI. This is untestable, but we assess trends prior to BPCI using dynamic reduced form coefficients for all outcomes of interest. These are obtained by estimating a reduced form version of Equation 3 where d_h is replaced by z_h. These patterns suggest that there are no deviations among hospitals with high potential gain prior to BPCI.

Goldsmith-Pinkham, Sorkin, and Swift (2020) suggest that the concern of differential trends also extends to correlates that can affect both the instrument and the outcomes of interest. In the next section, we assess reduced form dynamic coefficients for key correlates and find similarly reassuring evidence.

Monotonicity: This assumption, sometimes also called uniformity (Heckman, Urzua, and Vytlacil 2006), imposes that an increase in the potential for financial gain weakly increases the probability of participation for all hospitals. Figure 2 Panels (a), (b), and (c) show that this condition holds on average for baseline orthopedic revenue, episode spending, and also when we consider them jointly through the instrument. The assumption is untestable, since it is maintained at the hospital level in terms of potential outcomes.

4.3 Instrument validity

This section presents results from balance tests for the instrument in support of conditional independence. We test whether hospitals with high and low potential for financial gain trended differently prior to the introduction of BPCI. We estimate the nonparametric event study model presented in Equation 6 below.

$$C_{hmt} = \alpha_h + \alpha_{mt} + \sum_{s<2013} \gamma_s \, z_h 1(t = s) + \xi_{hmt}, \tag{6}$$

C_{hmt} is the correlate of interest for hospital h located in the market m in year t. ξ_{hmt} denotes unobserved factors. All other terms in the equation have been previously defined. We estimate these models at the hospital level, with the exception of predicted risk of SNF use, which we estimate at the patient level. The coefficients of interest are the γ_s for each year. We estimate these models using data from the pre-BPCI period, 2008–13. The coefficients are estimated in reference to 2013, which is set to zero. Figure 3 plots the corresponding coefficients obtained along with the 95% confidence intervals.

We test balance against six correlates. The first four are measures of patient volume and revenue, gradually converging on the procedure of interest, LEJR. Differential trends in these measures may indicate differential focus on orthopedics or LEJR at hospitals with more potential for gain. The fifth correlate is MA penetration in the hospital's catchment area, calculated as before. Differential trends in MA penetration may raise concerns about unobserved differential changes in patient mix at hospitals with high potential for gain. The final correlate is a measure of patient complexity, predicted use of SNF, obtained using observed patient demographics, utilization, and diagnoses history. Reassuringly, the figures do not indicate consistent patterns of pre-trends on any of these correlates.[13]

The event study plots are reassuring, but they are often imprecisely estimated. To improve precision, we estimate a more parsimonious model with a single coefficient of interest on the interaction of the instrument, z_h, and a linear trend variable, T_t. Equation 7 presents the model. Table 2 Column 1 presents the corresponding coefficient, θ, for each correlate, along with the standard errors.

$$C_{hmt} = \alpha_h + \alpha_{mt} + \theta \, z_h T_t + \epsilon_{hmt}. \tag{7}$$

These coefficients are statistically insignificant and confirm the inconsistent visual patterns in the event studies. To help interpret the magnitude of these estimated trends for hospitals with high potential financial gain, we compare them to the overall trend estimated for all the hospitals in the sample. These coefficients are estimated using a similar

13. We examined trends in a number of time-varying correlates apart from those presented in Figure 3, but did not present them in the interest of brevity. These include LEJR readmissions and mortality rates, predicted probability of complications, readmission, mortality, home health and inpatient rehabilitation use following LEJR, constructed in a manner similar to predicted use of SNF, and hospital compare quality outcomes such as heart attack mortality and readmissions. Across the board, we did not find evidence of differential trends.

model (without the market-year fixed effect), where the coefficient of interest is on the linear trend variable without any interaction. Table 2 Column 2 reports the corresponding values. Comparing the values in columns 1 and 2, we find that the differential trend for hospitals with high potential financial gain is also very small apart from being statistically insignificant. For example, while the total volume per bed decreased by 0.37 patients each year on average across all hospitals, the differential trend for hospitals with high potential gain is less than 0.02. Overall, these patterns are not surprising since the policy and payment formulas were not designed to favor specific types of hospitals and were exogenously determined from the perspective of hospitals.

A potential violation of the exclusion restriction is the possibility that hospital baseline orthopedic attributes can directly affect future spending changes through coincident changes in hospital-wide factors such as managerial or physician quality. If so, this instrument should also be correlated with future changes in episode spending for patients in other clinical bundles of BPCI, including those unrelated to orthopedics. This rationale inspires a series of falsification tests in the spirit of Angrist, Lavy, and Schlosser (2010). We replicate our IV analysis on episode spending for the five largest bundles after LEJR, none of them related to orthopedics, and find no relationship between the instrument and participation or spending changes. Section 5.6 presents the corresponding results after we present the main IV estimates.

5 Results

5.1 First stage

Table 3 presents the DD and LATE estimates on the effects of BPCI participation on episode spending. Columns 1 to 4 present the results using patient-level data. Furthermore, since the instrument leverages static variation between hospitals at baseline, we estimate equivalent models on cross-sectional hospital-level data using the mean difference in spending before and after BPCI as the outcome.[14] This model serves as a sensitivity check and helps to probe the sensitivity of our quasi-experimental design to the exclusion of all covariates, including fixed effects.

We begin by describing the first stage results. Panel B presents estimates of the first stage coefficient, π, from Equation 4. Column 1 presents the coefficient from a specification including only fixed effects; Columns 2 and 3 from specifications that add patient or hospital controls, respectively. Column 4 presents results from a model that includes both patient and hospital controls. Columns 5 and 6 are equivalent to columns 1 and 2, respectively, in terms of the covariates used, but present results from models estimated on cross-sectional hospital level data. Our preferred specification includes patient controls,

14. We collapse the data to hospital-level means over 2009–13 (pre-BPCI) and 2014–16 (post-BPCI) and take the difference between the two for both the outcome and covariate values. This generates a cross-sectional model at the hospital level since the hospital fixed effects drop out. The sample for this model has 2,437 hospitals instead of 2,529 since 92 never-BPCI hospitals performed LEJR episodes prior to 2014, but not in the post-BPCI period. We weight each hospital observation by the number of patients in the pre-BPCI period.

but not hospital controls (cols. 2 and 6), since some hospital covariates may themselves be affected by participation.

Within each type of model (patient- or hospital-level), including or excluding the covariates does not affect the first-stage coefficient materially, consistent with the treatment balance implied by conditional independence. We obtain f-test statistics in the range of 17-18 and 24-25 in patient- and hospital-level models, respectively, which exceed the conventional threshold for weak instruments in a just-identified model with a single endogenous variable (Andrews, Stock, and Sun 2019).

As shown in Angrist and Pischke (2008), the first stage coefficient in a model with a binary instrument without covariates (column 5) equals the proportion of complier hospitals in the sample. This implies that around 20% of the hospitals in the sample are financial compliers. In addition, we calculate that the proportion of treated compliers is approximately 28%.[15] Therefore, a majority of participants were not induced to participate by the potential for financial gain embodied in our instrument, also known as "always takers." Measurement error in the instrument could suppress the magnitude of the first stage coefficient and underestimate the true share of financial compliers. However, comments made by hospital administrators in interviews with CMS contractors, discussed in Section 2.3, also imply that innovation and learning were the main drivers for participation in BPCI (Lewin Group 2016).

We test the sensitivity of the first stage results to the relaxation of the key assumptions used to construct the instrument. Table 4 presents the corresponding results of these checks, with Column 1 repeating the baseline coefficients to facilitate comparison. These models are estimated on hospital-level data and correspond to the specification in Column 6 of Table 3. We test sensitivity to using the baseline observed rather than the predicted orthopedic values (Col. 2). Next, we use an equivalent form of the instrument based on LEJR rather than orthopedic baseline values (Col. 3). Finally, we test the sensitivity to moving beyond a binary instrument by using a continuously varying value of potential financial gain instead. We predict a potential financial gain value for each hospital using parametric assumptions on the percent reduction in episode spending. In line with our rationale for the instrument, the predicted potential gain for a hospital increases in value along with orthopedic revenue and episode spending at baseline.[16] In Column 4, the model uses an indicator of above-median predicted potential gain as the instrument. The model in Column 5 directly uses the continuously varying value in thousands of dollars as the instrument. We note that the instrument constructed using LEJR data (Col. 3) is marginally weak, and hence we attach less importance to this model. The other variants do not raise weak IV concerns. Across different instrument versions, the first stage coefficient implies

15. This is calculated as the first stage coefficient x $P(z_h = 1)/P(d_h = 1)$. In our setting, this is 0.195 x 0.243/0.172 = 0.275.

16. We simulate 500 percent spending reduction values for each hospital by drawing values from normal distributions with a mean of 5% and standard deviations varying between 1% and 4%. Rather than allowing spending cuts to be randomly distributed, we impose a correlation of 0.9 between the hospital's baseline spending level and the percent spending reduction to ensure that hospitals with higher levels of spending at baseline achieve greater percent cuts. In each run, we apply the simulated percent reduction to the hospital's baseline mean episode spending value and then scale by LEJR patient volume to obtain the aggregate gain.

that about 13–30% of the hospitals are financial compliers. The implied proportions of the treated compliers across these models vary between 17% and 40%. Formulations of the instrument that impose parametric assumptions and generate continuous variation in potential financial gain imply a larger complier group.

5.2 Complier hospitals

Since the LATE estimates the causal treatment effect for hospitals that comply with the instrument, characterizing the complier group helps to interpret the estimated effects. Table A.1 compares the entire sample, complier, and treated hospital groups, respectively, on a number of relevant attributes using data from the pre-BPCI period. We follow Abadie (2003) to compute the values for the complier group, since they cannot be directly identified. Column 2 presents the proportion of hospitals in the full sample that have the respective attribute described in Column 1. Similarly, Columns 3 and 5 present the corresponding proportions within complier hospitals and participating hospitals, respectively. Complier hospitals are likely to be teaching and privately owned at rates similar to those of the average participant. However, they are more likely to be standalone, for-profit-owned, and smaller based on number of beds than the average participating hospital. The difference in for-profit share is striking: 20% of compliers are for-profit owned against only 12% of participants. It is intuitive that for-profit hospitals are more responsive to financial inducement than nonprofit hospitals. Interestingly, the average participant is less likely to be for-profit-owned than the average hospital in the sample.

Compliers appear to serve patients of lower complexity at baseline, as reported in Panel B. For example, while 53% of participating hospitals had patients with below-median predicted complication risk, the corresponding number for compliers is 69%. Panel C shows that the complier group also had better quality of care performance at baseline than the participant group as a whole, although the differences on this dimension are relatively muted. Overall, this analysis suggests that the financial compliers are more profit-oriented, moderately sized, and serve a healthier patient population than the average participant.

5.3 Episode spending

Table 3 presents the DD and LATE estimates of the effects on (log) episode spending in Panels A and B, respectively. In the case of Columns 5 and 6, the outcome is the change in mean log episode spending between the post- and pre-BPCI periods. Across the different specifications, the DD estimate implies approximately a 3% reduction in episode spending. The stability of the estimates across different models, including without the use of any covariate at all (in Col. 5) is reassuring. Figure A.2 Panel (a) presents the event study plot corresponding to the DD estimate, obtained by estimating Equation 3. The event study indicates a flat trend in spending at participating hospitals before BPCI and a decrease in spending after 2014, which is consistent with the parallel trends assumption. Since the mean episode spending level at participating hospitals before BPCI was about $25,000, the DD estimate implies gross savings of $750 on average during 2014–16. We emphasize that this estimate of spending reduction, calculated against the trends for nonparticipants, does not necessarily represent the savings for Medicare from BPCI due to the program's convoluted calculation of target price and financial payoffs, discussed in Section 2.2.

We turn to the LATE estimate presented in Panel B. The LATE implies a much greater effect of 11-12% across all models in the table. In addition to the specification permutations presented here, we also estimate models flexibly, allowing the coefficient on patient risk indicators to differ in the post-BPCI period or by year. This addresses recent criticisms about the potential for confounding due to time-varying covariates (Zeldow and Hatfield 2021). The coefficient of interest remains undisturbed in both the DD and IV models.

Figure 4 Panel (a) presents dynamic reduced form effects on episode spending before and after the introduction of BPCI. These are obtained by estimating a variant of Equation 3 where the instrument, z_h, replaces the indicator for participation, d_h. The pre-BPCI coefficients confirm the lack of a consistent trend prior to BPCI, in fact, the estimates are typically close to zero and support the exclusion restriction. The figure also shows a noticeable change in the trajectory of spending starting in 2014. Episode spending reduces differentially at hospitals with greater potential financial gain, relative to the remaining hospitals. We perform an additional exercise examining reduced form dynamic trends and average pre-post coefficients separately for high potential gain participants (i.e., those with $z_h = 1, d_h = 1$) and high potential gain nonparticipants ($z_h = 1, d_h = 0$), relative to hospitals with low potential gain ($z_h = 0$). Consistent with the exclusion restriction, which holds that the instrument influences outcomes only through participation, we find a differential decline in spending only among high potential gain hospitals that participated.[17]

Table 4 presents the corresponding LATE estimates obtained using alternate formulations of the instrument, as discussed previously. These estimates tend to be within two standard errors of the baseline estimate in Column 1, with the exception of the estimate in Column 3 using the marginally weak instrument. Across the board, all IV estimates are much larger in magnitude than the DD estimate of 3%. In addition to presenting conventional heteroscadisticity-robust standard errors, we also present weak instrument-robust 95% confidence intervals (square brackets). Regardless of which standard errors we use, the confidence intervals typically do not include the DD point estimate.[18]

Applying our preferred LATE estimate in Column 2 (11%) to the mean episode spending of $25,000 in BPCI hospitals prior to the program, we obtain an implied savings of $2,750 per episode. Figure A.3 Panel (a) presents the distribution of episode spending in the pre-BPCI period, providing additional benchmarks (e.g., median spending is around $20,000). Next, we explore the sources of spending reduction by examining the effects on the various components of care. We divide total spending into two buckets: the amount billed for index LEJR surgery and for all post-discharge care (readmissions and post-acute care including home care). These contribute approximately 60% and 40% of total episode

17. The average pre-post reduced form coefficient for high potential gain participants is -0.038 (s.e. 0.007). In contrast, the corresponding coefficient among high potential gain nonparticipants is -0.005 (s.e. 0.004).

18. We follow Andrews, Stock, and Sun (2019) and use the STATA command *twostepweakiv* to estimate the weak instrument robust AR confidence intervals for each instrument type.

spending, respectively.

5.4 Utilization of care

5.4.1 Index stay

Since spending for the initial hospital stay is determined by the associated DRG, hospitals have limited avenues to reduce the amount spent on the LEJR surgery. The principal avenue available to reduce index stay spending is to shorten patient stay to below a duration threshold that defines a "short" stay. For these stays, CMS pays hospitals a lower amount than the standard prospective amount. During our sample period, this threshold was three days for routine LEJR surgeries. Under the standard FFS contract, hospitals have little incentive to use the short-stay option even if medically appropriate, but under bundled payment they can recover the difference in reimbursement as episode savings. Another payment distortion under FFS was the 3-day minimum stay requirement to discharge patients to a skilled nursing facility. CMS waived this requirement for BPCI participants. Hence, we hypothesize that participants would differentially increase the proportion of index stays less than 3 days in duration.[19]

Table 5 Columns 1 and 2 present the estimated effects on log length of stay for LEJR surgery and the probability that the stay is less than 3 days. To help interpret these coefficients, Figure A.3 Panel (b) presents the CDF of the index stay duration using data before BPCI (the median and mode were 3 days). The DD estimate in Column 1 implies a statistically significant 4% decline in length of stay. Figure A.2 Panel (b) presents the corresponding event study plot and suggests a sharp decrease after 2013. The LATE estimate is more than two times in magnitude and implies a 10% decrease in length of stay for patients at financial compliers. Figure 4 Panel (b) presents the dynamic reduced form coefficients on log length of stay, which indicate flat pre-trends and a noticeable decline starting in 2014.

Column 2 again shows large differences between the DD and LATE coefficients. Although the average participant increases the proportion of short stays by 5.5 percentage points, the corresponding increase among complier hospitals is nearly three times, at 16 percentage points. The increase at complier hospitals represents a nearly 70% increase relative to the mean level before BPCI. This suggests a meaningful change in treatment protocols at complier hospitals after BPCI participation. Figure A.4 Panel (a) presents the corresponding reduced form event study on the probability that the index stay is shorter than 3 days.

5.4.2 Post-discharge care

Table 5 Column 3 presents the effects on PAC use. The DD estimate implies a small decrease of 1.7 percentage points for the average participant, about 2% of the mean level

19. CMS defines short stays as those with length less than the geometric mean for that DRG. During this period, the applicable threshold was 3 days for DRG 470 which represents 95% of the total volume. In our sample, the mean payment for 470 cases of length ≤ 2 days is about \$1,200 less than cases longer than 2 days.

of PAC use prior to BPCI. Figure A.2 Panel (c) presents the associated event study plot, which is consistent with the DD coefficient. In contrast, the LATE estimate indicates a large decline in the use of PAC by 13 percentage points, approximately 15% of the mean level. Figure 4 Panel (c) presents the corresponding reduced form dynamic effects. The figure shows no differential patterns before BPCI and a noticeable change in trajectory after 2014.

Table 5 Columns 4–6 present the effects on the use of specific forms of PAC: home health (HH), inpatient rehabilitation (IRF), and skilled nursing (SNF), respectively. The sum of mean values across individual types of PAC is greater than the probability of using any PAC because some patients use multiple types of PAC in the same episode, such as SNF and then HHA after discharge from the SNF. The DD coefficients in Panel A indicate modest decreases in IRF and SNF but a small increase in the use of HH. If we aggregate the estimated reduction in the use of SNF and IRF at their mean cost, the DD estimates imply a decrease of approximately $600 per episode. The *increase* in HH use mitigates the saving to about $520.[20] This net amount accounts for about 70% of the total decrease in episode spending reported in the previous section. In general, these patterns are consistent with a story of substitution away from expensive institutional PAC toward home-based care by the average participating hospital.

However, the patterns in Panel B imply a different strategy by the compliers. We find significant declines in the use of HH (30% of the mean) and IRF (66% of the mean), but no net change in the use of skilled nursing. Figure A.4 Panels (b)–(d) present the event studies for effects on the three types of PAC use. The patterns are consistent with the IV estimates. The finding of no net effect on skilled nursing may mask two opposing effects that offset each other: an increase in SNF because it substitutes for more intense forms of care such as IRF and simultaneously a decrease in SNF because patients are discharged to HH instead.

The LATE estimates on the use of IRF and HH imply a reduction in episode spending of approximately $1,900, assuming that there are no changes on the intensive margin. Of this, inpatient rehabilitation accounts for a reduction of $1,450 and home health care about $435. Comparing this to the estimated effect on episode spending of $2,750, we infer that two-thirds of the reduction in complier spending is due to a reduction in PAC use. Hence, decreasing PAC use is the dominant channel for savings for both compliers and the average participating hospital, but compliers decrease PAC more aggressively and on different margins.

5.5 Patient health

An important concern with expanded use of prospective payment is that it can encourage providers to skimp on useful care, which can lead to adverse outcomes for patients (Cutler 1995). CMS anticipated this unintended consequence when it launched this alter-

20. We compute this by scaling down the mean spending on IRF at BPCI hospitals by the change on the extensive margin: $2,188\,(\text{mean IRF spend}) * [1 - (0.138\,(\text{pre-bpci mean level}) - 0.011\,(\text{estimated reduction}))/0.138] \approx 174$. We obtain the estimates for SNF (427) and HH (-81) using similar logic.

nate payment contract and included some stipulations around maintaining performance on quality metrics. However, payments are not directly linked to quality of care in the same way as in other CMS programs, such as the Hospital Readmissions Reduction Program. Previous studies have found no economically or statistically meaningful effects of bundling payments on quality of care. We also examine several quality measures. All our models include extensive patient-level controls to adjust for differences in observed risk. We explore concerns about potential changes in patient mix, including unobserved selection, in Section 5.6.

Table 6 presents the estimated DD and LATE coefficients on quality metrics in Panels A and B, respectively. In addition to the commonly studied quality metrics such as readmissions and mortality, we also examine the effects on the probability of patients undergoing a revision surgery within 90 days of the index procedure. Revision surgeries are performed to correct difficulties that arise in the artificial joint implanted during the index procedure (e.g., loosening of the joint resulting in dislocations) and are hence more closely linked to lapses in surgery quality than all-cause readmissions. In addition, these are unambiguously bad outcomes for patients (Katz 2006). Some experts had expressed concerns that the financial incentive to cut costs can lead to changes in implants used by hospitals, ultimately requiring more revisions (Koenig et al. 2018). These concerns appear to be well-founded since hospital administrators discussed device standardization and vendor consolidation as a strategy to reduce implant costs (Lewin Group 2017). Importantly, revision surgeries is not one of the quality metrics monitored by CMS. Hence, examining the effects on this "low stakes" outcome potentially helps us to assess quality more realistically (Jacob 2005).

The DD coefficients in Panel A suggest no meaningful effect on quality of care. There is a marginally significant decrease in readmission rates of 0.3 percentage points, but the estimated effects on all other outcomes are small and insignificant. We interpret these results as signaling no change in the quality of care at the average participating hospital. The LATE estimates in Panel B also generally imply no clear increase in adverse outcomes for patients, except for an increase in the frequency of revision surgeries. Although the coefficient is small in absolute terms, it implies an increase of 100% relative to the mean rate before BPCI. Figure 5 presents the corresponding reduced form event studies for each outcome presented in the table. The plots confirm that hospitals with high potential financial gain did not trend differently before BPCI. For most outcomes, there are no differential trends before or after the introduction of BPCI. Panel (c) presents dynamic effects on the probability of revision surgeries and confirms a gradual increase after BPCI.

The estimated effect on revision surgeries implies an expected increase in Medicare spending of approximately \$380 per episode.[21] This increase in cost is about 15% the size of the implied reduction in total episode expenditure of \$2,750 at the average complier hospital. However, this does not include the non-financial harm to patients. Patient surveys conducted by CMS suggest that LEJR patients experienced more adverse outcomes

21. We compute this as 1 percentage point of the average amount billed by hospitals to Medicare for a revision surgery episode, of about \$38,000. This is potentially conservative since it includes surgery and follow-up care only in the 30 days after discharge.

at BPCI hospitals after participation. For example, they reported a statistically significant worsening in mobility, hospital discharge experience (including whether they were discharged at the right time), and overall satisfaction with their recovery. These results are not specific to the financial compliers and are small in magnitude but are consistent with a decrease in quality of care (Lewin Group 2018b).

5.6 Threats to validity and alternative explanations

5.6.1 Falsification: Non-orthopedic bundles

A key concern is that baseline orthopedic attributes, used to determine the potential for financial gain, can be correlated with future spending changes through unobserved changes in hospital-wide production inputs. An example of this concern would be that hospitals with high potential financial gain may have also received boosts in their managerial quality in the period around 2014. We assess the importance of this concern through a falsification test. We test whether our instrument, designed to predict participation in LEJR or an orthopedic bundle only, also predicts participation and changes in spending in *non-orthopedic* bundles. These bundles are for episodes of heart failure, pneumonia, chronic obstructive pulmonary disease (COPD), sepsis, and urinary tract infection (UTI), in order of the number of hospitals participating in BPCI.[22] The purpose of this exercise is not to test whether BPCI delivered spending reductions in other bundles, but to assess the validity of the instrument.

Table 7 presents the corresponding point estimates. We present the estimates for LEJR in Column 1 for ease of comparison. The remaining columns present estimates for the different bundle types, ordered in decreasing degree of participation. Column 7 presents the results corresponding to a pooled sample of the five non-orthopedic bundles. Panel A presents the DD estimate on episode spending for these bundles, which confirms that we find similar results as in previous studies of BPCI. For example, like us, Lewin Group (2018a) Exhibit 13 reports statistically significant reductions in episode spending in the LEJR and UTI bundles. Hence, BPCI also produced spending reductions in other bundles besides LEJR.

Panel B presents coefficients from the IV models. First, we confirm that the instrument has negligible predictive power for participation in the non-orthopedic bundles. The only exception is heart failure where the f-statistic for the first stage is marginally higher than the conventional rule of thumb of 10. Since the first stage is typically weak, we focus on the reduced form coefficients instead of the LATE estimates. Reassuringly, the instrument does not predict changes in episode spending for these bundles, except in the case of COPD, where it predicts an *increase* in expenditure. When we consider a pooled sample of the five conditions, the reduced form coefficient is small and statistically insignificant, and the confidence interval excludes the coefficient we obtain for LEJR. Figure 6

22. Table A.2 Panels A and B show that the correlations between LEJR and these bundles on total and post-discharge spending were typically around 0.7 and 0.6, respectively. Hence, if hospital-wide factors drive changes in LEJR spending, they can plausibly do so for the other bundles.

presents the corresponding dynamic reduced form effects on episode spending for LEJR (thick red line) and each of the comparison conditions (gray lines). The figure illustrates that our orthopedic-based instrument is not correlated with the changes in spending for non-orthopedic bundles. This evidence reassures us that the instrument does not predict a reduction in episode spending where there is no basis for a relationship, supporting the exclusion restriction.

5.6.2 Contemporaneous hospital-wide changes

A threat is that our results may reflect contemporaneous changes in spending at participating hospitals that are unrelated to BPCI. For example, these could reflect responses to broader changes in the regulatory environment. We test the importance of this concern in two ways. First, in this spirit of falsification, we examine whether BPCI participants also produce similar spending reductions for an unrelated surgery, which would reflect hospital-wide changes. Second, we test the sensitivity of the baseline coefficients to using a modified model that allows hospitals with specific attributes to have a differential effect after 2013.

Unrelated surgical episode

We examine the effects on episode spending and PAC use for cholecystectomy (gallbladder removal) surgeries in hospitals participating in the LEJR bundle. This is a high-volume surgery that frequently results in the use of PAC and is performed by a different type of surgeon (general or gastrointestinal specialty), mitigating the potential for spillovers of clinical practice style within the hospital.

Table A.3 presents the corresponding results on episode spending and use of PAC for this surgery in Columns 1 and 2, respectively. Panel A reports the DD coefficients, which are small and statistically insignificant. The confidence interval of the spending estimate is precise enough to exclude the corresponding DD coefficient on LEJR spending. Panel B presents the corresponding LATE coefficients using our baseline instrument. These suggest small, positive, and statistically insignificant changes in the outcomes. The confidence intervals here also exclude the main LATE coefficients estimated in the case of LEJR. Figure A.5 presents the reduced form dynamic coefficients and confirms the absence of changes in trends after 2013. Hence, the decrease in the spending for LEJR episodes at participating hospitals or at financial compliers does not reflect hospital-wide trends.

Sensitivity checks

Since the instrument assigns hospitals with greater Medicare orthopedic revenue to greater potential financial gain, one may worry that the LATE partly reflects spending patterns at larger hospitals that adapted better to broader changes in regulation. We assess the importance of this concern by testing the sensitivity of the main DD or IV coefficient to including an interaction term of an indicator of above-median baseline hospital revenue with a flag for the post-BPCI period. Table A.4 presents the corresponding results. Panels A and B present the DD and IV estimates, respectively. For brevity, we focus on the primary specifications, which correspond to Columns 2 and 6 in Table 3, respectively. The upper row of each panel presents the baseline coefficients to facilitate comparison.

Panel A shows that the DD coefficients are essentially unchanged when we allow large hospitals to have a differential trend post-2013. Large hospitals reduce LEJR episode spending differentially after 2013. However, this does not affect the DD coefficient on BPCI participation. Similarly, we include this interaction term in the model estimated by 2SLS. The coefficients in Panel B show that the first stage and LATE coefficients attenuate by about 10-15%, but continue to be very precisely estimated and well within the confidence interval of the main estimate.

Another concern in the same vein is that the participating hospitals with more favorable historic trends disproportionately contribute to the average savings we estimate for BPCI. This is not necessarily inefficient, since both the DD and the IV estimators calculate savings as the result of a deviation in trends for participants relative to the trends of the counterfactual groups. If participants continued to progress along their pre-BPCI trends, this approach would estimate little or no spending reduction. However, it does affect the interpretation of the results.

We empirically test the sensitivity of the baseline coefficients to allowing hospitals that enjoyed a favorable historical trend to have a differential effect after 2013. Table A.5 presents the corresponding results. Panel A presents the results for the DD model. The top row repeats the baseline results from Table 3 Columns 2 and 6 to facilitate comparison. We find that hospitals with favorable historic trends reduced episode spending differentially, but this change in the model hardly affects the main DD coefficients. Hence, we conclude that the ATT estimate for BPCI is not driven by hospitals that already enjoyed a favorable spending trajectory.

We similarly examine the sensitivity of the IV estimates. We first check the correlation between the instrument, the potential for high financial gain, and the hospital's historic spending trend. Figure A.1 Panel B presents the share of hospitals in each quartile of expected gain deemed to have a high potential for financial gain. Reassuringly, shares are relatively similar across all quartiles, ranging between 13% and 16%. A formal regression of the quartile indicators on the instrument recovers coefficients that are jointly statistically insignificant, with a p-value of 0.39. The coefficients are also individually statistically insignificant at the 5% level. We then modify the IV model by including the interaction term for hospitals with favorable historical trends. The modified results show that, although these hospitals differentially reduced spending after 2014, this change does not significantly affect the LATE estimates. Hence, the causal effect for financial compliers is also not driven by hospitals with favorable historical trends.

5.6.3 Mean reversion

We assess whether the LATE estimate could be driven by mean reversion in episode spending. Our choice of using 2007 data to construct the instrument mitigates this possibility. Nevertheless we assess this concern and study the unadjusted time series trends in mean episode spending in Figure A.6. The figure presents the trends separately for hospitals with high and low values of potential financial gain. To focus on trends and abstract from levels, we normalize the values so that each curve is set at 100 in 2013. The two groups evolve in parallel over six years (from 2008 through 2013) and diverge only after the introduction of BPCI. Trends virtually overlap between 2010 and 2013. This evidence suggests that mean reversion is not empirically important in this setting.

5.6.4 Patient selection

We evaluate in multiple ways the possibility that BPCI participants selected patients based on risk to facilitate reducing episode spending. However, we find no evidence to support this channel, similar to previous studies (Finkelstein et al. 2018; Barnett et al. 2019; Einav et al. 2022). Table A.6 and Figure 7 present the corresponding point estimates and reduced form dynamic effects, respectively. We directly assess whether there is a differential change in the observed risk of patients in participating hospitals after BPCI. We summarize information on patient risk across a rich set of disease comorbidities and utilization history by generating a predicted risk of adverse outcomes such as readmissions and the use of different types of PAC.[23] Figure 7 Panels (a)–(d) present the corresponding reduced form event study plots. There is no discernible pattern before or after 2013. Table A.6 Panel A presents the corresponding coefficients. Although some DD coefficients are precisely estimated and suggest a change in predicted use of SNF or HH care, all LATE coefficients are statistically insignificant. More importantly, both the DD and LATE sets of coefficients are very small in magnitude when compared to the mean level of use or to the estimated effects on PAC use reported previously. For example, we can rule out a decrease in the predicted use of IRF at financial compliers of more than 0.5 percentage points, which is only about 5% of the corresponding estimated decrease in the use of IRF at compliers.

We then explore the possibility of patient selection on unobserved risk by examining changes in patient mix on other dimensions that could signal such selection, such as changes in patient distance to the index hospital, computed as the distance between the 5-digit zipcodes of the patient and the hospital where the index surgery was performed. Figure 7 Panels (e) and (f) present reduced form event studies with distance in levels and logs, respectively. The plots imply that there is no change in patient distance after BPCI. Table A.6 Panel B Columns 1 and 2 present the corresponding point estimates with respect to distance. The coefficients are imprecisely estimated and indicate small effects when the outcome is in levels.

23. We first estimate probit models using data from 2008 predicting the risk of a readmission, and the use of inpatient rehabilitation, skilled nursing or home health during the 90-day episode. We then use the model coefficients to generate predicted risk measures for patients in 2009–16.

Next, we use an alternate approach to test whether hospitals have changed their patient mix based on location and demographics. We estimate the patient propensity for treatment in a BPCI hospital based on zip code and observables using data from 2008. We apply these model coefficients to patients in the analysis sample and generate predicted probabilities of their choosing a BPCI hospital, which we refer to as "BPCI propensity." We use these probabilities as the outcome in this analysis and examine whether there is a change in the BPCI propensity of patients admitted to BPCI hospitals after 2013. We continue to find null effects, which are presented in Figure 7 Panels (g) and (h). Table A.6 Panel B Columns 3 and 4 present the corresponding point estimates, which support the conclusion of no change. Overall, the results imply that there is no change in patient complexity material enough to explain the large reduction in care utilization at financial compliers.

6 Discussion and Conclusion

There has been an unprecedented expansion of alternate payment contracts in US healthcare in order to align provider and social objectives. These contracts are usually voluntary in nature and offer financial incentives to drive both participation and changes in behavior. However, mixed evidence on their effects has led to an active debate on the wisdom of using financial incentives to improve performance. This paper informs this debate by studying an important voluntary payment reform, BPCI, which targeted savings for Medicare, the US federal health insurance program for the elderly.

Overall, our results provide some basis for optimism. We find little evidence of inefficient selection on baseline spending levels and trends, which suggests that thoughtful contract design can mitigate such inefficiencies in voluntary programs. Using a novel IV approach that leverages idiosyncratic variation in potential gain across hospitals due to the payment formula, we estimate the causal treatment effects for hospitals induced to participate by the financial reward. We show that these "financial compliers" constitute a substantial minority of participating hospitals, which is consistent with interviews of administrators suggesting that learning, not financial gain, was the primary motivation to participate. Compliers have some distinctive characteristics, such as a much higher probability of being for-profit-owned. Compliers obtain gross savings three times that of the average participating hospital by reducing medical services much more aggressively. We rule out several alternative explanations, including the possibility that compliers change patient mix to achieve these savings. Comparing the ATT and the LATE, the results further imply that always takers with regard to our instrument achieve very low savings. Although the specific estimates pertain to healthcare, the general takeaways are relevant to settings beyond healthcare since voluntary programs are popular policy tools in many sectors of the economy, including education, energy, and environment conservation.

However, not all is rosy. Although we do not find evidence of a decrease in quality of care at the average participating hospital and on most quality measures at compliers, we detect a relatively large increase in revision surgeries for patients at complier hospitals. This is an unambiguous signal of a decline in quality and suggests that achieving large savings in Medicare spending may involve a tradeoff with patient welfare. More work is needed, ideally using detailed clinical or survey data, to better understand the effects on patient welfare.

Although we have begun to shed light on this important topic, there are several avenues for future work. This paper focused on how hospitals respond to the use of financial incentives in voluntary contracts, which is the key lever pulled by healthcare regulators and insurers in such programs. However, our results suggest that they should also explore other inducement tools, as multiple factors influence provider decisions and could offer paths to attract participants more likely to succeed under the new contract. More evidence is also needed on the effectiveness of different contract designs and information interventions so that prospective participants are correctly matched to contracts to maximize social objectives.

References

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113 (2): 231–263.

Acemoglu, Daron, and Amy Finkelstein. 2008. "Input and technology choices in regulated industries: Evidence from the health care sector." *Journal of Political Economy* 116 (5): 837–880.

Agarwal, Rajender, Joshua M Liao, Ashutosh Gupta, and Amol S Navathe. 2020. "The Impact Of Bundled Payment On Health Care Spending, Utilization, And Quality: A Systematic Review: A systematic review of the impact on spending, utilization, and quality outcomes from three Centers for Medicare and Medicaid Services bundled payment programs." *Health Affairs* 39 (1): 50–57.

Alexander, Diane. 2020. "How do doctors respond to incentives? Unintended consequences of paying doctors to reduce costs." *Journal of Political Economy* 128 (11): 4046–4096.

Andrews, Isaiah, James H Stock, and Liyang Sun. 2019. "Weak instruments in instrumental variables regression: Theory and practice." *Annual Review of Economics* 11 (1).

Angrist, Joshua, Victor Lavy, and Analia Schlosser. 2010. "Multiple experiments for the causal link between the quantity and quality of children." *Journal of Labor Economics* 28 (4): 773–824.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics.* Princeton university press.

Barnett, Michael L, Andrew Wilcock, J Michael McWilliams, Arnold M Epstein, Karen E Joynt Maddox, E John Orav, David C Grabowski, and Ateev Mehrotra. 2019. "Two-year evaluation of mandatory bundled payments for joint replacement." *New England Journal of Medicine* 380 (3): 252–262.

Berlin, Nicholas L, Timothy A Peterson, Zoey Chopra, Baris Gulseren, and Andrew M Ryan. 2021. "Hospital Participation Decisions In Medicare Bundled Payment Program Were Influenced By Third-Party Conveners." *Health Affairs* 40 (8): 1286–1293.

Chopra, Zoey, Baris Gulseren, Karan R Chhabra, Justin B Dimick, and Andrew M Ryan. 2023. "Bundled payments for care improvement efficacy across 3 common operations." *Annals of Surgery* 277 (1): e16–e23.

Cutler, David M. 1995. "The incidence of adverse medical outcomes under prospective payment." *Econometrica* 63 (1): 29.

Darden, Michael, Ian Mccarthy, and Eric Barrette. 2023. "Who Pays in Pay-for-Performance? Evidence from Hospital Prices and Financial Penalties." *American Journal of Health Economics* 9 (3): 435–460.

Dummit, Laura A, Daver Kahvecioglu, Grecia Marrufo, Rahul Rajkumar, Jaclyn Marshall, Eleonora Tan, Matthew J Press, Shannon Flood, L Daniel Muldoon, Qian Gu, et al. 2016. "Association between hospital participation in a Medicare bundled payment initiative and payments and quality outcomes for lower extremity joint replacement episodes." *JAMA* 316 (12): 1267–1278.

Einav, Liran, Amy Finkelstein, Yunan Ji, and Neale Mahoney. 2022. "Voluntary regulation: Evidence from Medicare payment reform." *The Quarterly Journal of Economics* 137 (1): 565–618.

Finkelstein, Amy, Yunan Ji, Neale Mahoney, and Jonathan Skinner. 2018. "Mandatory Medicare bundled payment program for lower extremity joint replacement and discharge to institutional postacute care: interim analysis of the first year of a 5-year randomized trial." *JAMA* 320 (9): 892–900.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik instruments: What, when, why, and how." *American Economic Review* 110 (8): 2586–2624.

Gupta, Atul. 2021. "Impacts of performance pay for hospitals: The readmissions reduction program." *American Economic Review* 111 (4): 1241–83.

Heckman, James J, Sergio Urzua, and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The Review of Economics and Statistics* 88 (3): 389–432.

Jacob, Brian A. 2005. "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics* 89 (5-6): 761–796.

Katz, Jeffrey N. 2006. "Total joint replacement in osteoarthritis." *Best practice & research Clinical rheumatology* 20 (1): 145–153.

Koenig, Lane, Chaoling Feng, Fang He, and Jennifer T Nguyen. 2018. "The effects of revision total hip arthroplasty on medicare spending and beneficiary outcomes: implications for the comprehensive care for joint replacement model." *The Journal of Arthroplasty* 33 (9): 2764–2769.

Lewin Group. 2016. *CMS Bundled Payments for Care Improvement Initiative Models 2-4: Year 2 Evaluation & Monitoring Annual Report.* Technical report. October.

———. 2017. *CMS Bundled Payments for Care Improvement Initiative Models 2-4: Year 3 Evaluation & Monitoring Annual Report.* Technical report. October.

———. 2018a. *CMS Bundled Payments for Care Improvement Initiative Models 2-4: Year 4 Evaluation & Monitoring Annual Report.* Technical report. June.

———. 2018b. *CMS Bundled Payments for Care Improvement Initiative Models 2-4: Year 5 Evaluation & Monitoring Annual Report.* Technical report. October.

McWilliams, J Michael. 2022. "Pay for performance: when slogans overtake science in health policy." *JAMA* 328 (21): 2114–2116.

McWilliams, J Michael, Laura A Hatfield, Bruce E Landon, and Michael E Chernew. 2020. "Savings or selection? Initial spending reductions in the Medicare Shared Savings Program and considerations for reform." *The Milbank Quarterly* 98 (3): 847–907.

MedPAC. 2021. *Report to the Congress: Medicare and the healthcare delivery system.* Technical report. June. http://www.medpac.gov/docs/default-source/reports/jun21_medpac_report_to_congress_sec.pdf?sfvrsn=0.

Milad, Marina A, Roslyn C Murray, Amol S Navathe, and Andrew M Ryan. 2022. "Value-Based Payment Models In The Commercial Insurance Sector: A Systematic Review: Systematic review examines value-based payment models in the commercial insurance sector." *Health Affairs* 41 (4): 540–548.

Navathe, Amol S, Ezekiel J Emanuel, Atheendar S Venkataramani, Qian Huang, Atul Gupta, Claire T Dinh, Eric Z Shan, Dylan Small, Norma B Coe, Erkuan Wang, et al. 2020. "Spending And Quality After Three Years Of Medicare's Voluntary Bundled Payment For Joint Replacement Surgery: The spending and quality effects of Medicare's Bundled Payments for Care Improvement initiative among patients undergoing lower extremity joint-replacement." *Health Affairs* 39 (1): 58–66.

Navathe, Amol S, Joshua M Liao, Daniel Polsky, Yash Shah, Qian Huang, Jingsan Zhu, Zoe M Lyon, et al. 2018. "Comparison of hospitals participating in Medicare's voluntary and mandatory orthopedic bundle programs." *Health Affairs* 37 (6): 854–863.

Navathe, Amol S, Andrea B Troxel, Joshua M Liao, Nan Nan, Jingsan Zhu, Wenjun Zhong, and Ezekiel J Emanuel. 2017. "Cost of joint replacement using bundled payment models." *JAMA Internal Medicine* 177 (2): 214–222.

Norton, Edward C, Jun Li, Anup Das, and Lena M Chen. 2018. "Moneyball in medicare." *Journal of Health Economics* 61:259–273.

Ryan, Andrew M, Jason D Buxbaum, Sukruth A Shashikumar, and Zoey Chopra. 2023. "It's Time For A New Season Of Episode-Based Payment." *Health Affairs Forefront.*

Smith, Brad. 2021. "CMS innovation center at 10 years—Progress and lessons learned." *N Engl J Med* 384 (8): 759–764.

Weitzman, Martin L. 1980. "The" ratchet principle" and performance incentives." *The Bell Journal of Economics,* 302–308.

Werner, Rachel M, Jonathan T Kolstad, Elizabeth A Stuart, and Daniel Polsky. 2011. "The effect of pay-for-performance in hospitals: lessons for quality improvement." *Health Affairs* 30 (4): 690–698.

Zeldow, Bret, and Laura A Hatfield. 2021. "Confounding and regression adjustment in difference-in-differences studies." *Health Services Research* 56 (5): 932–941.

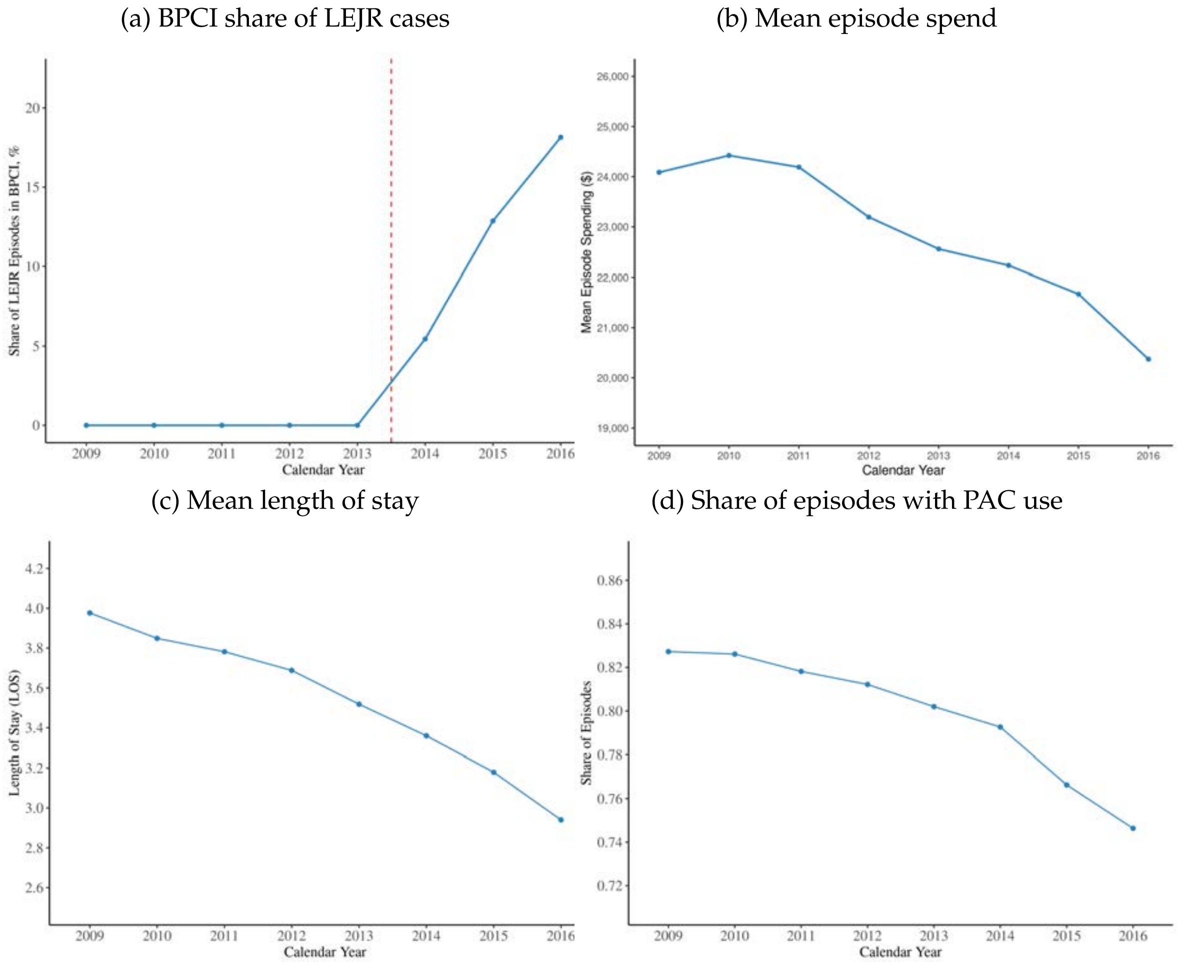
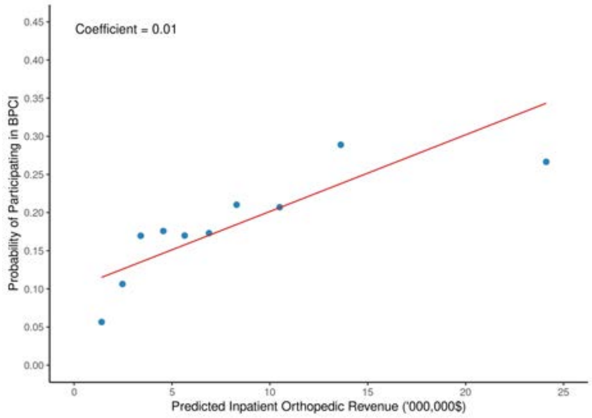


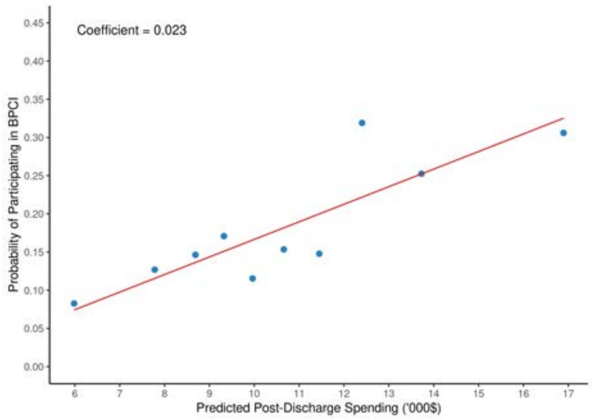
Figure 1: Trends in Medicare LEJR surgeries

Note: This figure presents trends in lower extremity joint replacement (LEJR) cases for Medicare fee-for-service patients (admissions for DRG 469 or 470). Panel (a) presents the share of LEJR episodes performed at BPCI hospitals over time across all markets in the US (except US territories and Maryland). Panel (b) presents the trend in mean 90-day episode spending, expressed in 2016 dollars. Panel (c) presents the mean length of stay for index LEJR surgeries in our sample. Panel (d) presents the mean share of episodes with any PAC use (inpatient rehabilitation, skilled nursing, or home health). Note that 2016 includes episodes only from the first three quarters of the year since we cannot observe the entire episode for surgeries performed in the last quarter.

(a) Orthopedic revenue



(b) Post-discharge episode spend



(c) Joint distribution

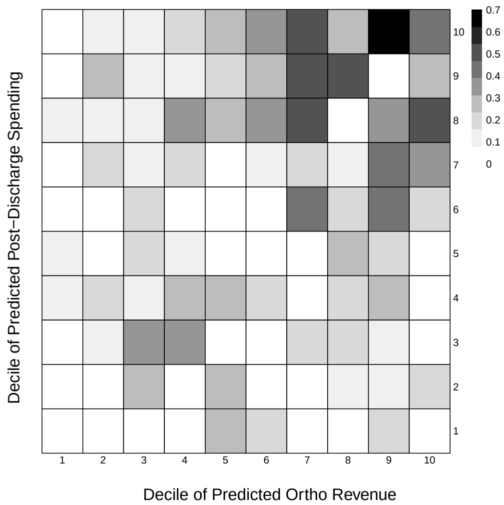


Figure 2: BPCI participation and baseline orthopedic performance

Note: This figure presents the association between BPCI participation between 2014–16 and baseline orthopedic performance, which is calculated using 2007 data. Panel (a) presents a binned scatter plot of the relationship between the hospital's Medicare baseline orthopedic revenue (millions of dollars) on the X-axis and participation on the Y-axis. Panel (b) replaces orthopedic revenue with mean post-discharge episode spending (thousands of dollars) in orthopedic episodes. These plots also present the slope coefficient obtained from an OLS regression of the corresponding baseline orthopedic value on an indicator of participation. Panel (c) presents the variation in participation against the joint distribution of orthopedic revenue and post-discharge spending. Orthopedic performance measures are obtained using data in 2007, described in Section A.3. Each hospital is weighted by the number of LEJR cases prior to BPCI.

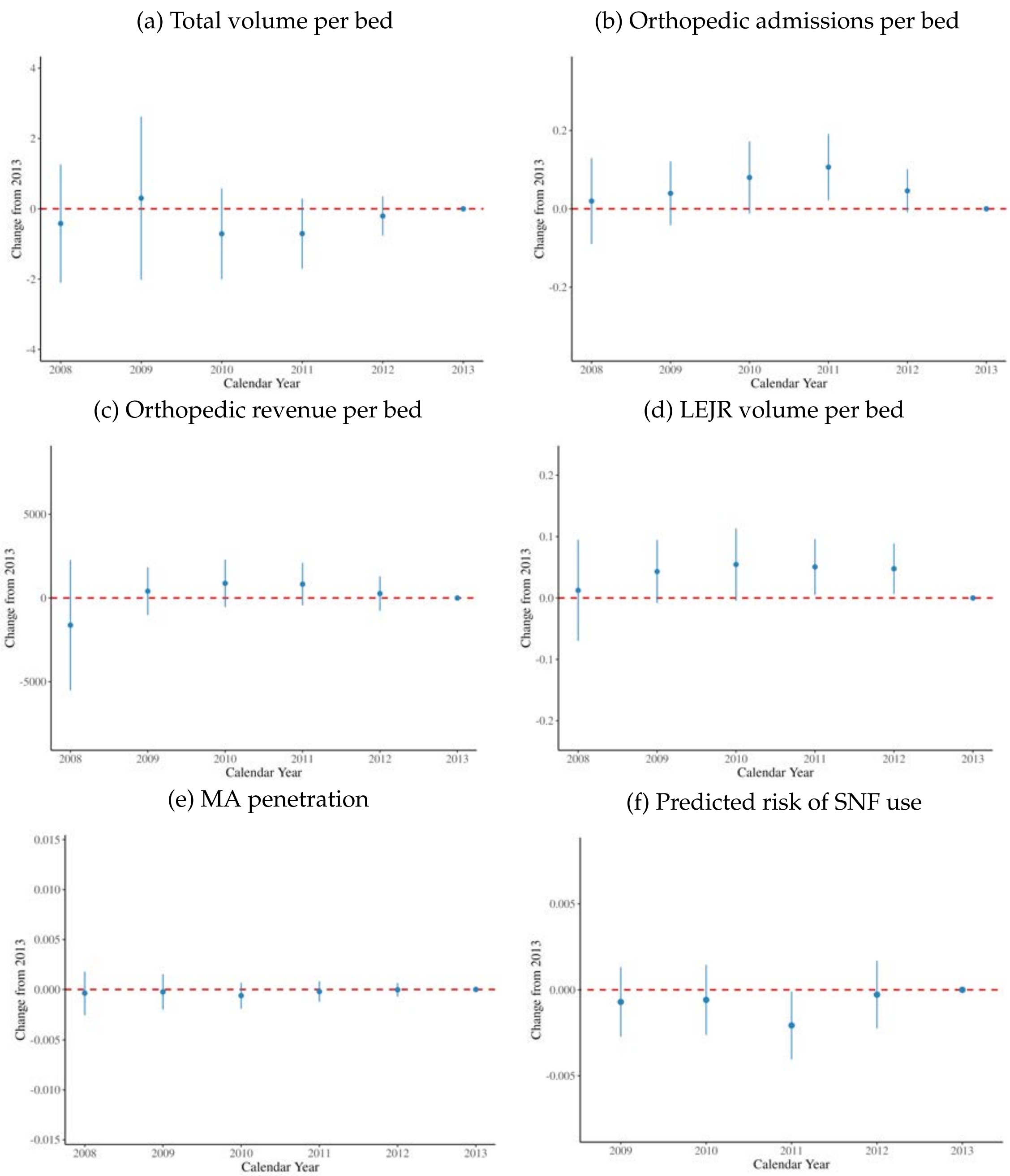
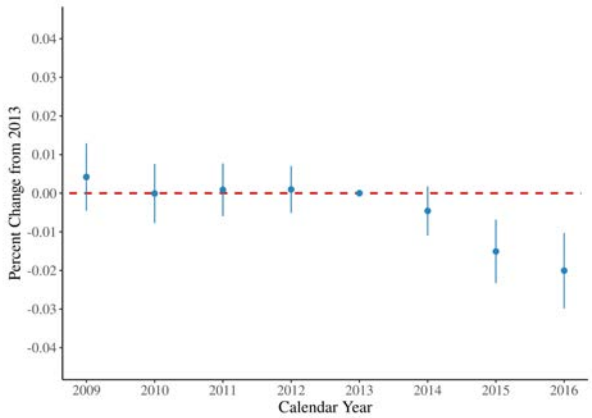


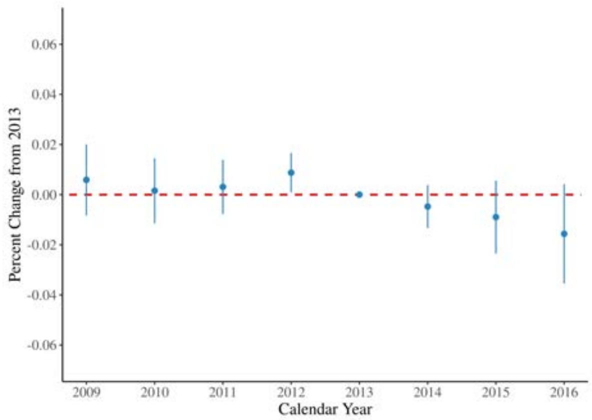
Figure 3: Trends prior to BPCI

Note: This figure presents estimated reduced form dynamic effects (i.e., comparing hospitals with low vs. high potential for gain) in time-varying hospital and patient characteristics prior to BPCI. The coefficients are obtained by estimating Equation 6. Total volume in Panel (a) is taken from annual American Hospital Association (AHA) surveys. Panels (b) and (c) present effects on Medicare fee-for-service orthopedic admissions and revenue per bed, respectively. Panel (d) presents trends in Medicare fee-for-service lower extremity joint replacement (LEJR) volume per bed. Panel (e) presents the trends in Medicare Advantage (MA) penetration, computed as described in Section 3.2. In Panel (f), predicted values from a probit model (fit on 2008 data) for each patient admitted with LEJR serve as our measure of predicted risk of SNF use. The sample for predicted risk of SNF use begins in 2009 in order to avoid reusing the 2008 cases used to estimate the probit model. The models in Panels (a)–(e) are estimated on hospital-year level data, while that in Panel (f) is estimated on patient-level data.

(a) Episode spending (log)



(b) LEJR length of stay (log)



(c) Use of any post-acute care

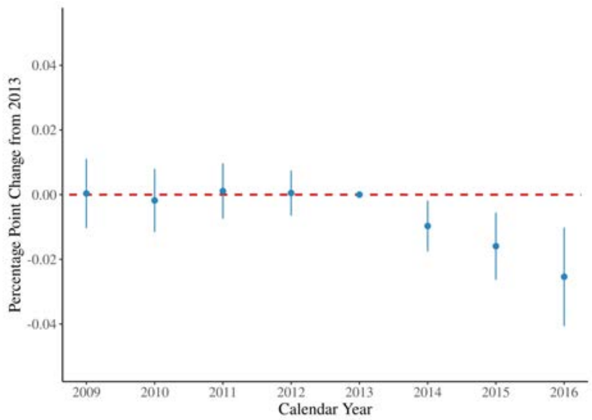


Figure 4: Effects on spending and utilization of care

Note: This figure presents reduced form dynamic effects (i.e., comparing hospitals with low vs. high potential for gain) on (log) episode spending (Panel a), index stay length in logs (Panel b), and the use of any post-acute care (Panel c). The coefficients are obtained by estimating a reduced form variant of Equation 3 where the participation indicator, d_h, is replaced by the instrument, z_h. In addition to hospital and market by year fixed effects, the models also include patient controls. Standard errors are clustered by hospital. Figure A.2 presents the corresponding DD event study plots obtained by comparing participants to nonparticipants.

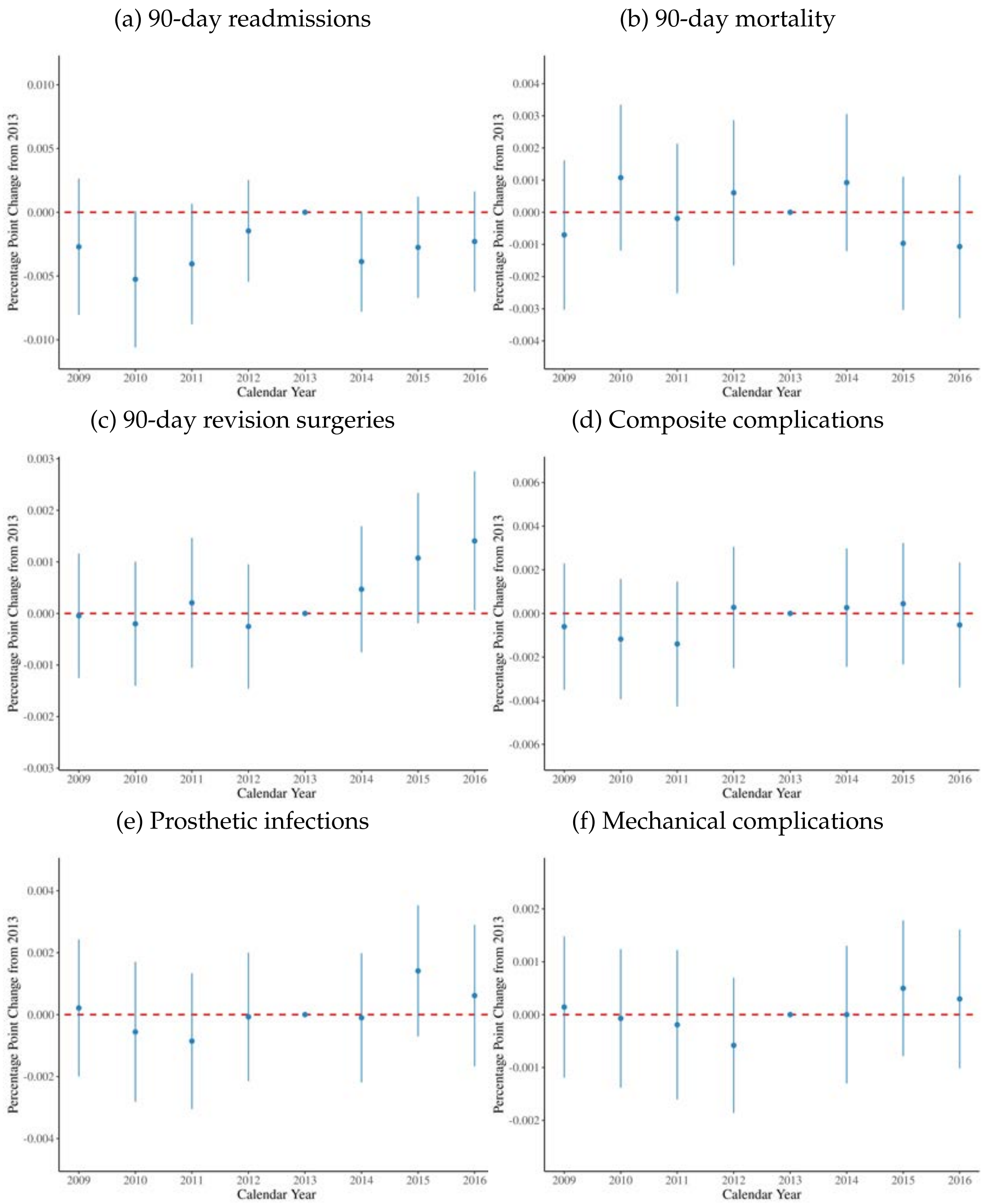


Figure 5: Effects on quality of care

Note: This figure presents reduced form dynamic effects (i.e., comparing hospitals with low vs. high potential for gain) on standard quality of care outcomes for LEJR patients. The coefficients are obtained by estimating a reduced form variant of Equation 3 where the participation indicator, d_h, is replaced by the instrument, z_h. Readmissions, presented in Panel (a), are defined to include revision surgeries. Revision surgeries are defined using DRGs 466, 467, and 468. Complication measures are constructed following CMS documentation, as described in Section A.1. The models include hospital and market by year fixed effects and patient controls. Standard errors are clustered by hospital.

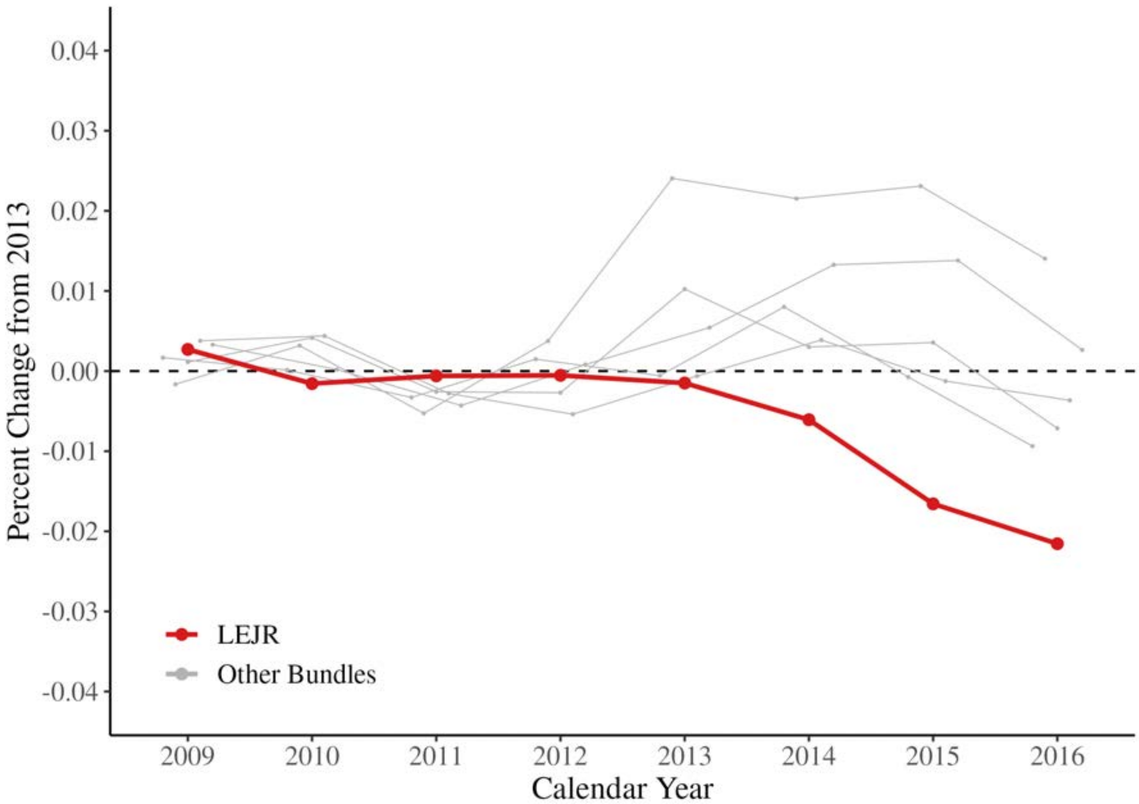
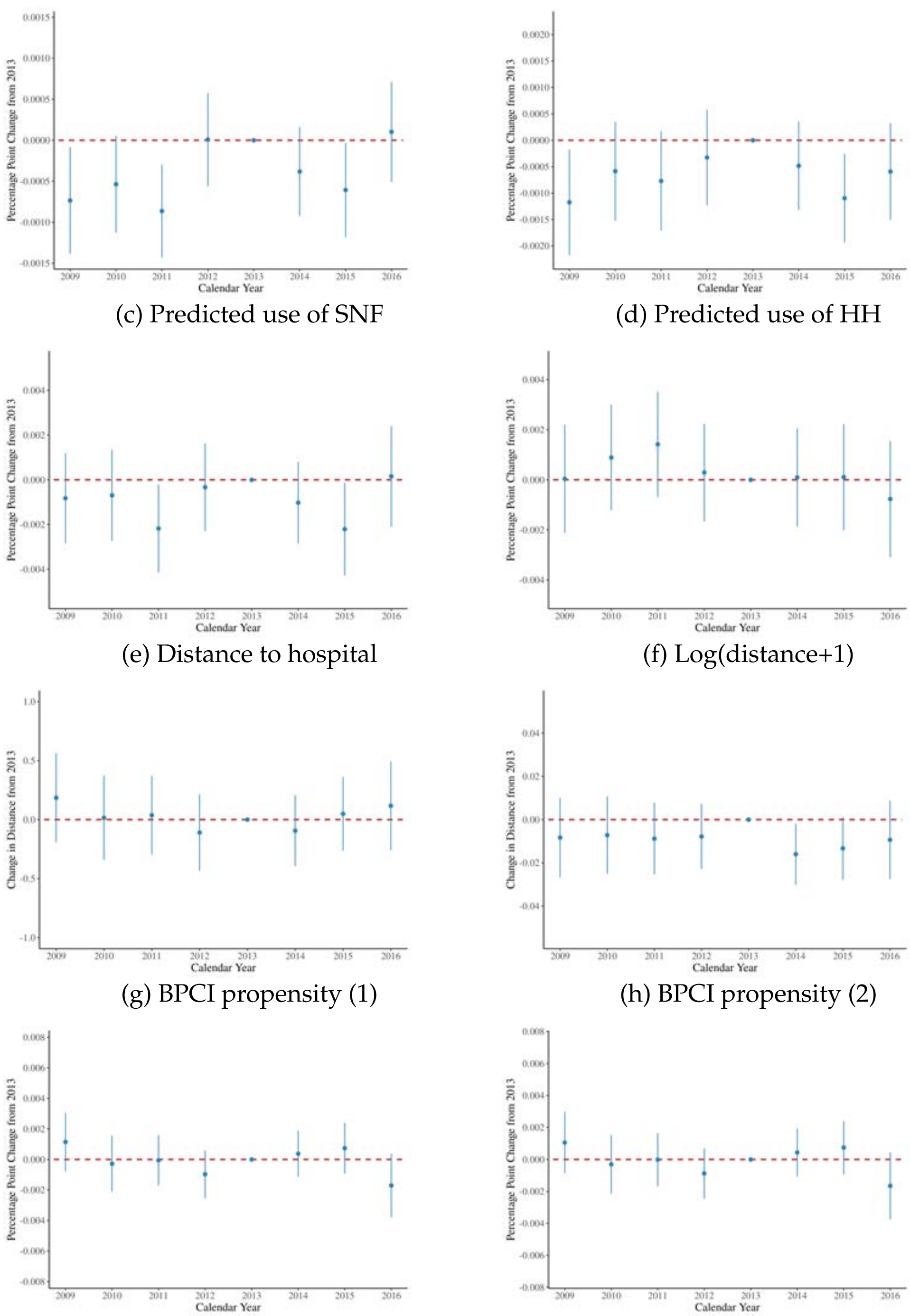


Figure 6: Joint replacement versus other clinical bundles

Note: This figure presents the estimated dynamic reduced form effects (i.e., comparing hospitals with low vs. high potential for gain) on 90-day episode spending for lower extremity joint replacement (LEJR) in bold red and five other clinical episode bundles (congestive heart failure, pneumonia, chronic obstructive pulmonary disease, sepsis, and urinary tract infections) in light grey. The coefficients are obtained by estimating a reduced form variant of Equation 3 where the participation indicator, d_h, is replaced by the instrument, z_h. We de-mean the coefficients by the 2009–12 cohort-specific mean so that the trends for all conditions are centered around zero in the pre-BPCI period. We do not present confidence intervals to facilitate clarity. The average reduced form coefficients are presented in Table 7 Panel B.



(a) Predicted readmission risk (b) Predicted use of IRF

(c) Predicted use of SNF (d) Predicted use of HH

(e) Distance to hospital (f) Log(distance+1)

(g) BPCI propensity (1) (h) BPCI propensity (2)

Figure 7: Patient selection

Note: This figure presents the estimated reduced form dynamic effects (i.e., comparing hospitals with low vs. high potential for gain) on measures of observed and unobserved patient complexity, as described in Section 5.6.4. The coefficients are obtained by estimating a reduced form variant of Equation 3 where the participation indicator, d_h, is replaced by the instrument, z_h. Panels (a)–(d) present results for selection on observed risk. Predicted risk scores are generated using demographics and historical utilization and spending. Panels (e)–(h) present results that help assess selection on unobserved risk. Distance between the patient and hospital is calculated using 5-digit zipcodes. BPCI propensity is predicted using 3-digit zipcode for Panel (g), while in Panel (h) we also use patient demographics. Section 5.6.4 describes the construction of the propensity measures.

Table 1: Summary statistics

	Never BPCI (1)	Ever BPCI (2)	All Hospitals (3)
A. Market-level			
% MA penetration	25.2	25.3	25.2
Geographic adjustment factor	0.99	1.02	0.99
Hourly wage	35.2	38.0	37.7
B. Hospital-level			
Number of hospitals	2,227	302	2,529
Number of beds	263.0	366.9	275.6
% Teaching hospital	33.2	47.1	34.9
% For-profit	20	16.6	19.6
% Government	16.9	7.2	15.7
Medicare LEJR cases per bed	0.77	0.683	0.759
Total admissions per bed	44.0	48.5	44.5
Medicare ortho. revenue per bed ('000$)	25.9	27.0	26.0
C. Patient-level			
Number of episodes	2,122,262	452,473	2,574,735
Length of index LEJR stay (days)	3.8	3.8	3.8
% using any PAC during episode	80.8	86.0	81.7
% using SNF during episode	41.7	43.8	42.0
% using HH during episode	36.7	38.9	37.1
% using IRF during episode	11.4	13.8	11.8
% readmitted during episode	10.5	11.7	10.7
% with no post-discharge spending	18.2	12.7	17.2
Mean episode spend ('000$)	23.3	25.3	23.7
Mean LEJR reimbursement ('000$)	13.7	14.5	13.8
Mean readmission spend ('000$)	1.4	1.5	1.4
Mean PAC spend ('000$)	8.3	9.2	8.4
Mean SNF spend ('000$)	5.1	5.5	5.1
Mean IRF spend ('000$)	1.8	2.2	1.8
Mean HH spend ('000$)	1.3	1.5	1.3
% Pred. risk of SNF use	40.4	40.3	40.4

Note: This table presents descriptive statistics on the main analysis sample. All spending values are expressed in 2016 dollars. Unless otherwise noted, all descriptive statistics pertain to January 2009 through December 2013, i.e., prior to the introduction of BPCI. Columns 1 and 2 present values for hospitals that do not participate in BPCI through the end of our sample ('never' BPCI) or do participate ('ever' BPCI), respectively. Panel A presents evidence on relevant market-level attributes. MA denotes Medicare Advantage, the portion of Medicare where beneficiaries receive benefits through private insurance plans. Panel B presents evidence on relevant hospital-level factors. Since participants are substantially larger in bed size, we present volume and revenue measures per bed. Panel C presents evidence on relevant aspects of episodes. LEJR denotes lower extremity joint replacement, the surgery that we focus on in this paper. PAC denotes post-acute care, SNF skilled nursing facility, HH home health care, and IRF inpatient rehabilitation facility. We generate predicted risk of SNF use as described in Section 3.2.

Table 2: Balance tests

	$z_h \cdot T_t$ (1)	Overall trend (2)
Total volume per bed	0.0158	-0.373***
	(0.221)	(0.066)
Ortho. admissions per bed	-0.001	-0.034***
	(0.010)	(0.006)
Ortho. revenue per bed	220	432***
	(293)	(81.9)
LEJR volume per bed	0.002	-0.010***
	(0.006)	(0.003)
MA penetration (%)	0.008	1.25***
	(0.022)	(0.020)
Predicted risk of SNF use (%)	-0.022	-0.233***
	(0.051)	(0.018)

Note: This table reports results from formal tests of differential trends in the pre-BPCI period (2008–2013) on six time-varying hospital, market, and patient attributes as outcomes. The same outcomes are also studied in Figure 3. Column 1 reports differential linear trends for hospitals that had high potential for financial gain. The coefficients in Column 1 corresponds to θ from Equation 7. Column 2 reports the coefficient γ corresponding to the overall linear trend across hospitals from the following model: $C_{ht} = \alpha_h + \gamma T_t + \varepsilon_{ht}$. Medicare Advantage (MA) penetration is computed as described in Section 3.2. We use predicted values from a probit model (fit on 2008 data) for each patient admitted with LEJR as our measure of predicted risk of SNF use. The sample for predicted risk of SNF use begins in 2009 in order to avoid reusing the 2008 cases used to estimate the probit model. All models are estimated on aggregated hospital-year level data, except for predicted SNF risk, which is estimated on patient-level data.

Table 3: Episode spending

	Log(spending)$_{iht}$				ΔLog(spending)$_h$	
	(1)	(2)	(3)	(4)	(5)	(6)
A. DD						
BPCI × Post	-0.032^{***}	-0.031^{***}	-0.032^{***}	-0.031^{***}	-0.041^{***}	-0.028^{***}
	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)
B. LATE						
BPCI × Post	-0.116^{***}	-0.112^{***}	-0.114^{***}	-0.108^{***}	-0.131^{***}	-0.126^{***}
	(0.038)	(0.037)	(0.038)	(0.037)	(0.034)	(0.031)
First stage	0.125^{***}	0.125^{***}	0.124^{***}	0.124^{***}	0.195^{***}	0.174^{***}
	(0.030)	(0.030)	(0.030)	(0.030)	(0.040)	(0.035)
F-statistic	17.7	17.7	17.4	17.4	24.3	25.3
Observations	2,574,735	2,574,735	2,572,788	2,572,788	2,437	2,437
$\bar{y}$	25.346	25.346	25.346	25.346	-0.084	-0.084
Patient Controls	N	Y	N	Y	N	Y
Hospital Controls	N	N	Y	Y	N	N

Note: This table presents estimated effects on log episode spending, our main outcome of interest. Panels A and B present the DD and IV coefficients, respectively. Patient controls include age, gender, race, disability status, and indicators for co-morbidities used to compute Elixhauser score. Hospital time-varying controls were sourced from the AHA survey and include beds, admissions, Medicare admission share, Medicaid admission share, Orthopedic share of medicare inpatient revenue, and an ACO participation indicator. We lose 1,724 observations when including hospital controls due to hospitals not observed in the AHA survey. Columns 1–4 present results from models estimated on patient-level data, while column 5–6 present results from models estimated on cross-sectional data. Standard errors are clustered by hospital in Columns 1–4 and are robust to heteroskedasticity in Columns 5 and 6. The outcome for the cross-sectional analysis is the long difference in mean log episode spending between 2014–16 and 2009–13. The cross-sectional sample drops 92 never-BPCI hospitals which do not perform LEJR surgeries post-2013. In the cross-sectional regression models, hospitals are weighted by pre-BPCI LEJR volume (2009–2013). The models in Columns 2 and 6 represent our preferred specifications. The mean values in Columns 1-4 present mean episode spending in thousands of dollars. In Columns 5 and 6 they present the mean long difference in log episode spending.

Table 4: Sensitivity to using alternate instruments

	Baseline (1)	Observed ortho. (2)	Observed LEJR (3)	Exp. gain > median (4)	Exp. gain (5)
1(BPCI)	-0.126^{***}	-0.141^{***}	-0.219^{***}	-0.179^{***}	-0.091^{***}
	(0.0308)	(0.0438)	(0.0653)	(0.0368)	(0.0289)
	[-0.207,-0.081]	[-0.282,-0.076]	[-0.479,-0.136]	[-0.275,-0.124]	[-0.139,-0.027]
First Stage	0.174^{***}	0.138^{***}	0.13^{***}	0.133^{***}	0.0318^{***}
	(0.0347)	(0.0353)	(0.0375)	(0.0232)	(0.00379)
F-Statistic	25.289	15.369	11.995	32.646	70.655
Observations	2,437	2,437	2,437	2,149	2,149
Z 1st	0	0	0	0	0.151
Z Mean	0.242	0.216	0.222	0.519	1.66
Z 99th	1	1	1	1	10.5

Note: This table presents the sensitivity of the BPCI per-episode-spending effect and first stage strength to using different formulations of the instrument for participation. Each column presents 2SLS estimates from a hospital level regression of change in log spending on an indicator for BPCI participation using a different measure of high potential financial gain as the excluded instrument. All regressions are weighted by hospitals' pre-BPCI volume (2009–2013). The spending outcome used in all columns is the long difference between average 2009–2013 and 2014–2016 log episode spending. All regressions control for the patient characteristics included in the baseline specification presented in Table 3 Column 6. Robust standard errors are presented in parentheses. We also present in square brackets weak instrument robust Anderson-Rubin 95% confidence intervals obtained using the STATA command "twostepweakiv" recommended by Andrews, Stock, and Sun (2019). Z – p1, Z – Mean, and Z – p99 denote the first percentile, mean, and 99^{th} percentile of each instrument, respectively. The instruments are described in detail in Section 5.1. Columns 4 and 5 use measures of expected financial gain, calculated as described in Section 5.1. These models have fewer observations since we limit the sample to hospitals with at least 25 LEJR cases. Column 4 presents results using an indicator for above median expected financial gain. Column 5 uses the expected financial gain value in $100,000 units.

Table 5: Utilization of care

	Index stay		PAC use			
	log(LOS) (1)	LOS ≤ 2 (2)	Any PAC (3)	HH (4)	IRF (5)	SNF (6)
A. DD						
BPCI × Post	-0.042^{***}	0.055^{***}	-0.017^{**}	0.021^{**}	-0.011^{**}	-0.034^{***}
	(0.010)	(0.010)	(0.008)	(0.009)	(0.005)	(0.007)
B. LATE						
BPCI × Post	-0.105^{*}	0.159^{**}	-0.131^{**}	-0.120^{*}	-0.090^{**}	0.021
	(0.062)	(0.073)	(0.051)	(0.064)	(0.038)	(0.048)
Observations	2,574,735	2,574,735	2,574,735	2,574,735	2,574,735	2,574,735
$\bar{y}$	3.813	0.22	0.860	0.389	0.138	0.438

Note: This table presents the estimated effects on utilization of care obtained using the primary specification, discussed in Section 4, and corresponding to the coefficients in Table 3 Col. 2. Columns 1 and 2 present effects on outcomes related to length of stay (LOS) for the index LEJR surgery. Column 3 presents effects on any use of post-acute care (PAC) in the 90 days following discharge. Columns 4–6 present the corresponding effects on the use of home health (HH), inpatient rehabilitation (IRF), and skilled nursing (SNF), respectively. Panel A presents DD coefficients, while Panel B presents the LATE coefficients. Patient controls include age, gender, race, disability status, and indicators for co-morbidities used to compute Elixhauser score. Standard errors are clustered by hospital and presented in parentheses.

Table 6: Quality of care

				CMS complications measure		
	90-day Readmissions (1)	90-day Mortality (2)	90-day Revisions (3)	Composite (4)	Joint Infection (5)	Mechanical (6)
A. DD						
BPCI × Post	-0.003^{*}	-0.001	0.0001	-0.0004	0.0001	-0.0004
	(0.002)	(0.001)	(0.0004)	(0.001)	(0.001)	(0.0004)
B. LATE						
BPCI × Post	-0.003	-0.004	0.008^{**}	0.005	0.007	0.003
	(0.012)	(0.005)	(0.003)	(0.006)	(0.005)	(0.003)
Observations	2,574,735	2,574,735	2,537,969	2,574,735	2,574,735	2,574,735
$\bar{y}$	0.101	0.028	0.008	0.042	0.021	0.008

Note: This table presents estimated effects on miscellaneous outcomes related to quality of care using the primary specification described in Section 4, which corresponds to Column 2 in Table 3. For 90-day revisions, we drop episodes where a subsequent joint replacement (469 or 470) occurs during the 90 days after discharge. Readmissions, presented in Column 1, include revision surgeries. Revision surgeries are defined using DRGs 466, 467, and 468. Complication measures are constructed following CMS documentation, as described in Section A.1. Panel A presents DD coefficients, while Panel B presents the LATE coefficients. Patient controls include age, gender, race, disability status, and indicators for co-morbidities used to compute Elixhauser score. Standard errors are clustered by hospital and presented in parentheses.

Table 7: Effect on spending for LEJR versus other bundles

	LEJR (1)	CHF (2)	PNA (3)	COPD (4)	Sepsis (5)	UTI (6)	Pooled (7)
A. DD							
BPCI × Post	-0.031^{***}	-0.006	-0.009	-0.004	-0.008	-0.019^{*}	-0.006^{*}
	(0.005)	(0.006)	(0.006)	(0.007)	(0.005)	(0.010)	(0.004)
B. LATE							
Reduced Form	-0.014^{***}	0.0002	-0.001	0.016^{***}	0.002	0.010^{*}	0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.005)	(0.003)
First stage F-stat.	17.7	12.2	3.2	3.0	0.9	1.1	5.6
Observations	2,574,735	2,103,288	2,388,107	1,663,846	2,424,028	1,270,679	9,849,948
Mean BPCI × Post	0.067	0.036	0.022	0.021	0.029	0.016	0.026
N. BPCI Hospitals	302	174	136	130	119	88	238
$\bar{y}$	22,872	19,155	18,366	14,899	24,613	17,590	19,386

Note: This table presents estimated effects on log episode spending for LEJR (Column 1) and five other clinical bundle episodes included in the BPCI program (Columns 2–6). These are, respectively, congestive heart failure (CHF), pneumonia (PNA), chronic obstructive pulmonary disease (COPD), sepsis, and urinary tract infections (UTI). These are the largest bundles in BPCI by hospital participation after LEJR. Column 7 presents results on a pooled sample across these 5 bundles. Panel A presents DD coefficients, while Panel B presents the reduced form coefficients and corresponding first stage f-statistics. Models include patient controls, hospital, and HRR by year fixed effects. Patient controls include age, gender, race, disability status, and indicators for co-morbidities used to compute Elixhauser score. Standard errors are clustered by hospital and presented in parentheses.

A Appendix

A.1 Sample and variable construction

Our primary data source is encounter-level data from the Medicare Provider Analysis and Review (MedPAR) files from 2006–2016, which contains all hospital and institutional post-acute care stays for fee-for-service Medicare beneficiaries. We link these hospital stays with additional beneficiary demographic information in the Master Beneficiary Summary File (MBSF) and fee-for-service claims from home health agencies (we have access to home health agency claims from 2007–2016). The claims data include beneficiary demographics (age, race, sex), enrollment information, death dates, and zip code of residence, as well as a record of medical services. Information about medical services includes provider identity, spending, encounter date, diagnosis, and procedure codes (International Classification of Diseases).

We remove all claims in the MedPAR and HHA files that are denied or rejected, have no covered charges, have zero Medicare payment, or involve a service for which the beneficiary left against medical advice. We also remove a small number of claims in which the admission date occurs after the discharge date or the beneficiary's death date. We exclude hospitals located in Maryland because they are not paid based on DRGs and those located in US territories.

To define episodes, we first identify candidate index hospitalizations in acute care hospitals paid under the inpatient prospective payment system (IPPS). To qualify as an index stay, we require the hospital stay:

1. Is paid as an IPPS claim.
2. Does not involve a primary payer other than Medicare.
3. Is not a hospital stay for more than 365 days.
4. Does not involve a beneficiary enrolled in Medicare for end-stage renal disease or enrolled in Medicare Advantage.
5. Is not a transfer from another inpatient hospital. In other words, when a transfer is involved, index hospitalizations are the first hospital in which the beneficiary receives care.
6. Have a beneficiary and hospital that can both be assigned a hospital referral region (HRR) by zip code.

We then impose several additional requirements for the inclusion of episodes in the analysis. To reliably track spending and comorbidities, we require beneficiaries to be enrolled in Medicare Parts A and B from 365 days before admission until death or 90 days post-discharge. We remove episodes if beneficiaries are enrolled in Medicare Advantage at any point from admission until 90 days post-discharge. We also remove episodes if any claims in the 90 days post-discharge involve a primary payer other than Medicare. To ensure that we observe adequate follow-up in both the MedPAR file and home health agency claims, we remove any episodes initiated after 9/30/2016.

Our main analysis focuses on lower extremity joint replacement (LEJR) episodes defined by MS-DRG 469 or 470.[24] In placebo tests, we additionally analyze episodes for patients admitted for cholecystectomy and a set of other conditions with substantial participation in BPCI. Cholecystectomy is defined by MS-DRGs 411–419; the BPCI conditions are defined by MS-DRGs 291–293 (congestive heart failure), 177–179 and 193–195 (pneumonia), 190-192 and 202–203 (chronic obstructive pulmonary disease), 870–872 (sepsis), 689–690 (urinary tract infections) and 483–484 (major joint replacement of the upper extremity).

In the main analytic sample, we further require that the beneficiaries be over 65 years old at the beginning of the episode. The treated hospitals are those that participated in BPCI Model 2. Only 77 hospitals participated in Models 1 and 4, of which 51 subsequently switched to Model 2 and are included in our sample. The remaining 26 are retained as comparison hospitals. Model 3 targeted post-acute care providers directly. Finally, we require that index hospitals eligible for inclusion have adequate pre-period data to reliably estimate average predicted post-discharge spend, defined as 25 LEJR cases in 2006–2007. We drop episodes at the remaining low-volume hospitals.

We use diagnosis codes from claims occurring in the year prior to index admission in order to determine beneficiaries' Elixhauser comorbidities. We determine whether admissions for LEJR involve a hip or femur fracture by examining the set of diagnosis codes used by CMS for the comprehensive care for joint replacement program (CJR).[25] In analyses of revisions to joint replacements, we define joint revision using DRG 466, 467, and 468.

We define complications following the definitions stipulated in the quality measure defined in National Quality Forum #1550, "Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)." The documentation is available from the NQF here. Within the documentation, the report, "2016 Procedure-Specific Measure Updates and Specifications Report Hospital-Level Risk-Standardized Complication Measure" lists the ICD9 diagnosis and procedure codes used to identify mechanical and prosthetic infection complications. Specifically, the codes are listed in Table D.1.10 on pages 47-48 of the report. We define an all-cause measure, "composite complications", which is an indicator set to 1 if a complication is detected for any of the reasons listed in Table D.1.10: heart attack, pneumonia, sepsis, surgical site bleeding, pulmonary embolism, death, and mechanical and infection complications.

24. For discharges occurring prior to 10/01/2007, LEJR was categorized as DRG 544.
25. The diagnosis codes are available from CMS here.

A.2 Testing for selection on historic trends

An evergreen concern with the design of voluntary incentive programs is that the firms that choose to join may receive a payoff under the new contract without making incremental efforts to fulfill the objective of the program. This is inefficient since the regulator or payer makes incentive payments without changing behavior. Einav et al. (2022) call this "selection on levels" and study it in another voluntary program called Comprehensive Joint Replacement (CJR).

This concern is also present in the case of BPCI, although the contract design mitigates the concern quite substantially. As discussed in Section 2.2, BPCI sets the target episode spending level for a hospital by projecting its historic spending level using the national average trend in spending. In theory, this introduces the possibility of arbitrage for hospitals that already enjoyed a steeper decline in spending than the national average during the baseline period. If they can maintain their historic performance improvements, then they will receive a payoff under BPCI without incremental efforts. We empirically test whether such hospitals were more likely to participate in BPCI. We perform the following steps.

- We restrict the sample to hospitals that have at least 10 Medicare FFS LEJR patients in each year from 2009 to 2013. This restriction is necessary since we will compute mean episode spending by hospital-year and use it to then compute the trend in spending. Very small samples could lead to measurement error and bias. This restriction drops 95 hospitals, but none of them are BPCI participants. The sample now has 2420 hospitals.

- We compute the compounded average annual change in mean episode spending for each hospital between 2010-11 (July-June) and 2012-13. Similarly, we also compute the national average annual change in mean episode spending during the same period.

- We use the national trend to project each hospital's mean episode spending value in 2013 to predict their target price in 2016. We also apply the 2% discount in this step, following the same formula used by CMS. We choose 2016 since it is the last year of our sample; however, the results are similar if we choose a different year.

- We also predict the hospital's expected spending level in 2016 assuming that it continues its own historic trend. The main differences between this expected spending level and the target price are 1) the use of the hospital's own trend instead of the national trend to project, and 2) we do not apply a discount.

- We then calculate the "expected gain" for each hospital in 2016 = target price - expected spending level. Positive values imply that the hospital is expected to make a profit if it continues on its historic trajectory. We assign hospitals to quartiles based on this variable, after dropping 48 hospitals with outlier values < 1 and > 99 percentiles, of which one is a BPCI participant. The mean expected gain in the top quartile is $2,747, while that in the bottom quartile is -$6,116. 1424 hospitals are expected to lose money while about 950 hospitals are predicted to have positive gain under BPCI.

We now have the necessary ingredients to formally test whether hospitals expected to receive a payoff under BPCI are also more likely to join the program. Figure A.1 Panel (a) presents a bar graph showing the fraction of hospitals in each quartile of expected gain that participate in BPCI at the end of our sample period. The red dashed line indicates the fraction of participating hospitals in the full sample. The figure suggests that the proportion of participants is quite stable across quartiles. We formally test this using an OLS regression model in which we regress indicators for quartiles of expected gain on an indicator for participation. The coefficients are insignificant individually and jointly. The *p*-value from the joint test of significance is 0.80. Hence, hospitals with more favorable historical spending trends are not more likely to join BPCI.

We also test whether our instrument, the indicator for high potential financial gain, is correlated with having a more favorable historical spending trend. Such a correlation could raise doubts about the conditional randomness of the instrument. We perform an equivalent analysis to test this empirically. Figure A.1 Panel (b) presents the corresponding bar graph. The Y-axis now plots the fraction of hospitals in each quartile of expected gain (based purely on historical trend in episode spending) that are tagged as having a high potential financial gain according to their LEJR volume and mean episode cost in 2007. Again, the fractions in each quartile are quite stable, with Q2 showing a slightly greater fraction than the others. The joint test of significance of coefficients on quartile indicators indicates that they are insignificant with a p-value of 0.39. The coefficient for Q2 is individually significant at the 10% level.

We also perform an alternative analysis in which we create a flag for the expected gain > 0 and regress this indicator on an indicator for participation or high potential for financial gain. In both cases, the coefficients are small and statistically insignificant. Taking this evidence collectively, we are reassured that hospitals did not sort into BPCI based on their historical trend in spending, and the historical trend is also not correlated with the instrument.

A.3 Construction of the instrumental variable

To capture baseline exposure to financial incentives in BPCI, we create two predicted values. The first, predicted post-discharge spending, is a proxy for higher episode spending in the pre-BPCI period. The second, predicted revenue, is a proxy for a higher LEJR volume in the pre-BPCI period. To create these predicted values, we use data from all orthopedic episodes during the 2007 calendar year. We choose to construct baseline values long before the intervention and use the full set of orthopedic episodes to limit the concern that predicted values are correlated with unobserved shocks or mean reversion in LEJR spending.

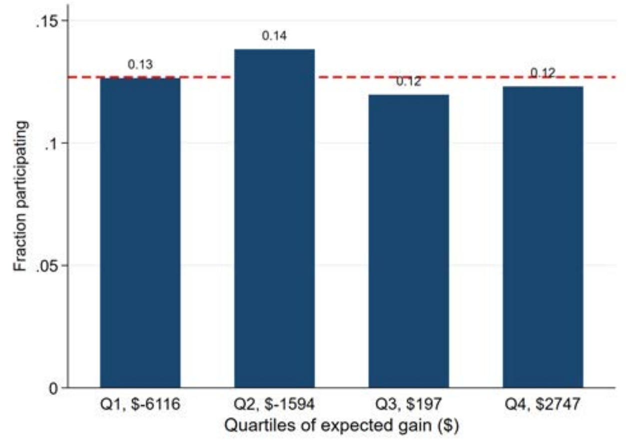
To create the predicted post-discharge spending measure, we use patient-level data to estimate linear regressions of observed post-discharge spending on hospital and patient attributes. The hospital attributes include beds, a teaching status indicator, a residency indicator, control type, and HRR fixed effects. Patient attributes include age, sex, race, Elixhauser comorbidities, disability indicator, and DRG fixed effects. Finally, we use the predicted values from this regression and create an average for each hospital. We use the hospital average as the predicted post-discharge spending measure.

To create the predicted revenue measure, we use hospital-level data to estimate unweighted linear regression models of orthopedic revenue on hospital attributes. The hospital attributes include beds, a teaching status indicator, a residency indicator, control type, core-based statistical area (CBSA) type (urban/rural), AHA primary service code in 2007, and non-hip and knee surgical volume. We also use AHA data to create an indicator equal to 1 if the hospital ever reports having orthopedic services and an indicator equal to 1 if the hospital ever is a surgical or orthopedic specialty hospital. Our predicted revenue measure for each hospital is the predicted value from this regression.

As discussed in Section 4.2.1 and shown visually in Figure 2, we find that participation is much more likely for the set of hospitals with high predicted revenue and predicted post-discharge spending. Consequently, we combine these two variables into a single binary exposure to BPCI incentives, which is equal to 1 if a hospital is above the median in both measures.

Additional figures and tables

(a) Participation in BPCI



(b) High potential for gain

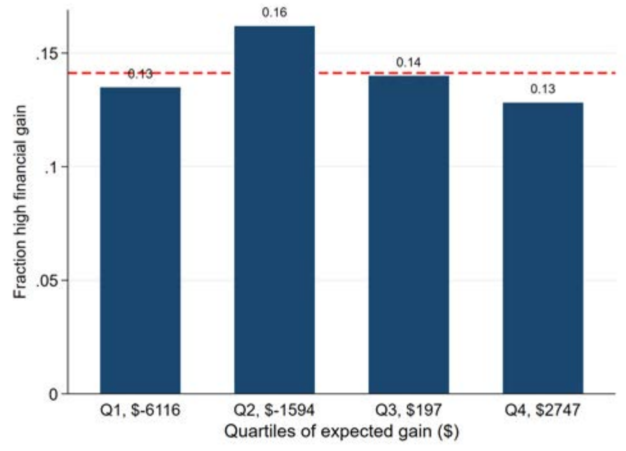
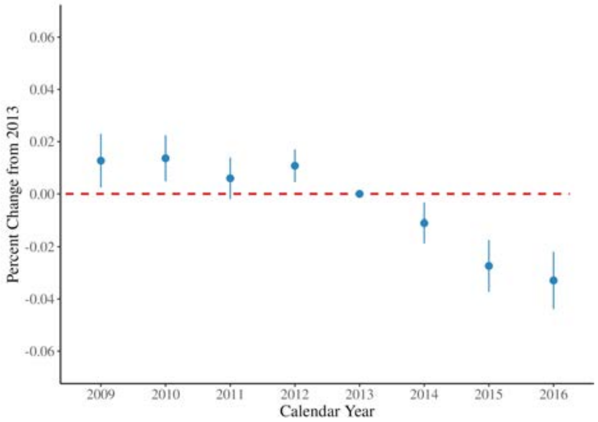


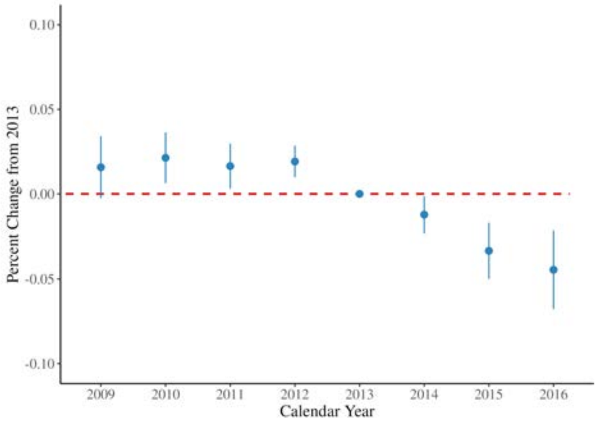
Figure A.1: Participation by expected gain based on historic trends

Note: This figure presents evidence on selection of hospitals based on their historic trends in episode spending. Section A.2 describes in detail how we compute the expected financial gain for each hospital in 2016, based purely on continuing their historical spending trend. Negative values of expected gain imply that the hospital's expected episode spending will exceed the target price under BPCI and it will lose money on average. We drop hospitals with $<$ 1% and $>$ 99% outlier values of historic spending trend from the sample and then assign hospitals to quartiles of expected gain. Hospitals in Q4 are predicted to gain the most and, those in Q1 to lose the most under BPCI if they continue on their historical spending trend. Panel (a) plots the share of hospitals in each quartile that participate in BPCI. The red dashed line indicates the share of hospitals participating in BPCI in the sample. A linear regression with quartile indicators finds that the coefficients are individually and jointly insignificant. The p-value from the joint test is 0.80. Panel (b) plots the share of hospitals in each quartile with high potential for gain, our instrument. The red dashed line indicates the share of hospitals with high potential financial gain in the sample. A linear regression with quartile indicators finds that all coefficients except for Q2 are individually insignificant. The coefficient for Q2 is significant at the 10% level. The joint test of significance fails to reject the null hypothesis with a p-value of 0.39. The plots mention the mean expected gain values within each quartile.

(a) Episode Spend (log)



(b) LEJR length of stay (log)



(c) Use of any PAC

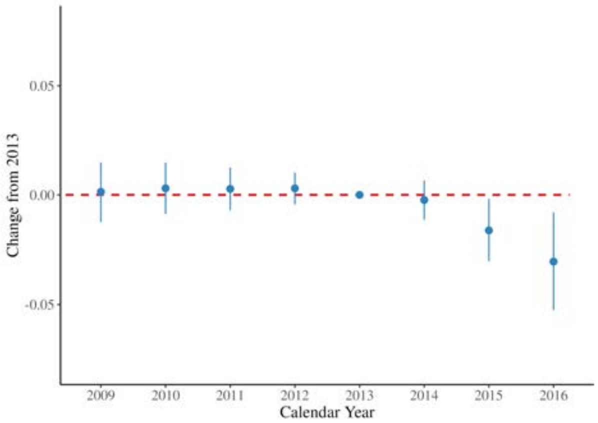
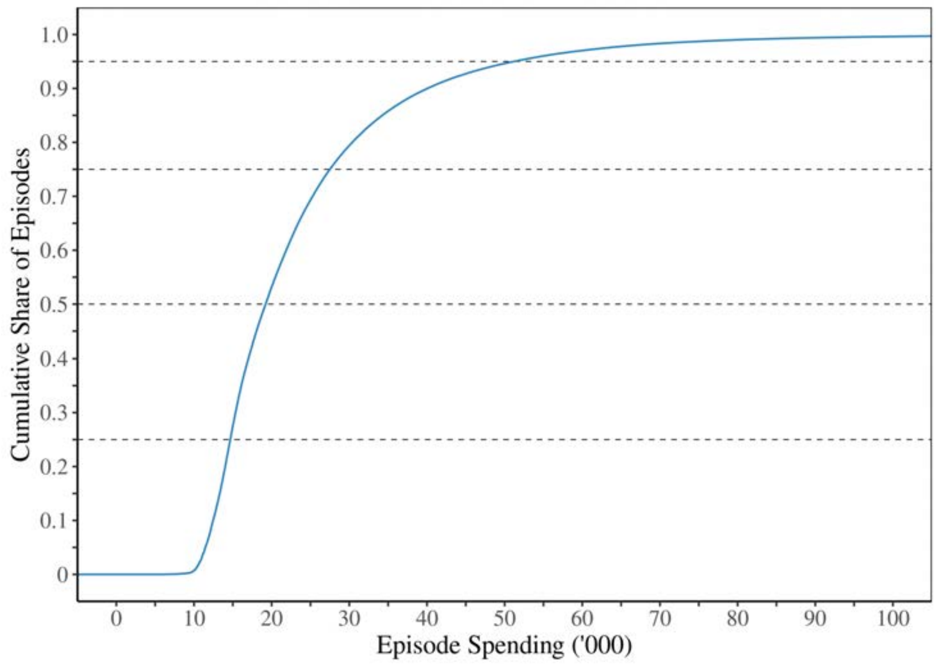


Figure A.2: Effects on spending and utilization (DD)

Note: This figure presents results on the dynamic effects estimated using the DD model (i.e., comparing hospitals participating in BPCI to nonparticipants) on the spending and utilization outcomes examined in Figure 4. The coefficients are obtained by estimating Equation 3. The error bars denote 95% confidence intervals. Standard errors are clustered by hospital.

(a) Episode spending



(b) Index surgery length of stay

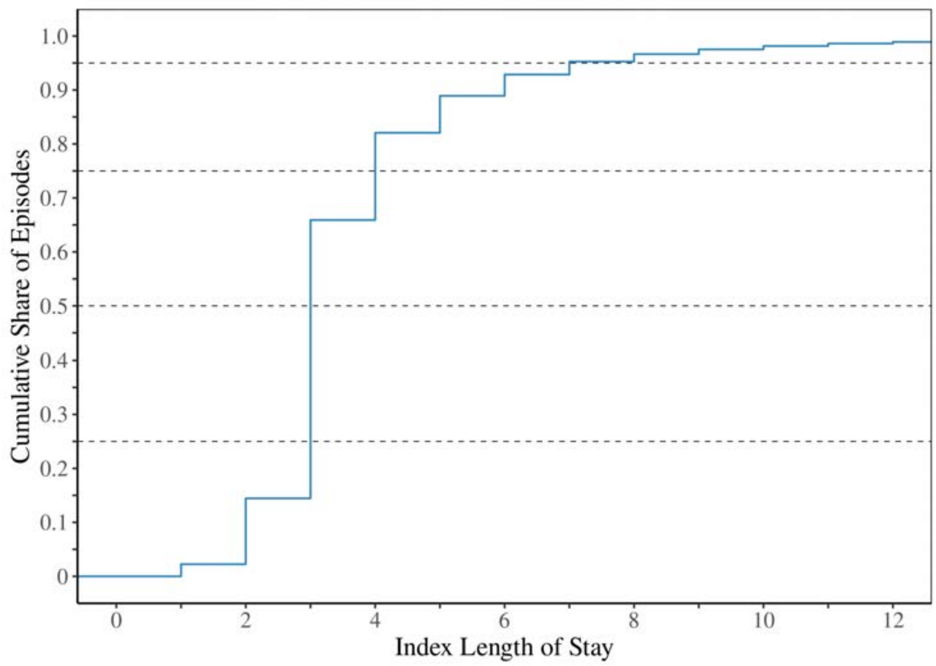


Figure A.3: Empirical CDFs in the pre-BPCI period

Note: This figure presents the cumulative distribution of episode spending (Panel a) and length of stay for LEJR surgeries (Panel b) during the 2011 and 2012 calendar years.

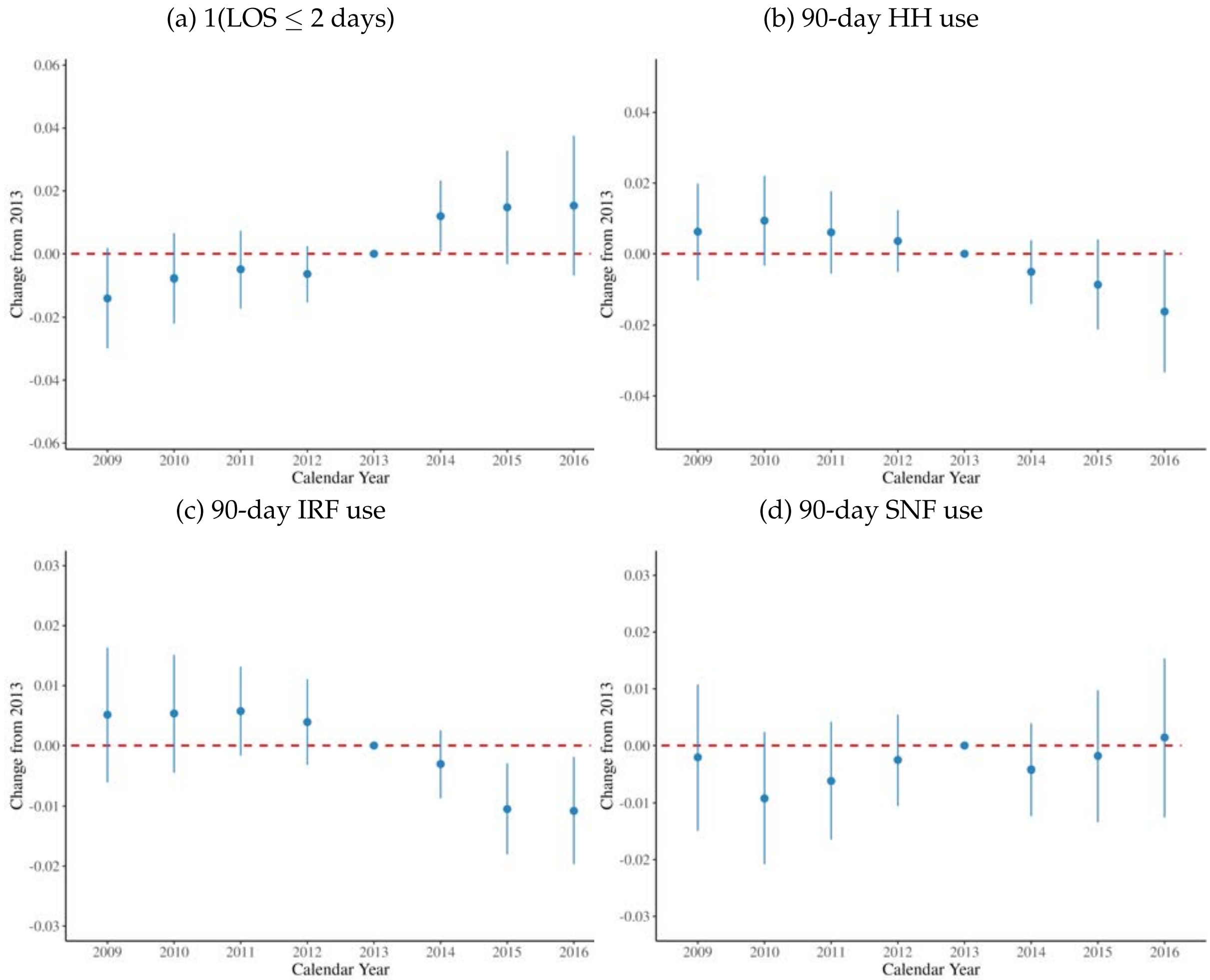


Figure A.4: Effects on utilization of care

Note: This figure presents results on the dynamic reduced form effects (i.e., comparing hospitals with low vs. high potential for gain) on index stay length and the use of post-acute care. The coefficients are obtained by estimating a reduced form variant of Equation 3 where the participation indicator, d_h, is replaced by the instrument, z_h. Panel (a) presents effects on the probability that length of stay is less than or equal to 2 days. Panels (b)–(d) present corresponding plots on the use of different types of post-acute care within 90 days following discharge from the LEJR surgery. The models include hospital and market by year fixed effects and patient controls. Standard errors are clustered by hospital. For reference, the cumulative distribution of length of stay is presented in Figure A.3b.

(a) Episode spending (b) 90-Day any PAC use

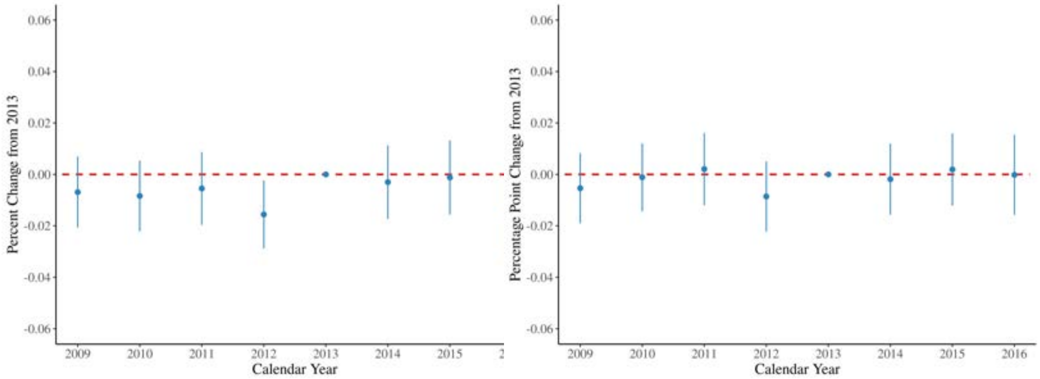


Figure A.5: Effect on cholecystectomy spending

Note: This figure presents the estimated dynamic reduced form effects (i.e., comparing hospitals with low vs. high potential for gain) on 90-day episode spending and PAC use in cholecystectomy episodes. The models include patient controls, hospital, and HRR by year fixed effects. Standard errors are clustered by hospital.

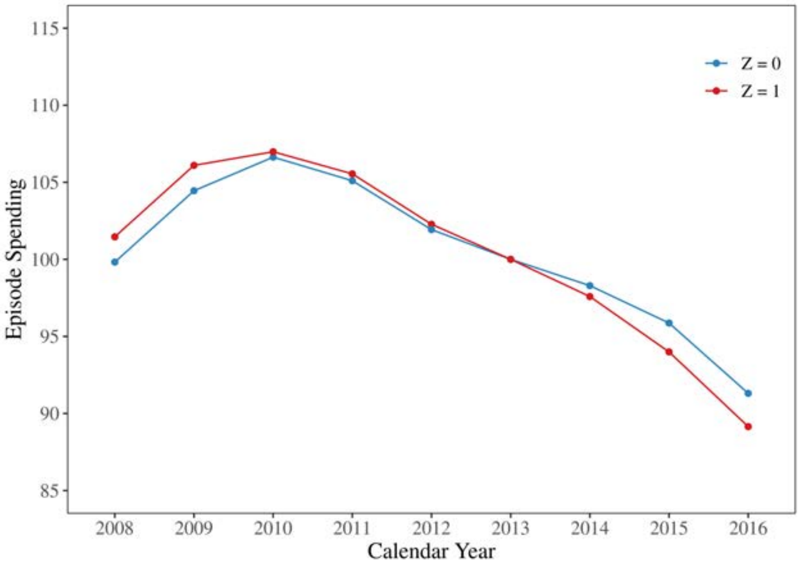


Figure A.6: Unadjusted trends in episode spend

Note: This figure presents trends in overall episode spending for LEJR cases among Medicare fee-for-service patients. It shows spending trends for hospitals grouped by the instrument, z_h, which is an indicator for high potential financial gain. Spending (in 2016 dollars) within each group is normalized such that 2013 spending equals 100.

Table A.1: Complier versus participant hospitals

	Full sample	Compliers		Treated	
Sub-Group	(1) Proportion of sample	(2) Proportion of sample	(3) Relative to full	(4) Proportion of sample	(5) Relative to full
A. Hospital characteristics					
Non-teaching	0.543	0.386	0.710	0.415	0.764
Teaching	0.457	0.614	1.345	0.585	1.281
Non-profit	0.716	0.722	1.009	0.826	1.154
For-profit	0.161	0.205	1.269	0.123	0.765
Government	0.123	0.064	0.519	0.050	0.411
Standalone hospital	0.277	0.299	1.080	0.204	0.737
2-10 Hospital system	0.357	0.383	1.072	0.396	1.109
$>$ 10 Hospital system	0.366	0.310	0.847	0.400	1.092
Beds: $<$ median	0.500	0.705	1.411	0.438	0.876
Beds: $>$ median	0.500	0.295	0.589	0.562	1.124
B. Risk in patient population					
Average age: $<$ median	0.501	0.641	1.280	0.524	1.048
Average age: $>$ median	0.499	0.359	0.719	0.476	0.952
Pred. readm. risk: $<$ median	0.500	0.694	1.389	0.547	1.094
Pred. readm. Risk: $>$ median	0.500	0.306	0.611	0.453	0.906
Pred. mortality risk: $<$ median	0.500	0.695	1.390	0.528	1.056
Pred. mortality risk: $>$ median	0.500	0.305	0.610	0.472	0.944
Pred. complication risk: $<$ median	0.500	0.689	1.377	0.534	1.067
Pred. complication risk: $>$ median	0.500	0.311	0.622	0.466	0.933
C. Hospital quality					
All Cause readm. rate: $<$ median	0.507	0.484	0.953	0.423	0.834
All Cause readm. rate: $>$ median	0.493	0.522	1.059	0.577	1.171
LEJR readm. rate: $<$ median	0.500	0.367	0.734	0.421	0.842
LEJR readm. rate: $>$ median	0.500	0.633	1.267	0.579	1.158
Mortality rate: $<$ median	0.500	0.591	1.182	0.524	1.047
Mortality rate: $>$ median	0.500	0.409	0.817	0.476	0.953
Complication rate: $<$ median	0.500	0.646	1.290	0.546	1.091
Complication rate: $>$ median	0.500	0.354	0.709	0.454	0.909

Note: This table presents patient-weighted shares of hospitals with certain attributes of interest in the full sample (col. 1), among compliers (col. 2), and among all BPCI participants (col. 4). Columns 3 and 5 present the likelihood of a hospital in the complier and treated groups, respectively, of having an attribute *relative* to the average hospital in the full sample. The proportion of each sub-group in the complier sample is unobserved and is estimated following the approach proposed in Abadie (2003) using data from the pre-BPCI period (2009-2013). Panel B presents the proportion of (patient-weighted) hospitals in each group that are below or above the median complexity on a number of predicted measures besides age. Panel C examines baseline performance on quality of care using a range of risk-adjusted measures, such as mortality and readmission rates.

Table A.2: Correlation of episode spending across clinical bundles

	LEJR	CHF	COPD	PNA	SEPSIS	UTI
A. Total Episode Spending						
LEJR	1	–	–	–	–	–
CHF	0.787	1	–	–	–	–
COPD	0.763	0.877	1	–	–	–
PNA	0.764	0.898	0.880	1	–	–
SEPSIS	0.760	0.872	0.856	0.889	1	–
UTI	0.731	0.844	0.836	0.875	0.838	1
B. Post-Discharge Spending						
LEJR	1	–	–	–	–	–
CHF	0.647	1	–	–	–	–
COPD	0.602	0.798	1	–	–	–
PNA	0.634	0.846	0.815	1	–	–
SEPSIS	0.562	0.763	0.728	0.825	1	–
UTI	0.592	0.777	0.756	0.820	0.737	1

Note: This table presents the within-hospital correlation in episode spending across different clinical episodes using data from 2010–12, i.e., prior to BPCI. Panels A and B present the correlations for total episode spending and post-discharge spending, respectively.

Table A.3: Effect on cholecystectomy spending

	Log(spending) (1)	Any PAC (2)
A. DD	−0.004	0.0001
	(0.005)	(0.005)
B. LATE	0.014	0.022
	(0.032)	(0.037)
$\bar{y}$	9.53	0.296
Observations	364,980	364,980

Note: This table presents the results from a falsification check: spending effects for a procedure not included under BPCI, cholecystectomy. We present estimated effects on log episode spending and post-acute care use at hospitals participating in the LEJR bundle using the primary specification. Panel A presents DD coefficients, while Panel B presents the LATE coefficients. Standard errors are clustered by hospital and presented in parentheses.

Table A.4: Differential effects at large hospitals

	$\text{Log(spending)}_{iht}$	$\Delta\text{Log(spending)}_h$
	(1)	(2)
A. DD		
Baseline		
BPCI × Post	-0.031^{***}	-0.028^{***}
	(0.005)	(0.005)
Alternate		
BPCI × Post	-0.030^{***}	-0.025^{***}
	(0.005)	(0.005)
Abv. mdn rev. × Post	-0.011^{***}	-0.023^{***}
	(0.003)	(0.004)
B. LATE		
Baseline		
BPCI × Post	-0.112^{***}	-0.126^{***}
	(0.037)	(0.031)
First stage	0.125^{***}	0.174^{***}
	(0.030)	(0.035)
Alternate		
BPCI × Post	-0.102^{***}	-0.108^{***}
	(0.044)	(0.033)
Abv. mdn rev. × Post	-0.004	-0.013^{***}
	(0.005)	(0.005)
First stage	0.107^{***}	0.156^{***}
	(0.031)	(0.036)
Observations	2,574,735	2,437
$\bar{y}$	9.908	-0.084
Mean BPCI × Post	0.067	–
Mean Ever BPCI	–	0.172
Patient Controls	Y	Y
Hospital Controls	N	N

Note: This table presents the results from a robustness check including an interaction term in the baseline specification. The term is the interaction of an indicator for above-median total hospital revenue and an indicator for post-2013. Hospital revenue is sourced from the Medicare cost reports. We compute the mean revenue during 2009–2013 for each hospital. Column 1 here presents results corresponding to the model in Table 3 Column 2. Column 2 here presents results corresponding to the model in Table 3 Column 6. Panel A presents the DD results and Panel B presents the 2SLS results. The top row in each panel presents the baseline coefficients from Table 3 to ease comparisons. Standard errors are clustered by hospital and presented in parentheses.

Table A.5: Differential effects at hospitals with favorable historic trend

	$\text{Log(spending)}_{iht}$	$\Delta\text{Log(spending)}_{h}$
	(1)	(2)
A. DD		
Baseline		
BPCI × Post	-0.031^{***}	-0.028^{***}
	(0.005)	(0.005)
Alternate		
BPCI × Post	-0.030^{***}	-0.027^{***}
	(0.005)	(0.005)
Hospital Spending Trend × Post	-0.019^{***}	-0.026^{***}
	(0.003)	(0.003)
B. LATE		
Baseline		
BPCI × Post	-0.112^{***}	-0.126^{***}
	(0.037)	(0.031)
First stage	0.125^{***}	0.174^{***}
	(0.030)	(0.035)
Alternate		
BPCI × Post	-0.115^{***}	-0.127^{***}
	(0.038)	(0.031)
Hospital Spending Trend × Post	-0.017^{***}	-0.025^{***}
	(0.004)	(0.004)
First stage	0.123^{***}	0.173^{***}
	(0.030)	(0.035)
Observations	2,571,479	2,397
$\bar{y}$	9.894	-0.084
Mean BPCI × Post	0.067	–
Mean Ever BPCI	–	0.172
Patient Controls	Y	Y
Hospital Controls	N	N

Note: This table presents the results from a robustness check including an interaction term in the baseline specification. The term is the interaction of a flag for having positive expected gain based on the spending trend during 2010–13 and an indicator for post-2013. Section A.2 describes in detail how we calculate the expected gain for each hospital assuming that it continues to decrease mean episode spending on its pre-BPCI trajectory. Column 1 here presents results corresponding to the model in Table 3 Column 2. Column 2 here presents results corresponding to the model in Table 3 Column 6. Panel A presents the DD results and Panel B presents the 2SLS results. The top row in each panel presents the baseline coefficients from Table 3 to ease comparisons. Standard errors are clustered by hospital and presented in parentheses.

Table A.6: Patient selection

Panel A	(1) Pred. risk readmission	(2) Pred. use IRF	(3) Pred. use SNF	(4) Pred. use HH
A1. DD	-0.0001	-0.0004	-0.002^{***}	0.002^{***}
	(0.0002)	(0.0003)	(0.001)	(0.001)
A2. LATE	0.001	-0.001	-0.003	-0.005
	(0.001)	(0.002)	(0.005)	(0.006)
Observations	2,574,735	2,574,735	2,574,735	2,574,735
$\bar{y}$	0.117	0.145	0.398	0.381

Panel B	Dist. to hospital	log(1+dist.) to hospital	Pr(BPCI) geographic	Pr(BPCI) geo. & patient
B1. DD	0.145	0.007	-0.004^{***}	-0.003^{***}
	(0.150)	(0.008)	(0.001)	(0.001)
B2. LATE	-0.067	-0.054	-0.001	-0.0003
	(1.082)	(0.054)	(0.006)	(0.006)
Observations	2,574,542	2,574,542	2,574,735	2,574,735
$\bar{y}$	19.955	2.387	0.169	0.169

Note: This table presents evidence on patient selection by the average participant (DD estimates) and average financial complier (LATE estimates). Panels A and B present results related to selection on observed and unobserved attributes, respectively. Panel A presents estimated effects using the specification from Table 3 Column 1 without patient controls. Dependent variables pertain to the 90 days following discharge from the LEJR surgery and are first predicted using demographics and historical utilization and spending (fit on 2008 data). The model coefficients are then applied to patient risk vectors to predict the risk of PAC use of various types. Panel B presents estimated effects on patient attributes using the primary specification, which corresponds to Table 3 Column 2. In Columns 1 and 2, we lose about 200 observations corresponding to beneficiaries in zip codes without centroid coordinates. Distance is winsorized at 100 miles. In Columns 3 and 4, the dependent variable, propensity for BPCI, is predicted from a probit model of 1(ever BPCI) on 3-digit zip code with and without patient characteristics fit on historic (2008) data. The propensity measures are described in more detail in Section 5.6.4. Standard errors are clustered by hospital and presented in parentheses.