NBER WORKING PAPER SERIES

SELECTION AND CAUSAL EFFECTS IN VOLUNTARY PROGRAMS: BUNDLED PAYMENTS IN 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

We are grateful to Rachel Werner and the LDI analytics center for help with access to the Medicare claims data. We thank Diane Alexander, Meredith Rosenthal, Seth Richards-Shubik, Zarek Brot-Goldberg and seminar participants at the University of Pennsylvania, Harvard School of Public Health, ASHEcon, Notre Dame, and the University of Chicago. All remaining errors are our own. Dr Navathe reports grants from Hawaii Medical Service Association, grants from Commonwealth Fund, grants from Robert Wood Johnson Foundation, grants from Donaghue Foundation, grants from the Veterans Affairs Administration, grants from Arnold Ventures, grants from United Healthcare, grants from Blue Cross Blue Shield of NC, grants from Humana, personal fees from Navvis Healthcare, personal fees from Singapore Ministry of Health, personal fees from Elsevier Press, personal fees from Medicare Payment Advisory Commission, personal fees from Analysis Group, personal fees from VBID Health, personal fees from Advocate Physician Partners, personal fees from the Federal Trade Commission, personal fees from Catholic Health Services Long Island, and equity from Clarify Health, personal fees and board membership for The Scan Group, and non-compensated board membership for Integrated Services, Inc. outside the submitted work in the past 3 years. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Selection and Causal Effects in Voluntary Programs: Bundled Payments in Medicare
Atul Gupta, Joseph R. Martinez, and Amol S. Navathe
NBER Working Paper No. 31256
May 2023
JEL No. H00,I10

ABSTRACT

Voluntary participation is a central feature of reforms being tested across US healthcare. Allowing choice can enhance effects if participants sort on unobserved treatment gains. However, selection may also bias program evaluation, misleading policymakers. We study this trade-off in the case of a national reform to reduce spending on surgeries. We exploit variation due to idiosyncratic program rules to instrument for participation. We detect considerable treatment effect heterogeneity but no evidence for sorting on treatment gains. In contrast, there is substantial selection on untreated outcomes in an unexpected direction, leading the difference in differences estimator to understate the causal effect.

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

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

Joseph R. Martinez
University of Pennsylvania
josma@wharton.upenn.edu

1 Introduction

Linking payments for providers to quality and efficiency has become the cornerstone of federal healthcare policy. The Affordable Care Act (ACA) authorized the launch of an innovation center within the Centers for Medicare and Medicaid Services (CMS) with the explicit goal of testing new payment models to improve productivity in healthcare. This center launched 54 different programs over the last decade involving more than a million providers, affecting care for 26 million patients, and representing a plethora of design choices (Smith, 2021). Fifty out of the 54 programs allowed providers to choose their preferred contract. Voluntary participation is therefore a central feature of regulatory reform experimentation in healthcare and is also common in commercial insurer initiatives (Milad et al., 2022). Voluntary programs have the putative advantage that they allow participants expecting the greatest treatment gains to sort in, thus delivering greater gains than mandating participation. However, unobserved selection may also bias standard program evaluation approaches in unpredictable ways, limiting the utility of such programs to guide future policy. Unfortunately, empirical evidence on this trade-off is limited. This paper attempts to fill this gap and provides novel evidence from one of the largest voluntary reforms introduced by Medicare, the US public insurance program for the elderly.

This paper studies the Bundled Payments for Care Improvement program (BPCI), an excellent setting to study voluntary programs. BPCI was prominent and policy-relevant – it was the largest voluntary payment program in US healthcare when it was introduced and the basis for the design of subsequent Medicare bundled payment contracts. Beginning in the fourth quarter of 2013, hospitals could enroll in BPCI for episodes of care initiated by inpatient stays. We study lower extremity joint replacement (LEJR) surgery since this was by far the most commonly selected clinical episode.[1] Our sample ends in December 2016, and by this time about 300 hospitals had enrolled in BPCI out of about 2,530 eligible hospitals nationwide.

The contract design and incentives for hospitals were transparent. If they kept episode spending – defined as the spending that patients incurred across all forms of care including during the surgery, associated hospitalization, and subsequent 90 days – below a target level determined by their historic mean episode spend, they received a payoff. Note that this required hospitals to optimize care costs over a much longer duration and coordinate with physicians in other firms. Hospitals did not necessarily have experience with these new tasks since under the default fee-for-service contracts they received fixed reimbursements for performing the surgery, without any concern for post-discharge utilization. Hence, it

[1] Hospitals could choose from among 48 surgical and medical episodes. Over 70 percent of hospitals participated in LEJR. The next largest clinical bundle, congestive heart failure, was selected by about a third of participating hospitals, with fewer than a third of the patients.

is unclear if they could accurately predict their ability to succeed under the new contract and participate accordingly. Prior studies of this program have noted that participating hospitals were noticeably different than non-participants (e. g., they were more likely to be large, academic medical centers). Selection on prior trends and sorting on treatment gains may therefore both be quantitatively important in this setting. However, previous studies of this program (and others) have typically ignored the endogenous participation decision and applied difference-in-differences (DD) as the baseline program evaluation approach.

We begin with a general conceptual framework that clarifies the role of unobserved selection on counterfactual outcomes and on treatment gains in biasing the DD estimate for the population average treatment effect, or ATE (following Heckman, Urzua and Vytlacil 2006). This organizes our empirical analysis, which seeks to quantify both types of biases and recover the ATE. All three objects (selection bias, selection on treatment gains, and ATE) are important inputs for future contract design. Finding limited sorting on treatment gains would suggest that hospitals did not advantageously sort into BPCI even though the incentives were relatively transparent and predictable, which may undercut one of the benefits of allowing choice. Finding meaningful selection bias would highlight the threat to internal validity and caution against the use of voluntary contracts in pilots or reforms more generally. Finally, the ATE is the relevant input for policymakers if they wish to predict the effect of mandating the new contract across all hospitals.

We propose an instrumental variable to circumvent endogenous participation by hospitals. The approach exploits variation in the incentive to participate due to the interaction of the program's contract design – exogenously set by the CMS innovation center – and pre-determined hospital attributes. Due to the new contract's design, hospitals that historically had higher episode spend levels expected greater per-episode payoffs. Since program implementation costs were largely fixed in nature, hospitals with greater LEJR volume enjoyed lower per-episode participation costs in addition to greater aggregate payoffs. In our preferred specifications, the instrument is an indicator for hospitals with above-median baseline values on both these dimensions, which we use to define hospitals with high potential financial gain. This approach is non-parametric and does not require us to make additional assumptions about hospital expectations over financial gains, though we also test robustness to using such a continuously varying measure as the instrument. We mitigate spurious factors in several ways when constructing the instrument, including for example, using data from well before the passage of the ACA to define the baseline values, and using predicted rather than observed values.

Identification relies on the exclusion restriction that hospitals with high and low values of potential financial gain would have progressed on parallel trends in the absence of BPCI. We present extensive evidence supporting the validity of this restriction. In addition to dynamic reduced form effects on all outcomes and relevant correlates to assess

pre-trends (Goldsmith-Pinkham, Sorkin and Swift, 2020), we also perform a falsification test and confirm that our orthopedic-specific instrument does not predict future changes in *participation* or *spending* in BPCI's other large bundles.[2] This vitiates the argument that unobserved hospital-level factors affecting both participation and performance may be spuriously correlated with the instrument (e.g., larger hospitals have more competent administrators that prefer to participate in BPCI and are better at reducing spend).

We interpret the instrument as quasi-randomly assigning hospitals to low or high potential financial gains. Compliers are hospitals that participate if the idiosyncratically chosen payment formula assigns them greater potential financial gains, otherwise they remain in the default fee-for-service contract. The local average treatment effects (LATE) are policy relevant since they capture the response of participants motivated by financial gains, the key policy design dimension employed by policymakers (and private insurers) to stimulate participation.

The instrument strongly predicts program participation, but also implies that less than 20% of hospitals are compliers, i.e., are persuaded to participate because of greater potential financial gain. The estimated size of the complier group remains small across a range of specifications and alternate instrument versions. This suggests that hospitals consider many factors in participation decisions and that designs relying on financial generosity alone likely have a low ceiling on participation. Other factors influencing participation may relate to mission, expectations about payment reform and the value of learning, or administrative costs (Dunn et al., 2021).[3]

The LATE estimate implies an 11% reduction in episode spend due to participation in BPCI, nearly four times the DD estimate of 3%. The main mechanism is a reduction in the use of post-acute care on the extensive margin, which accounts for two-thirds of the total spending decline. We find a large decline in the use of both inpatient rehabilitation and home health care. The remainder is largely explained by a reduction in the duration of the index LEJR inpatient stay, achieved by increasing the proportion of short stays, which are reimbursed at lower rates.[4] We find null effects on the quality metrics monitored by CMS (e.g., short-term mortality) but detect a large increase in the probability of a revision surgery, an unambiguously bad, but "low-stakes" outcome not monitored by CMS (Jacob, 2005). This result suggests that transmitting financial risk to providers may come at a cost

[2]We examine effects for the five largest bundles after LEJR based on number of hospital participants as well as on upper extremity joint replacement, another orthopedic surgery.

[3]While we cannot individually identify complier hospitals, we can bound the number of participating compliers at less than 90. Since about 300 hospitals participated in our sample, no more than 30% of participants are compliers. Hence, they likely constitute a small fraction of participants.

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

to some patients.

The LATE suggests that the causal effect on spending reduction is much larger than that implied by the DD estimate. However, since compliers form a small fraction of hospitals, the LATE estimate may not reflect the ATE, which is ultimately our objective. To recover population treatment effect estimates, we estimate marginal treatment effects (MTE) with our binary instrument using the separate estimation approach proposed by Brinch, Mogstad and Wiswall (2017). This allows us to predict treatment effects for all hospitals in the sample with minimal additional parametric restrictions. We are unable to reject the null hypothesis of a flat MTE curve, suggesting no detectable sorting on treatment gains. We estimate an ATE of 13% and an average treatment effect on the treated (ATT) of 10%, both quite similar to the LATE. This further implies that the DD estimate does not recover the ATT in this setting and substantially understates it. These patterns are robust to using alternate versions of the instrument.

The MTE analysis yields two additional insights. First, we quantify significant unobserved selection into the program. We use the model coefficients to predict the counterfactual changes in episode spend for each hospital in the absence of BPCI (Andresen, 2018). Our results indicate that participating hospitals would have experienced a smaller decline in episode spend relative to non-participants in the absence of BPCI. Hence, participants were adversely selected, biasing the naive DD estimate toward zero relative to the ATT.[5] Patterns in the raw data support this conclusion. Non-participants experienced a greater decline in the use of post-acute care in the years prior to BPCI (the key mechanism to reduce spending), despite starting off a lower base. Second, including additional covariates and fixed effects in the model prove largely ineffective in mitigating the unobserved selection. Across different permutations, we cannot reduce the unconditional variance in the counterfactual change in spend across hospitals by more than about 30%. This result casts doubt on the ability of conventional evaluation designs such as matching to overcome selection bias in voluntary programs.

This paper makes three contributions. Prior studies of voluntary reforms in the healthcare sector, including on accountable care organizations, have acknowledged concerns due to endogenous provider participation, but have largely relied on a DD approach, with or without the use of matching and provider fixed effects to mitigate provider selection (Werner et al., 2011; Dummit et al., 2016; Navathe et al., 2017; McWilliams et al., 2020; Alexander, 2020; Navathe et al., 2020). We exploit the idiosyncrasies of the contract design for identification and obtain policy-relevant, causal LATE estimates. In principle, our approach can be adapted for use in other settings to circumvent concerns about endogenous

[5]The difference between the mean counterfactual change in episode spend for participants and non-participants is 6.8% (2.4%), which explains nearly the difference between the DD and ATT estimates and is highly statistically significant.

participation. Using this approach, our study uncovers new mechanisms of savings (reducing index hospitalization spending via shorter stays) and patient harms (increased need for revision surgery) for Medicare's bundled payments.

Second, we shed light on the effectiveness of financial rewards in driving participation into new contracts. Our findings suggest that less than 20% of hospitals are motivated to participate by financial gains, implying a low ceiling on participation in programs that rely on the lure of financial rewards alone. This corroborates recent criticisms by experts that financial incentives are ineffective relative to intrinsic motivation in driving participation (Rathi and McWilliams, 2019). This result should give pause to regulators and insurers.

Third, we quantify the importance of unobserved selection versus sorting on treatment gains when evaluating voluntary pilot programs. We study hospital opt-in decisions made under uncertainty about potential ability on a new set of tasks – a common situation in payment reforms since these are often designed to change behavior or shift focus to a new outcome. We find that selection is much more important than sorting on gains and leads the DD to underestimate, rather than overestimate (as typically assumed by policymakers), the ATT. Our study complements Einav et al. (2022), who examine selection and sorting on gains on hospital decisions *after* 2 years of alternate contract participation experience for a similar clinical context, but substantively different Medicare program (CJR, Comprehensive Care for Joint Replacement).[6] Specifically, they study hospital decisions to *opt out* of the alternate contract mid-stream (due to an abrupt policy reversal), with the benefit of knowing their own and other hospitals' performance, mitigating uncertainty and therefore potentially leading to more favorable selection and sorting than under opt-in. They find that the stayers were advantageously selected on spending levels and to a lesser extent, on treatment gains. However, the nature of the selection is different than in our setting; stayers would have performed better in the counterfactual than hospitals that opt out. In our setting, participating hospitals would have actually performed worse under the counterfactual than non-participants. These differences contrast between opt-out and opt-in decisions, with opt-in decisions more reflective of voluntary programs in healthcare at large.

Our results also contribute to the broader evidence on voluntary regulation beyond healthcare. For example, our findings on the lack of advantageous selection and sorting reaffirm recent evidence on school choice. Abdulkadiroğlu, Pathak and Walters (2018) find lower quality schools were more likely to participate in a voucher program, hurting student outcomes. Abdulkadiroğlu et al. (2020) show that parents in New York did not choose schools based on treatment gains. These themes also arise in other sectors, for example, in

[6]CJR was a randomized mandatory market-level experiment implemented in 67 treatment and 104 control MSAs. Although only implemented in select geographies, it offers a potential benchmark for our ATE estimate. However, it set a substantively different financial incentive for treated hospitals. Specifically, hospitals mandated to participate with below-market average episode spend prior to the program could receive payoffs without improving further, dulling the incentive to reduce spend for half the treated hospitals.

designing conservation programs and pollution control incentives to avoid adverse selection (Jack and Jayachandran, 2019; Cicala et al., 2020).

The paper is organized as follows. Section 2 provides the necessary background on the new contract and incentives for hospitals. Section 3 presents a conceptual framework to organize the empirical analysis. Section 4 describes the data sources and presents descriptive evidence. Section 5 describes the instrumental variable and Section 6 presents the LATE estimates. Section 7 describes the marginal treatment effects approach and corresponding results. Section 8 concludes.

2 Background

2.1 BPCI

The Patient Protection and Affordable Care Act (ACA) was enacted in 2010 and introduced payment reforms to better align provider incentives with those of Medicare than the prevailing fee-for-service system. Recognizing the need for testing the effects of different policy designs before broad implementation, the ACA tasked CMS with launching an innovation center to test and evaluate various 'alternative' payment models and accordingly develop policy recommendations . The resulting Centers for Medicare and Medicaid Innovation (CMMI) has now tested more than 50 different initiatives offering alternate payment contracts to providers for services provided to Medicare patients (MedPAC, 2021). This paper focuses on the Bundled Payments for Care Improvement (BPCI), one of the largest alternate contract initiatives.

In 2013, CMMI launched the voluntary BPCI program, its first national bundled payment contract. Between October of 2013 and September 2018, Medicare tested bundled payments for medical and surgical episodes through BPCI across several different episode types. Lower extremity joint replacement surgery (LEJR) was by far the most frequently selected clinical episode, both in terms of number of participating hospitals and patients.[7] Hence, we focus on LEJR surgeries in this paper. Hospitals were accountable for episodes spanning the index hospitalization and 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 were performed at participating hospitals in 2016.

[7]LEJR includes major hip and knee replacement surgeries. It is billed under two MS-DRGs: 469 (complex) and 470 (base). About 300 hospitals participated in the LEJR bundle. The next largest was congestive heart failure, with about 170 hospitals. The difference was even larger in terms of patients – there were more than three times the number of LEJR patient episodes than 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 accounted for over 85% of all patient episodes in BPCI (Lewin Group, 2017).

2.2 Hospital incentives

From the perspective of the payer (Medicare), the goal of a bundled payment is to incentivize 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 for hospitals in 1983 which bundled payment for all inpatient services in a single hospital stay (Acemoglu and Finkelstein, 2008). However, BPCI is more ambitious since it assumes hospitals will coordinate with physicians across all the firms involved in the patient's care and optimize care costs over a *much* longer duration. The average LEJR surgery involves a stay of 3.8 days, while BPCI also considers spending over the next 90 days.

The target episode spend or price, $\hat{p}_{ht}$ for a hospital h in any year t is largely determined by the mean episode spend over 2009–12 for its patients, adjusted by the national trend in mean LEJR spend for Medicare patients between 2012 and t. Note this includes all spend on care during the index surgery and the 90 days following discharge. Intuitively, CMS predicted a hospital would track this price if it maintained its baseline performance and matched national spending trends.[8] The payoff per episode i at a participating hospital h, $gain_{hi}$, is simply:

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

where b_{iht} is the amount billed for patient i over the 90-day episode. The formula indicates that the payoff is positive if the hospital is able to limit spending below the target price. Importantly, the hospital can now also retain savings on the amount that would otherwise be billed by other providers, such as skilled nursing or home health care agencies. In the pre-BPCI period, inpatient care accounted only for about half the total episode spend (Dummit et al., 2016). Hence, hospitals could reap substantial rewards under BPCI if they reduced post-discharge care.

Unlike many other CMS programs, the hospital's target price is *not* adjusted for geographic differences in costs or risk-adjusted for differences in patient complexity, greatly simplifying the computation of the target price. The key source of uncertainty at the time of enrolling in the program is whether the hospital will succeed in changing the behavior of its physicians, nurses, and its clinical protocols to reduce average episode spend below the target price, chiefly by reducing payments to providers of post-discharge care but also by reducing inpatient costs if possible. Since hospitals did not have an incentive to meaningfully coordinate with other firms under fee-for-service contracts, it is unclear whether they had experience with these tasks and could accurately predict their ability to change and participate accordingly.

[8] To ensure CMS captures a fixed saving off the predicted spend, it also deducts a flat 2% off the trend-adjusted amount to arrive at the final target price.

CMS introduced other contract changes to facilitate changes in clinical protocols that could help reduce spend. We discuss only two here in the interest of brevity. First, hospitals were allowed to transmit incentives to physicians and post-acute care (PAC) providers by sharing the savings with them. 66% of participants submitted plans to exercise this 'gain-sharing' option and about $144 million was distributed through 2016 across all bundles. Second, Medicare waived the three-day minimum length of stay requirement to discharge patients to a skilled nursing facility (SNF). The goal was to allow hospitals to discharge patients sooner (and thus bill a lower amount) if they determined it was medically appropriate to do so (Lewin Group, 2018b).[9]

2.3 Participation and aggregate effects

CMS contractors conducted detailed interviews with hospital administrators, providing some insights 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. Relatedly, there was also some intrinsic motivation to improve. Some participants mentioned this was in keeping with their status as "leaders of innovation." In their official 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. Participants were confident in their ability to engage hospital staff and physicians to change practice patterns and therefore succeed. This was particularly true of LEJR surgeries which they viewed as a relatively predictable procedure with lower financial risk. Hospitals frequently 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 payoffs were more likely to participate.[10]

Using simulation analyses, we confirm that hospitals with greater financial gains predicted based on LEJR volume in 2013 and different assumptions on savings were modestly more likely to participate in BPCI. Hospitals at the 75^{th} percentile of expected gain were 1–5 percentage points more likely to participate than those at the 25^{th} percentile across dif-

[9]Hospitals had flexibility in structuring the gainsharing payments. In interviews with the Lewin group, hospitals indicated they used gainsharing payments to focus physician effort on reducing readmissions or changing the discharge destination. They also offered greater referral volumes 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.

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

ferent scenarios, against a mean participation rate of about 12 percentage points.[11] These results support the claim that while financial incentive was one of the key motivations, there likely were other important factors too. This raises concerns about key unobserved factors (such as intrinsic motivation and expertise) being correlated with both participation and changes in episode spend.

Hospitals indicated three principal cost-saving strategies in interviews. First, they would reduce the use of institutional post-acute care (PAC) by changing their discharge guidelines, engaging physicians, and collaborating with their PAC partners. Second, they would reduce the need for 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 spending for the Medicare program. They claimed these strategies entailed substantial new fixed costs and administrative overhead (Lewin Group, 2016).[12] Larger hospitals could bear these additional costs more easily.

Aggregate trends in Medicare utilization suggest that BPCI may have affected spending on LEJR episodes. Figure 1 Panel (b) presents the trend in mean LEJR episode spend for Medicare patients in the US over 2009–16, expressed in 2016 dollars. In 2013, mean episode spend was around $23,000 and it declined to about $21,000 by 2016. Similar observations can be made about 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, the aggregate trends are inconclusive.

[11] We assign each hospital a percent spending reduction value, γ_h, drawn from a normal distribution with a mean of 5%. We examine four scenarios by perturbing the distribution on two dimensions. First, we vary the standard deviation of γ_h (σ,low=1%, high=4%). This varies uncertainty in hospital ability to change physician behavior and clinical protocols, needed to achieve savings. Second, to allow some hospitals greater percent reductions than others based on observables, we vary the correlation between γ_h and baseline episode spend (ρ,low=0, high=0.9). When the correlation is set high, hospitals with higher baseline episode spend enjoy greater percent reductions. We obtain a percent spend reduction value for each hospital under each of the four scenarios. We then scale this value by hospital volume to obtain the aggregate saving in dollars. We repeat this exercise 500 times and generate the mean predicted savings for each hospital within each scenario. For brevity of presentation, we collapse the data to 15 equal sized hospital bins. Figure A.1 plots the mean probability of participation within each bin on the Y-axis against the corresponding mean predicted financial payoff (in thousands of dollars) on the X-axis. Table A.1 summarizes these patterns by presenting coefficients for each scenario from OLS models of participation regressed on the predicted gain and its square.

[12] For example they hired new analysts, nurses, and case managers for patients in BPCI episodes. They devoted considerable physician and staff time to training for new protocols, patient education, and discharge planning.

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 BPCI participation on episode spend and patient health. Prior 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 among Medicare fee-for-service beneficiaries. To our knowledge, the endogenous participation decision has not been explicitly addressed by previous studies.

CMS has also experimented with other bundled payment contracts, providing additional estimates of the causal effects of bundles. 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. This field experiment provides a rare opportunity to estimate the average treatment effect of bundling payments on spending. However, the incentives for participation and improvement in CJR were totally distinct from those imposed in BPCI and described in the previous section. In particular, the target price in CJR was largely determined by the market average episode spend.[13] Hence, hospitals with baseline spending below that of their market average would receive payments without any further reductions, an inefficiency pointed out by Einav et al. (2022). Moreover, the payment formula inadvertently dulled the incentive for about half the hospitals in every treated market. In contrast, all participants in BPCI had to reduce spend in order to receive payoffs. The sharper incentive together with the potential for advantageous sorting predict a greater treatment effect in BPCI.

However, early studies of CJR also report a causal spending reduction of 2–3%, comparable to that in BPCI (Finkelstein et al., 2018; Barnett et al., 2019). We find it puzzling that BPCI, with sharper incentives and endogenous participation, purportedly reduced spending to the same extent as CJR. This puzzle suggests there is value in re-examining BPCI using a different approach that accounts for the endogenous participation.

3 Conceptual framework

Difference in differences (DD) is the workhorse research design for program evaluation (Athey and Imbens, 2017). This is also true of prior studies that have examined voluntary

[13] 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 the benchmark.

payment reforms introduced by the federal government, including BPCI. Studies have used DD designs, with or without the additional use of matching to identify an appropriate comparison group, and provider fixed effects to control for unobserved invariant differences (Agarwal et al., 2020; Alexander, 2020; McWilliams et al., 2020).

The DD estimator recovers the average treatment effect on the treated units (ATT) under the assumption that treated and control units would progress on parallel trends absent treatment. In the presence of unobserved (to the analyst) treatment effect heterogeneity, the ATT may deviate from the ATE. Policymakers deciding whether to mandate a program after a pilot run should place more emphasis on the latter.

We illustrate these challenges below using the potential outcomes framework in equation 2 following Heckman, Urzua and Vytlacil (2006), extending their model for the panel setting. Y_T^j denotes the potential outcome under state j at time T. There are two possible states: BPCI participation, i.e., treated ($j = 0$) and non-participation, or untreated ($j = 1$). No hospital is treated in period $T = 0$ prior to BPCI, while in the period $T = 1$, some hospitals participate in BPCI.

$$
\begin{aligned}
Y_T^0 =& \mu^0(X_T) + U_T^0, \\
Y_T^1 =& \mu^1(X_T) + U_T^1 \\
Y_T =& Y_T^0 + D(Y_T^1 - Y_T^0), \\
Y_0 =& Y_0^0 = \mu^0(X_0) + U_0^0, \\
Y_1 =& \mu^0(X_1) + D(\mu^1(X_1) - \mu^0(X_1)) + \{D(U_1^1 - U_1^0) + U_1^0\},
\end{aligned}
\tag{2}
$$

where $\mu^j(X)$ denotes the expectation of the potential outcome conditioning on the exogenous covariate vector, X. Y_T denotes the realized outcome, which is observed. The DD estimator, β_{dd}, is defined below.

$$
\begin{aligned}
\beta_{dd} = &\{E[Y|D=1,T=1,X] - E[Y|D=1,T=0,X]\} - \\
&\{E[Y|D=0,T=1,X] - E[Y|D=0,T=0,X]\}
\end{aligned}
$$

Substituting values in from the right hand side of equation 2 and noting that the intercept term drops out, we get

$$
\begin{aligned}
\beta_{dd} = &(\underbrace{\mu^1(X_1) - \mu^0(X_1)}_{ATE(X)}) + E(\underbrace{U_1^1 - U_1^0}_{\Delta U_1}|X, D=1) + E(\underbrace{U_1^0 - U_0^0}_{\Delta U^0}|X, D=1) - \\
&E(\underbrace{U_1^0 - U_0^0}_{\Delta U^0}|X, D=0) \\
&= \underbrace{ATE(X) + \underbrace{E(\Delta U_1|X, D=1)}_{\text{Unobserved gains for treated}}}_{ATT} + \underbrace{E(\Delta U^0|X, D=1) - E(\Delta U^0|X, D=0)}_{\text{Selection bias}},
\end{aligned} \tag{3}
$$

The right hand side of Equation 3 can be organized into three terms. The first term captures the ATE conditional on attributes, X. The second term captures the mean unobserved treatment gain for the treated group. Together, the first two terms give the ATT. If there is no treatment effect heterogeneity, i.e., all hospitals with the same values of X experience the same treatment effect from participation, then the second term reduces to zero. Alternatively, this term also reduces to zero if hospitals do not sort on unobserved gains from treatment. This could occur if hospitals are unaware or uncertain about their treatment effects.

The last two terms together capture the unobserved differential trend in *untreated* or counterfactual outcomes between the treated and untreated groups. If this term does not reduce to zero, it is a failure of the parallel trends assumption, and the DD estimate is biased for the ATT. This confounder has been called "selection bias" in the treatment effects literature, and we also adopt this nomenclature (Angrist and Pischke, 2008; Kowalski, 2021). Appendix Section A.1 further describes the potential bias in DD due to the presence of unobserved selection using illustrative examples.

The goal of our analysis is to empirically estimate the population treatment effects of BPCI as well as the bias terms in Equation 3. The ATE provides an estimate of the expected reduction in episode spend if the BPCI contract were mandated across all hospitals, which is interesting in and of itself. Confirming the presence of sorting on treatment gains (i.e., a non-zero second term in Equation 3) helps establish whether hospitals can accurately predict their performance under such alternate payment contracts and choose accordingly. If this is indeed the case, then it increases the appeal of permitting providers to choose their contract. Finally, if we detect economically meaningful selection bias, it will serve as a cautionary tale arguing against using such voluntary contracts in pilots, the results of which may be used to decide on future policy making.

4 Data and descriptive evidence

4.1 Data sources and sample construction

Our primary data source is the universe of Medicare facility claims accessed through the Medicare Provider Analysis and Review (MedPAR) files over 2006–16. For all hospital and post-acute care stays (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 care, start and end dates, a vector of diagnosis and procedure codes, and the discharge destination.[14] We link the MedPAR files with the Master Beneficiary Summary File (MBSF) using the beneficiary's unique identifier to observe enrollment information, further 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 a small component of total episode spend. Based on prior studies, these dimensions of care account for about 11% of the 90-day episode spend.[15]

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 index LEJR cases (MS-DRGs 469 and 470) following the same approach used by CMS when it computes episode spend and target prices for hospitals under BPCI.[16] We exclude about 600 hospitals with fewer than 25 index LEJR cases during the baseline period of 2006–07 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.[17] Appendix Section A.2 provides more details on our sample construction approach.

[14] We also link the MedPAR to home health care claims files to access some variables not available in the MedPAR.

[15] 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 spend for these episodes. Hence, these components account for 11% of total spend. Lewin Group (2018b) Exhibit 10 documents a $56 decrease in part B spend for LEJR episodes in the BPCI program (about 2% of the baseline).

[16] Routine cases are billed under MS-DRG 470, while patients with complications or co-morbidities are billed under MS-DRG 469. About 95% of all cases are billed under 470. Our estimates remain undisturbed if we limit the analysis only to 470 episodes. Prior to 2008, CMS used a different version of DRG classification in which all LEJR cases were billed under DRG 544. Index cases are defined such that the patient does not have a 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.

[17] We exclude hospitals located in Maryland because they are not paid based on DRGs and those located in US territories.

4.2 Descriptive statistics

Table 1 presents descriptive statistics from our analysis sample. Panels A and B, respectively, present hospital-level and episode-level statistics over 2009–13, before hospitals could enroll in bundled payments. We present statistics separately for hospitals that did not participate (column 1) and those that participated in BPCI (column 2), respectively. All spending values are deflated and expressed in 2016 dollars. We note two patterns about participants in Panel A. First, participating hospitals are substantially larger (40% more beds) and obtained nearly $1,200 more in Medicare revenue per bed from orthopedic procedures. This may reflect fixed costs of participation which make entry more favorable for larger hospitals or those with a larger orthopedics practice. Second, participants are 40% more likely to be teaching hospitals, suggesting that academic facilities were more enthusiastic about the new contract.

Market factors also affect a hospital's episode spend, for example, through the average risk or complexity of the local patient pool. Medicare beneficiaries have the option to enroll in traditional Medicare (TM) which allows them to receive care at any provider that accepts Medicare payments, or in Medicare Advantage (MA), where they enroll with a private health plan and access care through its contracted network of providers. MA grew by 7 million enrollees during 2009–16, increasing from 23% to 31% of all Medicare beneficiaries (Kaiser Family Foundation). If MA plans advantageously select healthier beneficiaries, then the remaining beneficiary pool in TM in markets with higher MA penetration will become progressively sicker over time (Newhouse et al., 2015). If BPCI hospitals are disproportionately located in such markets, they may be at a disadvantage. On the other hand, studies have also found substantial treatment style spillovers of the utilization controls employed by MA plans, resulting in lower utilization of hospital care among TM patients in markets with greater MA penetration (Baicker et al., 2013; Baicker and Robbins, 2015). Table 1 reports that BPCI participants and non-participants are located in markets with similar baseline MA penetration *levels*.[18] We examine the differences in penetration trends in Section 4.3 below.

Panel B shows that we observe about 2.5 million episodes over the entire period, with about 18% performed at hospitals that ever participated in BPCI. Participating hospitals had higher treatment intensity prior to the program – an episode cost Medicare about $2,000 more at participating hospitals on average ($25,000 vs. 23,000). About 40% of this gap is explained by higher billing for the joint replacement surgery. The remainder is largely explained by higher post-acute care (PAC) spend for patients in the 90 days following

[18] We compute this value for each hospital-year as the weighted average MA share across the hospital service areas (HSA) contributing patients to the hospital. The weights are the HSA's share of the hospital's total LEJR patients in 2007, held constant. The Dartmouth Atlas group defined HSAs as contiguous zip codes whose residents receive most of their inpatient care from the hospitals in the same area. There are approximately 3,400 HSAs in the US, slightly more than the number of counties.

discharge from the surgery. The difference in PAC cost is driven both by greater use on the extensive margin (86% vs. 81%) as well as on the intensive margin. To summarize observed patient risk, we construct an index of propensity for SNF use.[19] Participating and non-participating hospitals appear to have treated patients at similar observed risk of SNF use.

4.3 Trends prior to participation

Since parallel trends is the key identifying assumption in the DD design, this section presents results from formal tests for differential trends in the time-varying correlates between participating and non-participating hospitals in the pre-BPCI period. We estimate the non-parametric event study model presented in equation 4 below.

$$Y_{hmt} = X'_{hmt}\,\delta + \sum_{s \neq 2013} \beta_s \, d_h 1(t = s) + \epsilon_{hmt}, \tag{4}$$

Y is the correlate of interest for hospital h located in market m in year t. X includes hospital and market by year fixed effects, allowing markets to trend differentially. We do not include any other covariates in X for this exercise. d is an indicator for participating hospitals, it is interacted with an indicator for each year, and ϵ denotes unobserved factors. Following prior studies, we use hospital referral regions (HRR) to define hospital markets. 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 β_s, the differential trends in Y at BPCI hospitals relative to non-participants.

Figure 2 plots the corresponding coefficients obtained over 2008–13, along with 95% confidence intervals. The figure suggests that participating hospitals trended differentially in the pre-BPCI period on a number of outcomes related to LEJR episode spending. Panels (a)–(d) examine different measures of patient volume, gradually converging to Medicare fee-for-service LEJR volume, the procedure in question. Panels (a) and (b) suggest that participating hospitals differentially increased total patient volume and Medicare fee-for-service orthopedic admissions, respectively, during this period. However, the coefficients are imprecisely estimated. Panel (c) investigates trends in orthopedic revenue from traditional Medicare patients, again finding an increasing trend. Panel (d) shows an increasing trend in LEJR volume at participating hospitals. This suggests participating hospitals may have had greater managerial and clinical focus on LEJR surgeries leading up to BPCI. Panel

[19] We use 2008 data to estimate a linear model predicting the probability of having a SNF stay in the 90 days following discharge on patient risk controls. We apply the coefficients of this model to the data over 2009–13 to generate predicted SNF risk values for each patient. We also performed the same exercise to predict the risk of inpatient rehab and home health care use to capture changes in risk on other margins. Those statistics are not presented in the interest of brevity but are qualitatively similar.

(e) indicates participating hospitals drew patients from communities that experienced a differential decline in the share of MA beneficiaries. Note that there was a secular trend of increasing MA penetration during this period, hence this result implies participating hospitals drew patients from areas with slower MA growth. Panel (f) implies that LEJR patients at participating hospitals experienced a differential decline in predicted risk of SNF use, based on observed attributes.

The event study plots suggest pre-trends but they are often imprecise. Table 2 summarizes these trends into a single coefficient for each correlate to improve precision. We report coefficients from separate regressions estimating a variant of Equation 4 in which we include a linear trend interacted with the participation indicator to estimate the differential trend for participants (instead of the β_s coefficients). Column 1 reports the coefficient on the overall trend and Column 2 reports the coefficient on the differential trend among participants. The coefficients are consistent with the patterns in Figure 2. Participating hospitals had a statistically significant increasing trend in orthopedic and LEJR Medicare admissions per bed and a declining trend in MA penetration and predicted SNF risk. In all cases, the differential trends for participating hospitals are substantial compared to the trend for non-participants. For example, non-participating hospitals experienced a reduction of 0.05 LEJR cases per bed over 2008–13 (5*0.01). In contrast, participating hospitals experienced an increase in LEJR cases of 0.02 per bed (5*(0.014-0.01)).[20]

5 Instrumental variable

5.1 Potential financial gain

To circumvent concerns related to endogenous participation, which may also relate to the differential pre-trends discussed in the previous section, we develop an instrumental variable approach. The design exploits the idiosyncratic choices made by CMS in setting the target price, which generate plausibly exogenous variation across hospitals in the ex-ante potential for net financial gain from participation. We utilize differences in two lagged attributes across hospitals to harness this policy-driven variation.

As discussed in Section 2, the conventional wisdom in the hospital industry is that reducing episode spend typically involves incurring investments such as hiring additional nurses, case managers, and data analysts, improving the functionality of electronic health records to better coordinate with external providers, redesigning clinical protocols, and

[20]We examined trends in a number of time-varying correlates apart from those presented in Table 2, 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. Among these, we found differential trends at participating hospitals in the case of LEJR readmissions.

training staff on the new protocols (Lewin Group, 2016). Since these actions, which we interpret as the 'technology' of spend reduction, largely involve incurring fixed costs, scale economies come into play, making it easier for larger hospitals or those with greater LEJR volume to justify incurring these. Hospitals with a larger LEJR practice at baseline also naturally expect a greater aggregate payoff from moving to bundled payments. Hence, the first attribute we use is a measure of hospital size at baseline – LEJR revenue from Medicare.

CMS chose a contract design such that each hospital's target price was determined by its baseline episode spending level. Hospitals with greater baseline spend likely had higher baseline rates of PAC use and readmissions. Assuming that the spend reduction technology has diminishing returns, hospitals with higher baseline spend levels had the potential to obtain larger reductions per unit cost. Accordingly, the second attribute we use is a measure of the potential for reduction at baseline – mean post-discharge spend per episode. Note that the aggregate potential for financial gain increases in both dimensions individually, and in their interaction.

We use data from 2007 to compute baseline values of these two measures for each hospital in the sample. This was much before the announcement of the ACA, much less BPCI. We take two more steps to mitigate the potential for spurious correlations between these variables and future outcomes of interest. First, we use data across all orthopedic procedures rather than limiting to LEJR surgeries alone.[21] Second, instead of using the observed values, we project them on hospital attributes such as teaching status, number of beds, etc. and patient risk measures and obtain predicted values. Appendix Section A.3 describes the steps in detail. This approach sacrifices some of the variance in the observed values, but mitigates spurious correlation. Reassuringly, we find similar estimates if we relax these two restrictions.

Figure 3 presents bin-scatter plots showing the mean participation rate on the Y-axis against the mean predicted baseline orthopedic revenue (panel a) and post-discharge spend (panel b), respectively, in each decile bin on the X-axis. The plots show that these baseline attributes predict hospital participation in BPCI 7 years later. Panel (c) shows that the interaction of these two variables is more strongly predictive of participation since hospitals which are above the median on *both* dimensions are more likely to participate. We incorporate this pattern in our empirical approach to avoid weak instrument problems. Our instrument is an indicator, z_h, for hospitals with greater than median baseline orthopedic revenue and episode spend. For ease of exposition, we refer to this as hospitals with greater potential financial gain. A strength of this non-parametric approach is that we circumvent the need for additional assumptions to convert baseline attributes into dollars of predicted

[21] 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 spend over the 90 days following these stays.

gain. In sensitivity tests, we do make such assumptions to compute continuously varying predicted gain dollars. We find qualitatively similar results using this alternate instrument.

The cross-sectional nature of the instrument implies we do not have identifying variation to predict differential time of entry. Hence, we abstract away from staggered entry, assuming that treatment begins in 2014 regardless of when hospitals joined BPCI. Once a hospital enters BPCI, we assume it participates through the end of our sample, ignoring early exit. Hence, the IV analysis recovers the average effect of ever participating in BPCI.[22]

Equation 5 below presents the first stage equation, which closely corresponds to the event study model presented in equation 4. The vector X now includes patient controls in addition to the fixed effects, and in some models we include time-varying hospital controls as well.[23] 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 6 presents the outcome model with β as the coefficient of interest. We estimate both models simultaneously using two stage least squares.

$$d_h \cdot \mathbf{1}(t \geq 2014)_t = X'_{ihmt}\,\delta_1 + \pi\, z_h \cdot \mathbf{1}(t \geq 2014)_t + \eta_{ihmt}, \tag{5}$$

$$Y_{ihmt} = X'_{ihmt}\,\delta_2 + \beta\, d_h \cdot \mathbf{1}(t \geq 2014)_t + \epsilon_{ihmt}. \tag{6}$$

The coefficient β estimates the local average treatment effect (LATE) of BPCI participation on complier hospitals. Most voluntary contracts in healthcare attract participation by offering financial rewards. Since compliers are hospitals that take-up BPCI because of greater potential financial rewards proxied by the indicator z_h, their response – captured by the LATE – is of much policy interest.

5.2 Identification assumptions and instrument validity

A causal interpretation of the IV estimate β relies on two untestable identification assumptions.

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

[23] 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.

ASSUMPTION 1 (Conditional independence): $\mathbb{E}(\epsilon_{ihmt}\, z_h | X_{ihmt}) = 0$.
This assumption is standard in IV designs and implies the instrument is as good as randomly assigned after conditioning on the covariates and fixed effects, i.e., within a market-year cell. 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 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.

Goldsmith-Pinkham, Sorkin and Swift (2020) suggest that the concern of differential trends also extends to correlates that may affect both the instrument and the outcomes of interest. We re-examine pre-BPCI trends for the same factors we assessed in Section 4.3, but assess correlation with the instrument instead of with participation. Figure A.2 presents the corresponding reduced form dynamic coefficients for the same correlates. Reassuringly, we find no evidence of pre-trends on any of these correlates. Table 2 column 3 presents the corresponding differential linear trends associated with hospitals with higher potential financial gain. These coefficients confirm the patterns in the event studies. Not only are the coefficients statistically insignificant, but they are also typically much smaller in magnitude than those in column 2 discussed previously. These patterns are not surprising since the policy and payment formulas were not designed to favor any specific types of hospitals and were exogenously determined from the perspective of hospitals.

A violation of the exclusion restriction is the possibility that the hospital's baseline orthopedic attributes may directly affect future spending changes through hospital-wide production inputs such as managerial quality or a greater focus on value based payments. If so, this instrument should also be correlated with future changes in episode spend for patients in other BPCI clinical bundles, 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 spend for six other bundles offered through BPCI and find no relationship between the IV and participation or spending changes. Five of the six clinical bundles are the largest bundles after LEJR, by the number of participating hospitals. In addition, as a stricter test of the presence of spurious effects, we also test on an orthopedic surgical procedure – upper extremity joint replacement (UEJR). These surgeries are often performed by the same surgeons belonging to the same departments at the hospital. Section 6.4 describes the corresponding results.

ASSUMPTION 2 (Monotonicity): $Pr(d_h = 1|z_h = 1) \geq Pr(d_h = 1|z_h = 0)$ for all h. 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 3 panels (a), (b), and (c) show that this condition holds on average for baseline orthopedic revenue, episode spend, 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.

6 Local average treatment effects

6.1 First stage

Table 3 presents the OLS and 2SLS estimates on the effects of BPCI participation on episode spending. Columns 1–4 present results from the corresponding models estimated on patient-level data. Since the instrument exploits static variation across hospitals at baseline, in principle we should be able to estimate the IV models on cross-sectional hospital-level data. We therefore also present results from models estimated on first-differenced, hospital-level data in columns 5 and 6.[24] Since hospital fixed effects drop out of these models, they help us probe the validity of our quasi-experiment without including any covariate.

We begin by describing the first stage results. Panel B presents the first stage coefficient, π, from Equation 5. Column 1 presents the coefficient from a specification with only fixed effects; columns 2 and 3 from specifications that add patient or hospital controls, respectively. Column 4 presents results from a model including both patient and hospital controls. Columns 5 and 6 are equivalent to columns 1 and 2, respectively, in terms of the controls used, but present results from models estimated on cross-sectional, hospital-level data. Our preferred specification is the one including patient controls, but not hospital controls (cols. 2 and 6), since some of the hospital attributes may be affected by participation.

Within each type of model (patient- or hospital-level), including or excluding the controls does not affect the first-stage coefficient meaningfully, consistent with treatment balance implied by Assumption 1. We obtain f-test statistics in the range of 17-18 and 24-25 in the patient-level 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).

[24] 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 perform 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.

Regardless of the controls or the type of model, the first-stage coefficient is less than 0.2. The first stage coefficient in a model with a binary instrument, particularly without covariates (column 5), is equivalent to the proportion of complier hospitals in the sample (Angrist and Pischke, 2008), i.e., hospitals that participate because of greater potential financial gain, but not otherwise. The proportion of complier hospitals appears low and suggests that voluntary programs relying on financial incentives may have a low participation ceiling. We tested the sensitivity of the coefficient and predictive strength of the first stage to alternate formulations of the instrument, and the results are presented in Appendix Table A.2.[25] We test sensitivity to using the observed rather than predicted baseline orthopedic values (col. 2), directly using LEJR rather than orthopedic baseline values (col. 3), and using expected financial gains, either dichotomized (col. 4) or as a continuously varying value (col. 5). We note that the instrument constructed using LEJR data is marginally weak, and hence we attach less importance to this variant. The other variants do not raise weak IV concerns. All the different permutations imply that less than 20% of the hospitals are compliers.

6.2 Episode spend

Table 3 also presents the DD and LATE estimates of the effects on (log) episode spend in Panels A and B, respectively. In the case of columns 5 and 6, the outcome is the change in mean log episode spend between the post- and pre-BPCI periods. Across the different specifications, the DD estimate implies approximately a 3% reduction in episode spend while the LATE implies a much greater effect of 11-12%.[26] The stability of the estimates across different controls, including without the use of any covariate (col. 5) is reassuring. Appendix Table A.2 also presents the corresponding LATE estimates obtained by using alternate formulations of the instrument. 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 greater 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 OLS point estimate.[27]

Figure 4 Panel (a) presents dynamic reduced form effects on episode spend before and

[25] These models were estimated on hospital-level data and correspond to the specification in Table 3 col. 6.

[26] In addition to the specification permutations presented here, we also estimated models flexibly allowing the coefficient on patient risk indicators to differ in the post-BPCI period, or by year. This addresses recent critiques about the potential for confounding due to time-varying covariates (Zeldow and Hatfield, 2021). The coefficient of interest remains undisturbed in both the OLS and IV models.

[27] 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.

after the introduction of BPCI. These are obtained by estimating a variant of Equation 4 on the full sample period where we use the instrument, z_h, instead of 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. 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. The trajectory is consistent with the gradual increase in participation over time.

Applying our preferred LATE estimate in column 2 (11%) to the mean episode spend of $25,000 at BPCI hospitals prior to the program, we obtain an implied saving of $2,750 per episode. Figure A.4 Panel (a) presents the distribution of episode spending in the pre-BPCI period, providing additional benchmarks (e. g. the median spend is around $20,000). Next, we explore the sources of reduction in spending by examining the effects on the various components of the episode. We split the episode into two buckets: the index LEJR surgery and post-discharge care (readmissions and post-acute care including home care). These account for approximately 60% and 40% of the total episode spend, respectively.

6.2.1 Index stay

Since both the target price and spending for the initial hospital stay are determined by the DRG associated with the initial admission, hospitals have limited avenues to reduce the amount spent on the LEJR surgery. Hospitals could attempt to secure higher benchmarks by increasing use of the more lucrative DRG for complicated cases; however, this 'upcoding' option is likely applicable to a very small set of marginal patients. Nevertheless, in unreported analyses, we confirm that the share of patients billed to the complicated DRG remained stable following BPCI at participating hospitals.

Thus, the principal available avenue to reduce index stay spending is to shorten the stay for patients to below a threshold duration 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 fee-for-service 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 fee-for-service was the 3-day minimum duration requirement to discharge patients to SNF. 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.[28]

[28]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 accounts for 95% of the total volume. In our sample, the mean payment for 470 cases of length $\leq$ 2 days is about $1,200 less than cases longer than 2 days.

Table 4 columns 1 and 2 present the estimated effects on log length of stay for the LEJR surgery and the probability that the length of stay is less than 3 days. To help interpret these coefficients, Figure A.4 panel (b) presents the CDF of the index stay's duration using data prior to BPCI (the median and mode were 3 days). The LATE estimate in column 1 implies a 10% decline in length of stay, but it is only marginally significant. Column 2 shows that there is a 16 percentage point differential increase in the proportion of 1- or 2-day stays at participating hospitals, a nearly 70% increase in the prevalence of such stays relative to the pre-BPCI mean level. This suggests a meaningful change in treatment protocols at the hospitals following BPCI participation. Figure 4 panel (b) presents the dynamic reduced form coefficients on log length of stay and confirms a noticeable decline starting in 2014. Figure A.5 Panel (a) presents the corresponding event study on the probability that the index stay is shorter than 3 days.

6.2.2 Post-discharge care

Table 4 columns 3–6 present the effects on the use of post-acute care. The overarching takeaway from these results is we find a large decline in the use of PAC on the extensive margin. The LATE in column 3 implies a 13 percentage point decrease in the use of any PAC, about 15% of the mean level of 80%. Figure 4 Panel (c) presents the corresponding reduced form dynamic effects on the use of any post-acute care. The figure shows no patterns prior to BPCI and a noticeable change in trajectory after 2014. Table 4 columns 4–6 present effects on the use of specific forms of PAC – home health, inpatient rehabilitation, and skilled nursing, respectively. We find large declines in the use of home health (30% of the mean) and inpatient rehab (90% of the mean), but no effect on the use of skilled nursing. Appendix Figure A.5 panels (b)–(d) present the event studies for effects on the three types of PAC use. The patterns are consistent with the IV estimates, with noticeable changes in trajectory after 2014. The finding of no net effect on skilled nursing may mask two opposing effects that offset each other – an increase in skilled nursing because it substitutes for more intense forms of care such as inpatient rehab, and simultaneously a decrease in skilled nursing because patients are discharged to home health care instead.[29]

The estimated LATE effects on use of inpatient rehab and home health imply a reduction in episode spend of approximately $1,900, assuming there are no changes on the intensive margin. Of this, inpatient rehab 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 spending is due to a reduction in PAC use.[30]

[29]The sum of mean values across the 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.

[30]We compute this by scaling down the mean spend on inpatient rehabilitation at BPCI hospitals

6.3 Patient health

An important concern with the expansion of prospective payment is that it may incentivize providers to skimp on useful care, which may lead to adverse outcomes for patients (Cutler, 1995). CMS anticipated this unintended consequence when it launched this alternate payment contract and included some stipulations around maintaining performance on quality metrics.[31] However, the 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 (HRRP). Previous studies have found no economically or statistically meaningful effects of bundled payments on patient quality. We also examine several quality measures using our IV approach. All our models include extensive patient-level controls for differences in risk. Hence, they account for changes in observed patient risk. We explore concerns about potential changes in patient mix, including unobserved selection, in Section 6.4.

Table A.3 presents the estimated DD (panel A) and IV (panel B) coefficients for a number of quality metrics. 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 hence are more tightly linked to lapses in surgery quality than all-cause readmissions. Further, these are unambiguously bad outcomes for patients (Katz, 2006). Some physicians had expressed concerns that the financial incentive to cut costs may lead to changes in the implants used by hospitals, ultimately requiring more revisions (Koenig et al., 2018). These concerns appear well founded since hospital administrators discussed device standardization and consolidation of vendors as a strategy to reduce implant costs (Lewin Group, 2017). Importantly, revision surgeries was not one of the quality metrics monitored by CMS. Hence, examining effects on this "low-stakes" outcome potentially helps us assess quality more realistically (Jacob, 2005).

Taken together, the LATE estimates imply no meaningful change in adverse outcomes for patients, except an increase in the frequency of revision surgeries. While the coefficient is small in absolute terms, it implies a doubling of the frequency post-BPCI. Figure A.6 presents the corresponding reduced form event studies for each outcome presented in the table. The plots confirm that hospitals with high potential financial gain were not differentially trending prior to 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

by the change on the extensive margin: $2,188$ (mean IRF spend) $*$ $[1 - (0.136$ (pre-bpci mean level) $-$ 0.09 (estimated reduction)$)/0.136] \approx 1,450$. We obtain the estimate for home health care using similar logic.

[31] Hospitals were only able to earn savings bonuses if performance on certain quality metrics (e.g., complications after surgery) did not worsen.

revision surgeries and confirms a gradual increase post-BPCI.

The estimated effect on revision surgeries implies an expected increase in Medicare spending of approximately $380 per episode.[32] Hence, this is a small increase in cost relative to the implied reduction in total episode spend. This value does not incorporate the loss in utility for the affected patients since it does not incorporate potential loss in functional status and decline in care experience that were indicated in patient surveys conducted by CMS. LEJR patients treated at BPCI hospitals reported a statistically significant worsening in mobility, discharge experience at the hospital (including whether they were discharged at the right time), and overall satisfaction with their recovery. The effects were small in magnitude, but paint a picture consistent with our results (Lewin Group, 2018b).

6.4 Alternate explanations and robustness

We perform falsification tests to assess the importance of two potential violations of the exclusion restriction. A key concern is that the hospital's baseline orthopedic attributes may directly affect future spending changes through hospital-wide production inputs such as managerial quality, intrinsic motivation to learn or improve (cited in hospital interviews), or a greater focus on value based payments. As discussed in Section 5.2, we assess the importance of this concern by testing whether the instrument also similarly predicts participation in other bundles and reductions in spending.

We repeat our IV analysis on hospital participation and episode spend for six other bundles offered through BPCI. We select bundles which enjoy high participation and where episode spending was highly correlated with LEJR prior to BPCI. The first five – heart failure, pneumonia, chronic obstructive pulmonary disease (COPD), sepsis, and urinary tract infection (UTI) – are the largest bundles after LEJR, ranked by the number of participating hospitals. The sixth is an orthopedic surgical procedure – upper extremity joint replacement (UEJR) – often performed by the same surgeons belonging to the same department as those performing LEJR.[33]

Table 5 presents the corresponding point estimates. Panel A presents the DD estimate on episode spend, mostly for completeness and to be consistent with the format used in previous tables. We repeat the estimates for LEJR in column 1 for ease of comparison. The remaining columns present estimates for the different clinical episodes, ordered in decreasing extent of participation in the bundle. Panel B presents the reduced form coefficients and f-statistic from the corresponding first stage model. First, we find that, with the exception

[32] 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 the surgery and follow-up care only in the 30 days following discharge.

[33] Appendix Table A.4 Panels A and B show that the correlations between LEJR and these bundles on total and post-discharge spend were typically around 0.7 and 0.6, respectively. Hence, if hospital-wide factors drive the post-BPCI changes in LEJR spend, they may plausibly do so for the other bundles as well.

of heart failure, the instrument has negligible predictive power for participation in the non-LEJR bundles. We therefore focus on the reduced form coefficients instead of the LATE estimates. The instrument also does not predict changes in episode spend for any of the comparison conditions, except in the case of COPD, where it predicts an *increase* in spend. When we consider a pooled sample of the six conditions, the reduced form coefficient is nearly zero and precise enough to rule out the effect we obtain for LEJR.

Figure 5 presents the corresponding dynamic reduced form effects on episode spend for LEJR (thick red line) and each of the comparison conditions (grey lines). The figure clearly illustrates that the reduction in episode spend for LEJR is uniquely negative and large. The figure also shows that the increase in spend for COPD occurred in 2013, prior to the introduction of BPCI, and is therefore spurious. Taken together, this evidence reassures us that the instrument does not predict a reduction in episode spend where there is no basis for a relationship, supporting the exclusion restriction.

Another threat to exclusion is whether our results partially capture coincident hospital-wide changes in treatment unrelated to BPCI, but perhaps a response to changes in the broader regulatory environment or changes in the expense associated with inpatient surgical care more generally. If so, we would expect to find similar effects in procedures not included under BPCI at the participating hospitals. To assess this concern, we examine the effects on episode spend and PAC use for cholecystectomy (gallbladder removal) surgeries at the hospitals participating in the LEJR bundle. This is a high-volume surgery performed by a different type of surgeon (general or gastrointestinal specialty), mitigating the potential for within-hospital spillovers in clinical protocols or practice styles. Furthermore, CMS has not yet offered a bundled payment contract for this procedure, nor is it subject to other performance-based penalties. Appendix table A.5 presents the corresponding OLS and LATE point estimates. Unlike the IV estimates for LEJR, these are statistically insignificant and much smaller in magnitude. These estimates are also precise enough to rule out our main effects. Figure A.7 confirms the absence of any changes in trends after 2013.

We then present evidence from multiple robustness checks. Recall that the point estimates on all outcomes remain nearly unchanged to the inclusion of a rich set of patient- and hospital-level controls. This mitigates concerns of mis-specification bias. We assess whether the LATE estimate is partly driven by mean reversion in episode spend. Our research design of using 2007 data to develop instruments mitigates this possibility. Nevertheless we assess this concern and study the unadjusted time series trends in mean episode spend in Figure A.8.[34] The figure presents the trends separately for hospitals with high and low values of potential financial gain, respectively. The two groups evolve on parallel trends over six years (from 2008 through 2013) and diverge only following the introduction

[34]To focus on trends and abstract away from levels, we normalize the values so that each curve is set to 100 in 2013.

of BPCI. Indeed, between 2010 and 2013, the two curves virtually overlap.

We assess in multiple ways the possibility that complier hospitals responded by advantageously selecting patients based on risk and find no evidence to support this channel, similar to previous studies (Finkelstein et al., 2018; Barnett et al., 2019; Einav et al., 2022). Appendix Table A.6 and Figure A.9 present the corresponding point estimates and dynamic effects, respectively. We directly assess whether there was a differential change in observed patient risk at participating hospitals post-BPCI. We summarize information on patient risk across a rich set of disease co-morbidities and utilization history by generating a predicted risk of adverse outcomes – readmissions and use of different types of PAC.[35] We also 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 and in their predicted probability of choosing a BPCI hospital based on their location and attributes. We find no meaningful change on any of these dimensions. Appendix Section A.4 describes these tests in more detail.

6.5 Discussion

The LATE estimates imply that treated complier hospitals differentially reduced episode spend by 11-12% post-BPCI, relative to untreated complier hospitals. We find reductions in the index surgery length of stay and increases in revision surgeries not previously described in literature. We do not find any changes in patient mix that could explain these patterns, so we interpret the results as real reductions in care utilization. The relatively large reduction in care utilization suggests inefficiencies under fee for service, that are only partially offset by the small increase in revision surgeries. The compliers are hospitals that participated because of greater potential for financial gains. Since most new voluntary contracts in healthcare attract participation through generous financial rewards, the LATE estimates are highly policy relevant.

That being said, the LATE estimates are not externally valid for population treatment effects such as the ATE without strong additional assumptions (Imbens and Angrist, 1994; Heckman et al., 2006). For example, if the compliers represent a distinct sub-group of hospitals that is well-suited to succeed under BPCI, the LATE may overstate the ATE. The complier group does have some distinctive features - they are more likely to be standalone, for-profit owned, and below-median in size than the average participating hospital.[36] These

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

[36] Table A.7 compares the full sample, complier, and treated hospital groups, respectively, on a number of relevant attributes using data from the pre-BPCI period. 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 corresponding proportions among complier hospitals and participating hospitals, respectively. We followed the approach proposed in Abadie (2003) to compute the values for the complier group.

attributes may make it easier for the compliers to adapt to new clinical protocols and generate savings. Next, we build on this IV approach to estimate the selection bias and sorting-on-gains terms discussed in Section 3 and obtain population average treatment effects.

7 Population Treatment Effects

This section presents empirical estimation of the three objects discussed in Equation 3: the ATE, sorting on treatment gains, and selection bias. In particular, policymakers considering the merits of switching entirely from fee for service to bundled payment contracts need an estimate of the ATE. To do so, we estimate marginal treatment effects (MTE) using the separate estimation approach proposed by Heckman and Vytlacil (2007) and adapted for binary instruments by Brinch, Mogstad and Wiswall (2017). We then integrate the MTE estimates across the appropriate hospital sub-groups to obtain the ATE, ATT, and quantify selection. For brevity, we focus on episode spend, the key outcome of interest.

7.1 Estimation

To simplify estimation and focus on selection and treatment effect heterogeneity at the hospital level, we perform the MTE analysis on hospital-level data. We previously discussed LATE estimates of BPCI participation on episode spending using this data (Table 3 cols. 5 and 6). The outcome is therefore the change in log episode spend, denoted as ΔY_h.

Since the separate estimation approach has been described in detail previously in Brinch, Mogstad and Wiswall (2017) and elsewhere, we discuss the necessary details in Appendix Section A.5. Here, we briefly note the two additional assumptions that are required to extrapolate causal treatment effects beyond the compliers and obtain population-level treatment effects. First, we assume additive separability between the observed and unobserved components of the marginal treatment effect. This restricts the slope of the MTE curve against the unobserved resistance, U_D, to be independent of X (patient risk factors), considerably easing estimation of the MTE. Although restrictive, this is weaker than the separability assumption implicit in Equation 6 and routinely used in IV analysis.[37] Second, we assume a linear functional form when estimating the MTE. Both assumptions are standard in the MTE literature, the latter specially when using binary instruments, as is the case here (Kowalski, 2021). The assumptions are formally defined in Section A.5.

[37] The assumption in the MTE analysis is weaker since we allow treatment effects to differ by X and by U_D, but not by the interaction of the two. In contrast, treatment effects are constant across X values in the 2SLS model.

7.2 Results

Figure 6 panel (a) presents histograms of the estimated propensity scores for hospitals in the ever-BPCI and never-BPCI groups. The distributions overlap approximately over the range 0.03–0.66, and reassuringly, most hospitals fall within this range. Following convention in the MTE literature, we limit the analysis sample to the 2,312 hospitals (out of about 2,440) that overlap in treatment propensity values.[38] We interpret the propensity score as capturing the variation in participation probability due to differences in financial gain predicted by observed factors.

Figure 6 panel (b) presents the first set of results from the MTE analysis. We plot the MTE estimates of the change in log episode spend on the Y-axis against values of the unobserved resistance, U_D, on the X-axis. Hospitals with higher values of U_D need greater financial inducement to participate (implied by a higher value of p), or alternatively, have greater unobserved resistance to participating in BPCI. The shaded grey region indicates the corresponding 95% confidence interval at each value of U_D. Importantly, we cannot reject the null hypothesis of a flat MTE curve or constant marginal treatment effects, thus the MTE analysis implies that sorting on treatment gains is less important in this setting. Quantitatively, we estimate a downward sloping MTE curve, implying the hospitals most likely to participate obtained lower spending reductions than the hospitals with greater reluctance. In other words, we find directional evidence of reverse sorting on treatment gains.

Figure 6 Panel (c) presents the treatment effects obtained by appropriately integrating the MTE estimates over different ranges. For ease of comparison, Panel A columns 1 and 2 present the OLS and LATE estimates, respectively, corresponding to the estimates presented in Table 3 column 6. The two sets of estimates differ slightly since the MTE analysis was performed on the truncated set of hospitals that overlap in treatment propensity.

We perform a specification check on the linear functional form assumption by aggregating the MTE values using IV weights and computing the IV estimate (Panel A col. 3). While the 2SLS estimate is robust to non-linearity in the true data generating process, the MTE values are sensitive to the choice of the functional form (Cornelissen et al., 2018). However, the MTE values perform well in approximating the 2SLS estimate (10.9% vs. 12.2%), reassuring us that the linear specification is reasonable in this setting.

We then use the MTE estimates to recover population treatment effect estimates, presented in Panel B. Column 1 presents the estimated ATE of a 13.5% reduction in episode spend, slightly larger in magnitude than the 2SLS estimate (12.2%). These results imply that the LATE approximates the average effect on episode spend we would obtain if bundled payments were mandated across all hospitals. Column 2 shows the ATT (10.7%),

[38] We consider hospitals within the 0.1 ($p = 0.032$) and 99.9 percentiles ($p = 0.657$) of overlapping propensity scores.

estimated to be smaller than the ATE. This is to be expected since the ATT places more weight on hospitals most likely to participate while the ATE places the same weight on all hospitals. The difference in magnitude between the ATT and ATE (-2.8%) quantifies the sorting on treatment gains term in Equation 3, except that it is negative and suggests reverse sorting. Similarly, the average treatment effect for the untreated group (ATUT) is slightly greater in magnitude than the ATE at 14.2%. The difference in magnitude between the ATT and ATUT (-3.5%) provides an alternate measure of the magnitude of treatment effect heterogeneity. We are unable to reject the null hypothesis that this difference is significantly different from zero (s.e. of 2.9%), consistent with the inability to detect sorting on treatment gains.

Figure 6 panel (d) plots the weights associated with the ATT, ATUT, and the LATE and helps explain the relative magnitudes. The ATT puts greater weight on hospitals with lower resistance to treatment, which also have smaller estimated treatment effects. Since the LATE weights more heavily hospitals with moderate resistance to treatment, it lies in between the ATT and ATUT, and so on. The difference between the ATT and OLS estimates, about 8%, estimates the selection bias term in Equation 3. These results clarify that selection bias is quantitatively more important than sorting on gains in this setting.

Table A.8 presents robustness of the implied population treatment effect estimates to using alternate versions of the instrument. We test sensitivity, using the same variants investigated in Table A.2. Reassuringly, the point estimates across the checks are typically less than two standard errors away from the corresponding baseline value. The exception is when we use the instrument in row 3 computed using data on LEJR patients and which is marginally weak. The baseline point estimates tend to be at the lower end in magnitude, raising the possibility that they may be understating the treatment effects on spending. The key qualitative patterns remain consistent across all variants of the instrument – the predicted average treatment effects for non-participants and participants have overlapping confidence intervals and the estimated ATUT is greater than the estimated ATT.

We follow Andresen (2018) to recover predicted potential changes in spend for *each* hospital in the untreated and treated states, respectively.[39] We then examine heterogeneity in the predicted change in untreated (counterfactual) episode spend, ΔY_h^0, across hospitals and investigate selection into BPCI. Similarly, we also examine heterogeneity in the predicted treatment effect, $\Delta Y_h^1 - \Delta Y_h^0$.

Figure 7 presents the corresponding histograms of selection and treatment effect heterogeneity in panels (a) and (b), respectively. We plot the histograms for participant and non-participant hospitals separately in each panel. The dashed lines mark the correspond-

[39] We use parameters estimated from the MTE analysis to predict potential outcomes for each hospital. Specifically, we have the treated outcome $\Delta Y^1 = \Delta X'\beta^1 + \frac{d-p}{1-p}K^1(p)$ and the untreated outcome $\Delta Y^0 = \Delta X'\beta^0 + \frac{p-d}{p}K^0(p)$.

ing patient-weighted mean values for both groups. Panel (a) shows considerable overlap between the two groups in their counterfactual change in spend values, but the distribution for non-participants is clearly more negative than that for participants. This implies the participant hospitals were adversely selected relative to non-participants, by which we mean that non-participants are predicted to reduce spending more than participants, absent treatment. The difference between the means of the two groups is 0.068 and is highly statistically significant (s.e. of 0.024). This difference term is approximately equal to the difference between the ATT and DD estimates, which is the selection bias term in Equation 3.[40] Figure 7 panel (b) shows that the treatment effect distribution for non-participant hospitals is more negative than that for participants, but with significant overlap.

The finding that non-participant hospitals were "on track" to reduce spending by a greater proportion than participant hospitals runs counter to conventional wisdom, however it is supported by patterns in the raw data. As discussed in Section 6, we detect two mechanisms to reduce episode spending – length of stay of the index hospital stay and use of post-acute care following discharge, with the latter being quantitatively more important. We find a greater decline on both measures among non-participants than among participants in the years prior to BPCI. PAC use declined by 1.64% at BPCI hospitals over 2010–13, but it declined by 2.64% at non-BPCI hospitals over the same period. Note that BPCI hospitals had a greater level of PAC use at baseline (86% versus 81%). Hence, the decline at non-BPCI hospitals was even greater in proportionate terms. The corresponding difference in length of stay between the two groups is small, but is qualitatively similar. Length of stay declined by 0.3 days on average at BPCI hospitals during this period versus 0.31 days at non-BPCI hospitals. Finding a smaller difference on this dimension is intuitive since there were multiple barriers to reducing length of stay prior to BPCI, as discussed previously in Section 6.

The results of the MTE analysis imply substantial unobserved selection of hospitals into BPCI. This suggests caution in relying on the results of such voluntary pilot programs in the future. We test whether we can reduce the magnitude of unexplained variation in both the counterfactual change in spend (ΔY^0) and the treatment effect across hospitals by controlling for a large array of patient, hospital, or market attributes. Here, we test the efficacy of including a number of potentially relevant additional controls that were hitherto held out from the analysis. For example, the number of employees or employees per bed, Medicare and Medicaid share of admissions, whether the hospital entered into an Accountable Care Organization (ACO) arrangement, and the number of technological services offered by the hospital at baseline are seemingly relevant factors that could help

[40] Recall, $DD = ATT + E(\Delta U^0|D = 1) - E(\Delta U^0|D = 0)$. Substituting the empirical estimates, we get $DD \approx -0.11 + (-0.01 - (-0.08))$. Note that this result is not a check for, nor does it inform us about parallel trends with respect to the instrument.

explain variation in hospital decisions to participate or their treatment effects.

Table A.9 presents the results of this analysis. Column 1 presents the share of variance in the estimated ΔY^0 across hospitals that can be explained away by including different factors as controls in the MTE estimation. With the exception of the trend in predicted SNF use, none of the factors can individually explain more than 1% of the unconditional variance. Market fixed effects have the most explanatory power: they eliminate about 20% of variance. This suggests that about 80% of the variation in untreated spend changes is within market. Including market fixed effects along with the most predictive hospital controls still explains only about 30% of the total variance. Column 2 presents the corresponding analysis for the variance in estimated treatment effects across hospitals, with very similar patterns. We conclude that it is difficult to mitigate selection bias in such settings by controlling for the covariates typically available to researchers in standard datasets.

8 Discussion and Conclusion

This paper studies the trade-off inherent in all voluntary programs: the benefit of advantageous sorting on treatment effects versus the cost of introducing selection bias. Our setting is an important national voluntary payment reform for hospitals which aimed to reduce spending on a high-volume surgery. We overcome bias due to endogenous hospital participation by using an instrumental variable approach which exploits plausibly exogenous variation in the financial attractiveness of the program due to the interaction of idiosyncratic program rules and pre-determined hospital attributes. The LATE on spend reduction is three times as large as the naive DD estimate, suggesting large spend reductions, at least for complier hospitals. However, compliers are estimated to be less than 20% of all hospitals.

We then estimate marginal treatment effects to recover the average treatment effect on participants and average treatment effect across all hospitals. This allows us to quantify both unobserved selection bias and sorting on gains. The results imply substantial treatment effect heterogeneity across hospitals, but we cannot reject the null hypothesis of no sorting on gains. This suggests that even sophisticated firms like hospitals may not be aware of their treatment effects with sufficient certainty to benefit from having the option to choose their contract. In contrast, we find evidence of substantial unobserved selection bias. Participants were adversely selected, i.e., the hospitals that chose to participate would have otherwise done worse on spending. On the one hand, this biased the DD estimate in standard program evaluations to understate the ATT, rather than overstate it (which is what policymakers generally suspect in such settings). On the other hand, however, the design garnered participation from hospitals that would otherwise have negligibly reduced spending, so in this respect policymakers succeeded in recruiting the appropriate

hospitals. Taken together, however, program evaluation suffered, with meaningful policy consequences, without the potential benefit from sorting.

Healthcare regulators and insurers have typically limited their design focus on payment reforms to varying financial generosity, assuming that this is sufficient to drive participation into the new contracts. However, a key implication of our results is that they should incorporate other dimensions since multiple factors influence participation into alternate payment contracts, and financial incentives move only a small fraction of firms to join such programs. This insight is relevant to settings beyond healthcare, since voluntary programs are popular policy tools in many sectors of the economy, including education, energy, and environment conservation. Quantifying the role of other factors that may drive participation, such as reputation concerns or an intrinsic motivation to improve, were outside the scope of this study and left as an exercise for future research.

References

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

Abdulkadiroğlu, Atila, Parag A Pathak, and Christopher R Walters, "Free to choose: Can school choice reduce student achievement?," *American Economic Journal: Applied Economics*, 2018, *10* (1), 175–206.

_ , _ , Jonathan Schellenberg, and Christopher R Walters, "Do parents value school effectiveness?," *American Economic Review*, 2020, *110* (5), 1502–39.

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

Agarwal, Rajender, Joshua M Liao, Ashutosh Gupta, and Amol S Navathe, "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*, 2020, *39* (1), 50–57.

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

Andresen, Martin Eckhoff, "Exploring marginal treatment effects: Flexible estimation using Stata," *The Stata Journal*, 2018, *18* (1), 118–158.

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

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

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

Athey, Susan and Guido W Imbens, "The state of applied econometrics: Causality and policy evaluation," *Journal of Economic Perspectives*, 2017, *31* (2), 3–32.

Baicker, Katherine and Jacob A Robbins, "Medicare payments and system-level healthcare use: the spillover effects of Medicare managed care," *American Journal of Health Economics*, 2015, *1* (4), 399–431.

_ , Michael E Chernew, and Jacob A Robbins, "The spillover effects of Medicare managed care: Medicare Advantage and hospital utilization," *Journal of Health Economics*, 2013, *32* (6), 1289–1300.

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

Brinch, Christian N, Magne Mogstad, and Matthew Wiswall, "Beyond LATE with a discrete instrument," *Journal of Political Economy*, 2017, *125* (4), 985–1039.

Cicala, Steve, David Hémous, and Morten Olsen, "Adverse selection as a policy instrument: unraveling climate change," Technical Report, Universities of Chicago, Zurich, and Copenhagen 2020.

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg, "Who

benefits from universal child care? Estimating marginal returns to early child care attendance," *Journal of Political Economy*, 2018, *126* (6), 2356–2409.

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

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., "Association between hospital participation in a Medicare bundled payment initiative and payments and quality outcomes for lower extremity joint replacement episodes," *JAMA*, 2016, *316* (12), 1267–1278.

Dunn, Abe, Joshua D Gottlieb, Adam Shapiro, Daniel J Sonnenstuhl, and Pietro Tebaldi, "A denial a day keeps the doctor away," Technical Report, National Bureau of Economic Research 2021.

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

Finkelstein, Amy, Yunan Ji, Neale Mahoney, and Jonathan Skinner, "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*, 2018, *320* (9), 892–900.

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

Heckman, James J and Edward J Vytlacil, "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments," *Handbook of Econometrics*, 2007, *6*, 4875–5143.

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

Imbens, G and J Angrist, "Estimation and identification of local average treatment effects," *Econometrica*, 1994, *62*, 467–475.

Jack, B Kelsey and Seema Jayachandran, "Self-selection into payments for ecosystem services programs," *Proceedings of the National Academy of Sciences*, 2019, *116* (12), 5326–5333.

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

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

Koenig, Lane, Chaoling Feng, Fang He, and Jennifer T Nguyen, "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*, 2018, *33* (9), 2764–2769.

Kowalski, Amanda E, "Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform," *The Review of Economics and Statistics*, 2021.

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

_ , "CMS Bundled Payments for Care Improvement Initiative Models 2-4: Year 3 Evaluation & Monitoring Annual Report," Technical Report October 2017.

_ , "CMS Bundled Payments for Care Improvement Initiative Models 2-4: Year 4 Evaluation & Monitoring Annual Report," Technical Report June 2018.

_ , "CMS Bundled Payments for Care Improvement Initiative Models 2-4: Year 5 Evaluation & Monitoring Annual Report," Technical Report October 2018.

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

MedPAC, "Report to the Congress: Medicare and the healthcare delivery system," Technical Report June 2021.

Milad, Marina A, Roslyn C Murray, Amol S Navathe, and Andrew M Ryan, "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*, 2022, *41* (4), 540–548.

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

_ , 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., "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*, 2020, *39* (1), 58–66.

_ , Joshua M Liao, Daniel Polsky, Yash Shah, Qian Huang, Jingsan Zhu, Zoe M Lyon, Robin Wang, Josh Rolnick, Joseph R Martinez, and Ezekiel J Emanuel, "Comparison of hospitals participating in Medicare's voluntary and mandatory orthopedic bundle programs," *Health Affairs*, 2018, *37* (6), 854–863.

Newhouse, Joseph P, Mary Price, J Michael McWilliams, John Hsu, and Thomas G McGuire, "How much favorable selection is left in Medicare Advantage?," *American Journal of Health Economics*, 2015, *1* (1), 1–26.

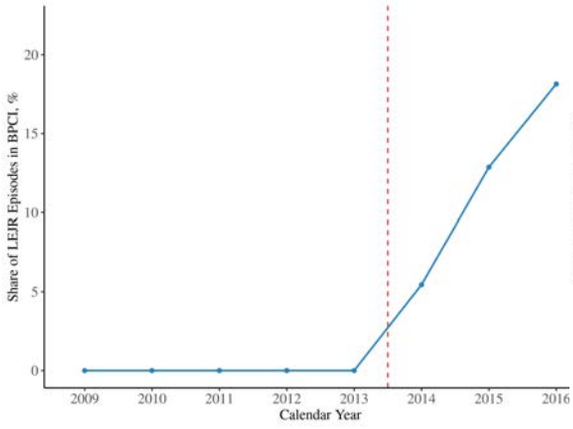
Rathi, Vinay K and J Michael McWilliams, "First-year report cards from the Merit-based Incentive Payment System (MIPS): what will be learned and what next?," *JAMA*, 2019, *321* (12), 1157–1158.

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

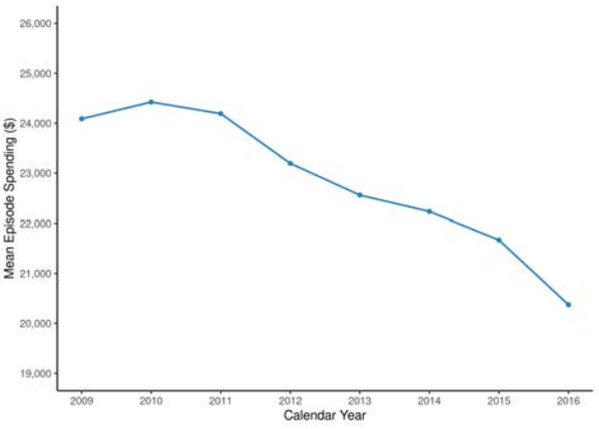
Werner, Rachel M, Jonathan T Kolstad, Elizabeth A Stuart, and Daniel Polsky, "The effect of pay-for-performance in hospitals: lessons for quality improvement," *Health Affairs*, 2011, *30* (4), 690–698.

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

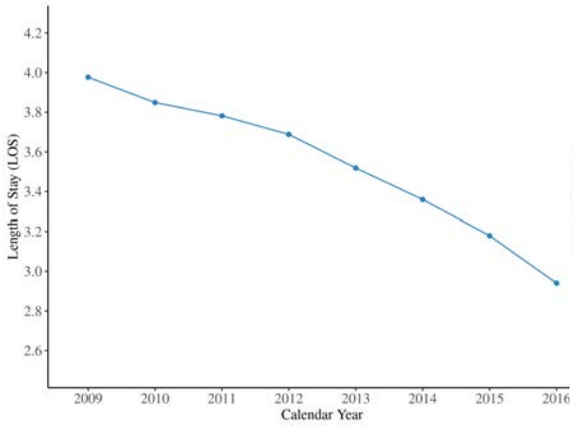
(a) BPCI share of LEJR cases



(b) Mean episode spend



(c) Mean length of stay



(d) Share of episodes with PAC use

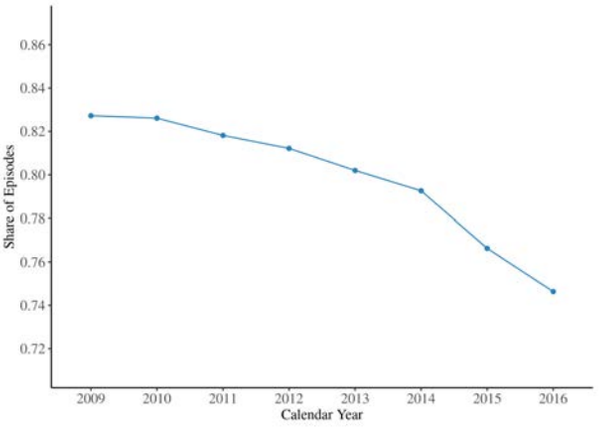


Figure 1: Trends in Medicare LEJR surgeries

Note: This figure presents trends in 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 spend, 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 rehab, 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.

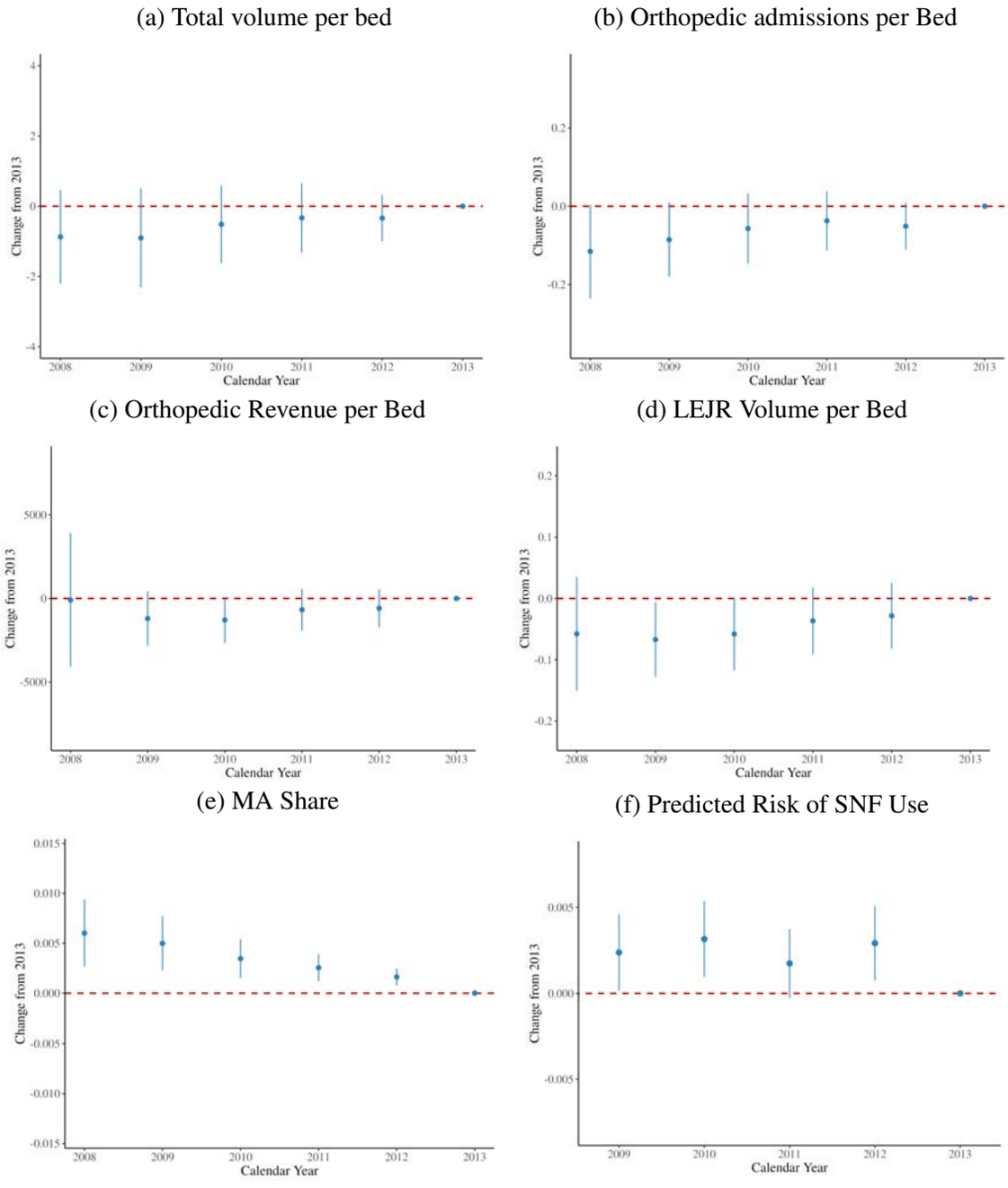
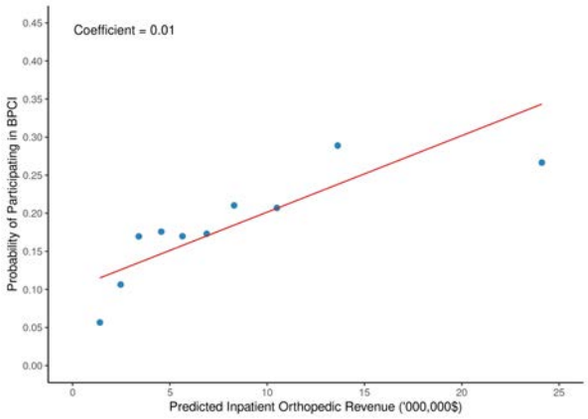


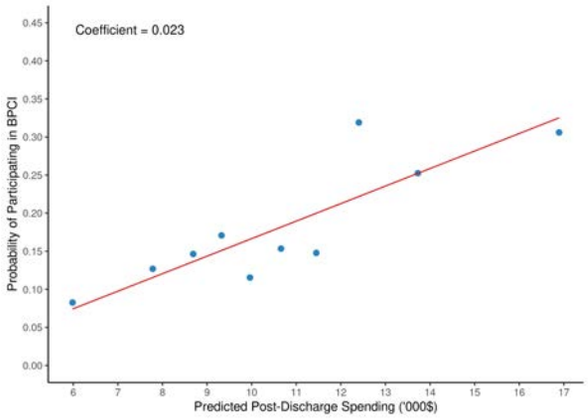
Figure 2: Trends prior to BPCI

Note: This figure presents the estimated dynamic trends in time-varying hospital and patient characteristics prior to BPCI for participating hospitals. The coefficients are obtained by estimating Equation 4. Total volume in panel (a) is taken from annual 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 fee-for-service LEJR volume. Panel (e) presents the trends in MA penetration, computed as described in Section 4.2. 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) Orthopedic revenue



(b) Post-discharge orthopedic episode spend



(c) Joint distribution

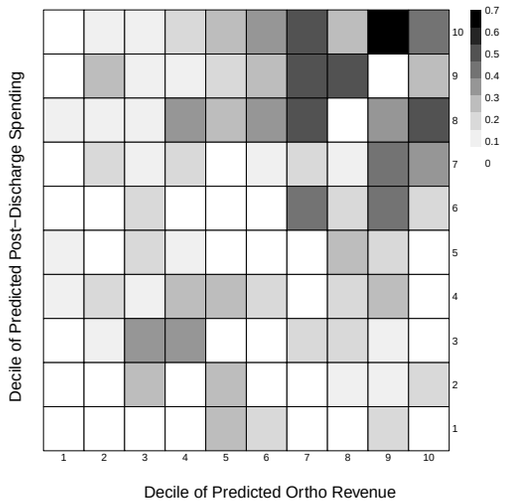
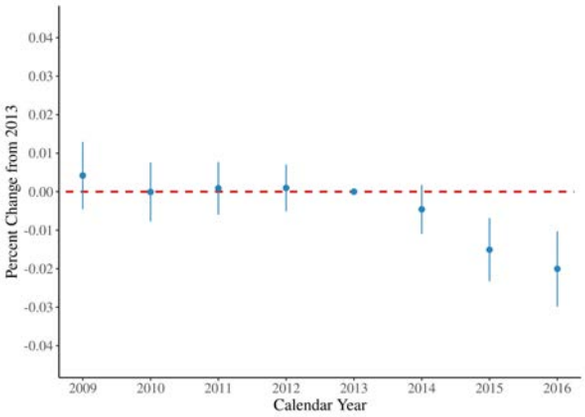


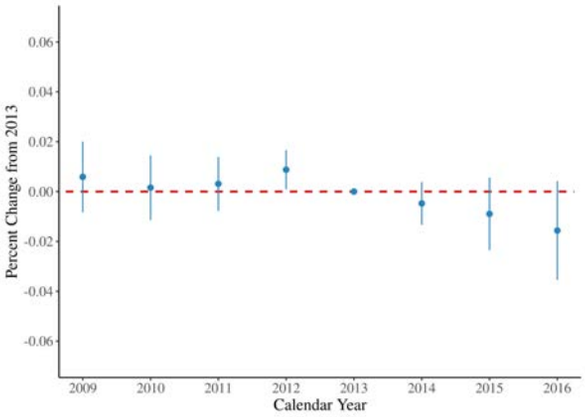
Figure 3: Baseline orthopedic performance and BPCI participation

Note: This figure presents the association between baseline orthopedic performance and BPCI participation over 2014–16. Panel (a) presents a binned scatter plot of the relationship between the hospital's Medicare orthopedic revenue on the X-axis and participation on the Y-axis. Panel (b) replaces orthopedic revenue with mean post-discharge episode spend in orthopedic episodes. Panel (c) presents the variation in participation against the joint distribution of orthopedic revenue and post-discharge spend. Orthopedic performance measures are predicted values generated using patient and hospital attributes from data in 2007, described in Section 5. Each hospital is weighted by the number of LEJR cases prior to BPCI.

(a) Episode spend (log)



(b) LEJR length of stay (log)



(c) Use of any PAC

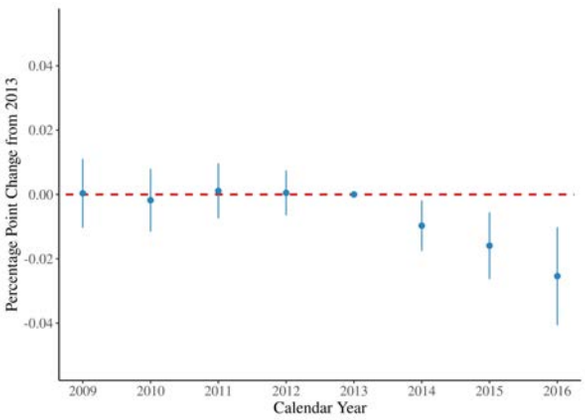


Figure 4: Effects on spend and utilization of care

Note: This figure presents dynamic effects on episode spend (panel a), index stay length (panel b), and the use of post-acute care (panel c). The coefficients are obtained by estimating a reduced form variant of Equation 4 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.

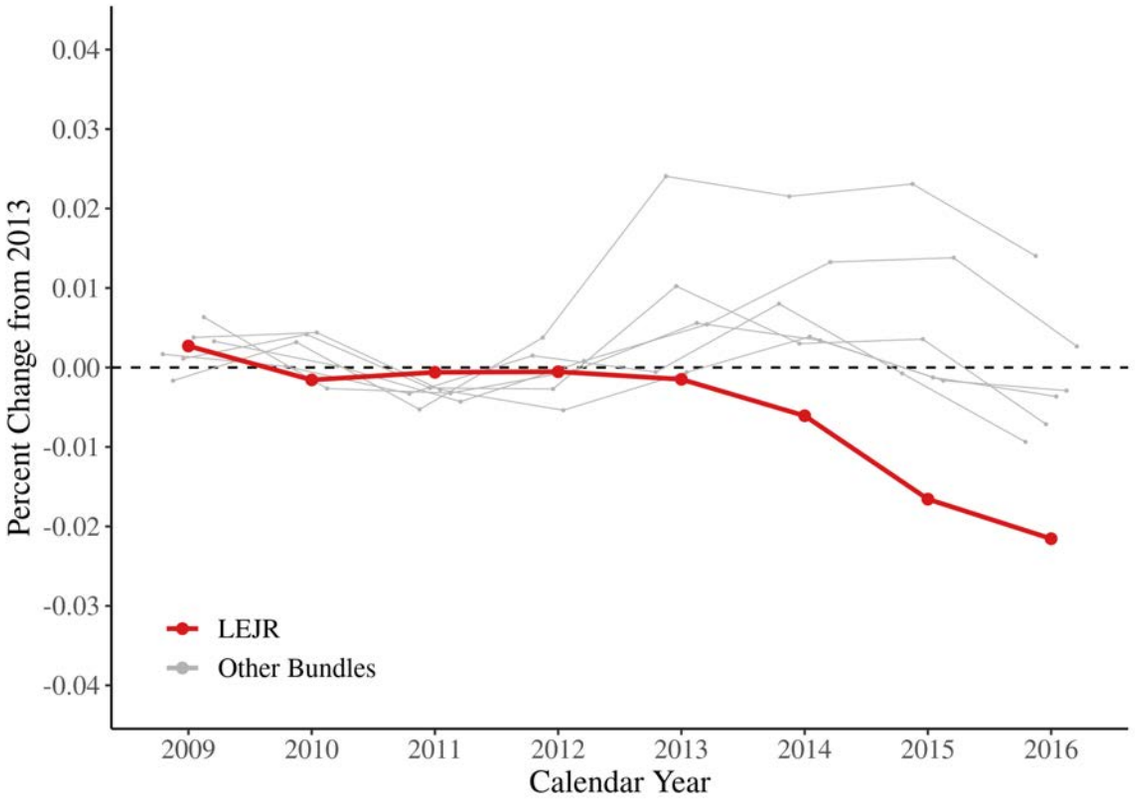
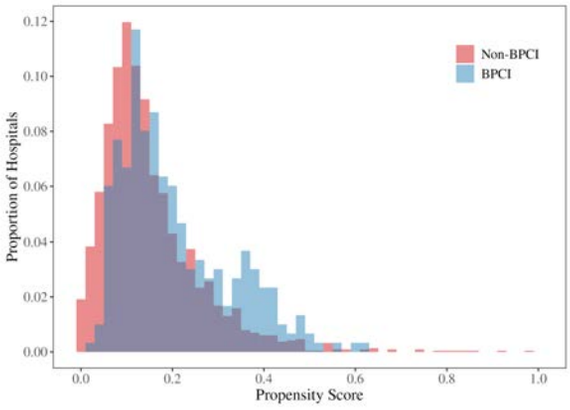


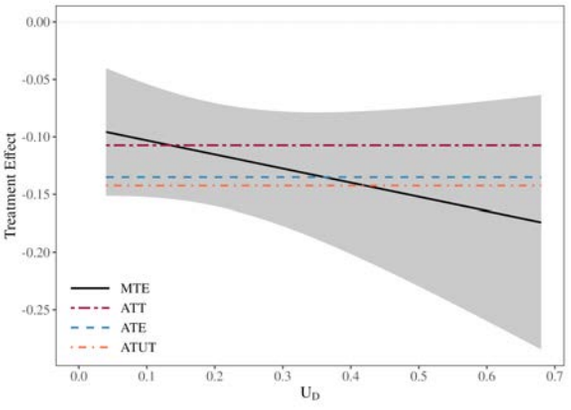
Figure 5: LEJR versus other clinical bundles

Note: This figure presents the estimated dynamic reduced form effects on 90-day episode spending for lower extremity joint replacement (LEJR) in bold red and six other clinical episode bundles (congestive heart failure, pneumonia, chronic obstructive pulmonary disease, sepsis, urinary tract infections, and upper extremity joint replacement) in light grey. The models include patient controls, hospital and HRR by year fixed effects. We de-mean the coefficients by the 2009–12 condition-specific average value 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 5 Panel B.

(a) Treatment propensity



(b) MTE curve



(c) Point estimates

A. OLS & IV

OLS	2SLS	MTE with IV-Weights
-0.027^{***}	-0.122^{***}	-0.109^{***}
(0.005)	(0.03)	(0.027)

B. Population TEs

ATE	ATT	ATUT
-0.135^{***}	-0.107^{***}	-0.142^{***}
(0.032)	(0.027)	(0.036)

(d) Weights

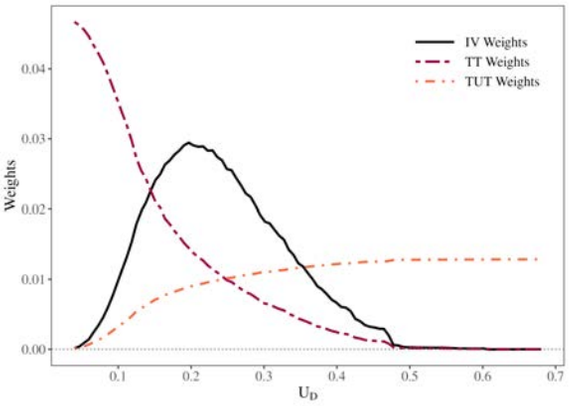
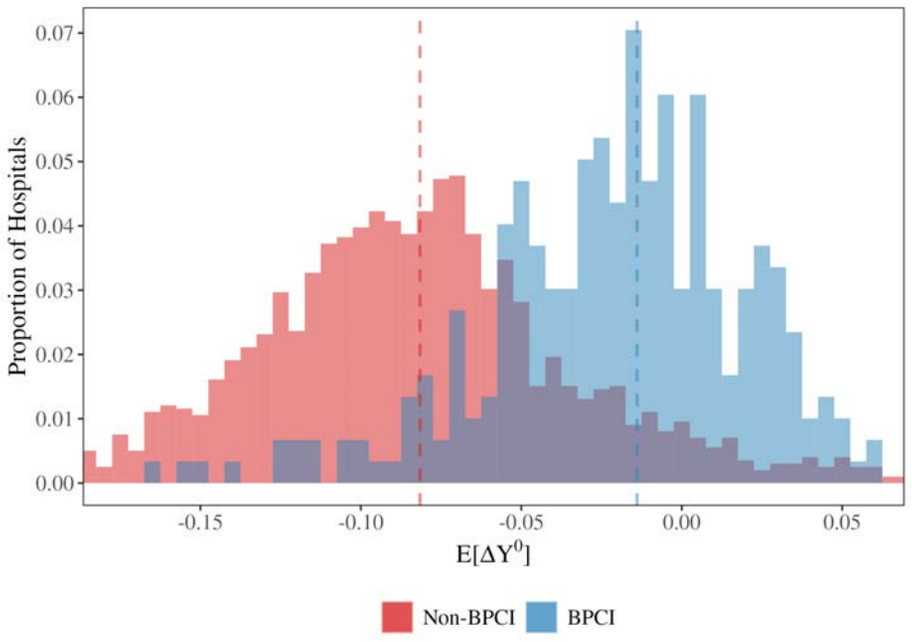


Figure 6: Marginal Treatment Effects

Note: This figure presents results from the MTE analysis described in Section 7. All models are estimated on a cross-sectional hospital-level sample of 2,312 hospitals that perform surgeries pre- and post-BPCI, and fall within the common overlap region of participation propensity. The dependent variable is the pre-post change in log episode spend for each hospital, $\Delta logY_h = log\bar{Y}_{h,post} - log\bar{Y}_{h,pre}$. The covariates are similarly computed for each hospital. Panel (a) plots the histogram of predicted propensity of treatment (BPCI participation) for the treated and comparison groups, obtained by estimating Equation 8 via probit (see Section A.5). Panel (b) presents the estimated MTE curve and shows the point estimates for the population treatment effects: ATT, ATE, and ATUT. Panel (c) displays various point estimates with standard errors obtained using block bootstrap. Panel (c)A presents the OLS, 2SLS, and IV estimate obtained as a weighted average of MTE coefficients. Panel (c)B presents the population treatment effect estimates, all computed as weighted averages of the MTE coefficients. Panel (d) plots the weights associated with the ATT, ATU, and LATE parameters.

(a) Predicted untreated counterfactual outcomes



(b) Predicted treatment effects

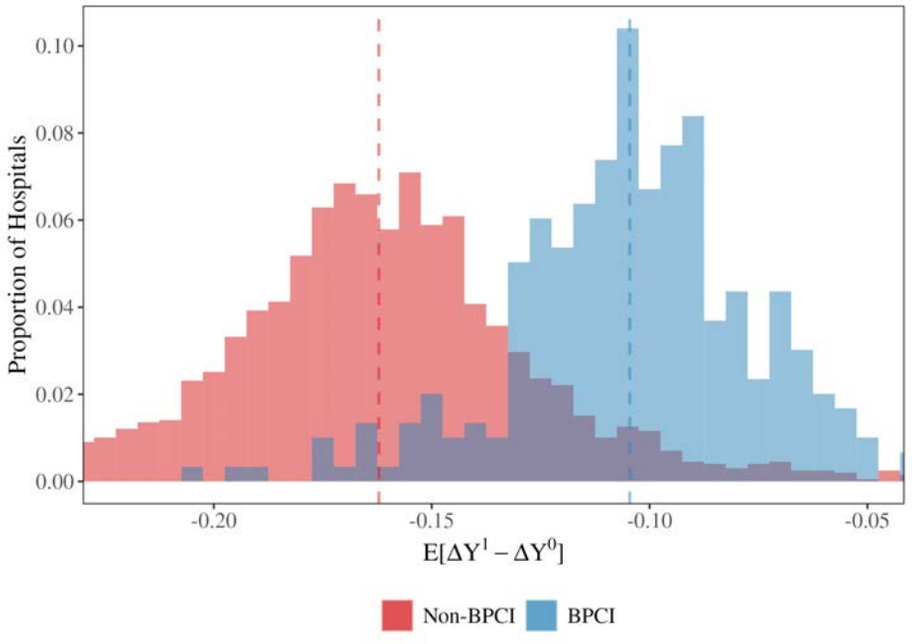


Figure 7: Selection and treatment effect heterogeneity

Note: This figure presents posterior estimates of hospital-specific counterfactual values computed using the MTE model coefficients following Andresen (2018). Panel (a) presents histograms of the potential untreated outcomes (ΔY^0) for the treated and comparison hospitals, where the difference is taken between the post-BPCI (2014-16) and pre-BPCI (2009-13) periods. Panel (b) presents the histograms of the predicted treatment effect ($\Delta Y^1 - \Delta Y^0$) for the treated and comparison hospitals.

Table 1: Summary statistics

	Never BPCI (1)	Ever BPCI (2)	All Hospitals (3)
A. 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
Annual Medicare LEJR cases per bed	0.77	0.683	0.759
Annual total volume per bed	43.947	48.478	44.499
Annual Medicare Ortho. revenue per bed	25,861.3	27,039.6	26,004.6
% MA Penetration	25.21	25.33	25.22
B. 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 Institutional PAC spend ('000$)	6.9	7.8	7.1
Mean HH spend ('000$)	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. Since participants are substantially larger in bed size, we present volume and revenue measures per bed.

Table 2: Balancing tests

	Overall Trend (1)	$d_h \cdot T_t$ (2)	$z_h \cdot T_t$ (3)
Total volume per bed	-0.373***	0.179	0.0158
	(0.066)	(0.153)	(0.221)
Ortho. admissions per bed	-0.034***	0.020*	-0.001
	(0.006)	(0.011)	(0.010)
Ortho. revenue per bed	432***	82.5	220
	(81.9)	(317)	(293)
LEJR volume per bed	-0.010***	0.014**	0.002
	(0.003)	(0.007)	(0.006)
MA penetration (%)	1.25***	-0.117***	0.008
	(0.020)	(0.035)	(0.022)
Predicted risk of SNF use (%)	-0.233***	-0.127**	-0.022
	(0.018)	(0.052)	(0.051)

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. Column (1) reports the coefficient γ corresponding to the overall linear trend across hospitals from the following model: $Y_{ht} = \alpha_h + \gamma\, T_t + \varepsilon_{ht}$. Columns 2 and 3 report differential linear trends for hospitals that participated or had high potential for financial gain based on baseline orthopedic performance, respectively. The coefficients in column 2 corresponds to β from the following specification: $Y_{ht} = \alpha_h + \alpha_{m(h)t} + \beta\, T_t \cdot d_h + \epsilon_{ht}$. Column 3 presents the coefficient from a variant of this model in which we replace the participation indicator, d_h, with the instrument, z_h. Medicare Advantage (MA) penetration is computed as described in Section 4.2. 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 were estimated on aggregated hospital-year level data, except for predicted SNF risk, which was estimated on patient-level data.

Table 3: Episode spending

	$\log(\text{Spending})_{iht}$				$\Delta\log(\text{Spending})_h$	
	(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
$\bar{y}$	9.908	9.908	9.908	9.908	-0.084	-0.084
Mean BPCI × Post	0.067	0.067	0.067	0.067	–	–
Mean Ever BPCI	–	–	–	–	0.172	0.172
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. 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 column 5–6. The outcome for the cross-sectional analysis is the change in mean log episode spend between 2014–16 and 2009–13. The cross-sectional sample drops 92 never-BPCI hospitals which do not perform LEJR surgeries post-2013. Hospitals included in the cross-sectional regression models are weighted by pre-BPCI LEJR volume (2009–2013).

Table 4: Utilization of care

	Index stay		PAC use			
	log(LOS) (1)	LOS ≤ 2 (2)	Any PAC (3)	HHA (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
$\bar{y}$	1.137	0.22	0.799	0.376	0.103	0.402

Note: This table presents the estimated effects on utilization of care obtained using the primary specification, discussed in Section 5, and corresponding to the coefficients in Table 3 Col. 2. Columns 1 and 2 present effects on outcomes related to length of stay for the index LEJR surgery. Columns 3–6 present effects on PAC use in the 90 days following discharge. 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 5: Effect on spend for LEJR versus other bundles

	LEJR (1)	CHF (2)	PNA (3)	COPD (4)	Sepsis (5)	UTI (6)	UEJR (7)	Pooled (8)
A. DD								
BPCI × Post	-0.031^{***}	-0.006	-0.009	-0.004	-0.008	-0.019^{*}	-0.023^{*}	-0.006
	(0.005)	(0.006)	(0.006)	(0.007)	(0.005)	(0.010)	(0.012)	(0.004)
B. LATE								
Reduced Form	-0.014^{***}	0.0002	-0.001	0.016^{***}	0.002	0.010^{*}	-0.002	0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.005)	(0.005)	(0.003)
First stage F-stat.	17.7	12.2	3.2	3.0	0.9	1.1	0.1	5.5
Observations	2,574,735	2,103,288	2,388,107	1,663,846	2,424,028	1,270,679	225,808	10,075,756
Mean BPCI × Post	0.067	0.036	0.022	0.021	0.029	0.016	0.014	0.026
N. BPCI Hospitals	302	174	136	130	119	88	37	247
$\bar{y}$	22,872	19,155	18,366	14,899	24,613	17,590	19,577	19,390

Note: This table presents estimated effects on log episode spending for LEJR (Column 1) and six other clinical bundle episodes included in the BPCI program (Columns 2–7). These are, respectively, congestive heart failure (CHF), pneumonia (PNA), chronic obstructive pulmonary disease (COPD), sepsis, urinary tract infections (UTI), and upper extremity joint replacement (UEJR). The first five are the largest bundles in BPCI by hospital participation (after LEJR) while the sixth pertains to orthopedics. Column 8 presents results on a pooled sample across these 6 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 Bias in DD due to unobserved selection

As discussed briefly in Section 3, the DD estimate can deviate from the ATE due to two sources of unobserved selection. Figure A.3 illustrates this identification problem under three potential scenarios using a stylized model. Suppose the outcome, ΔY, is the change in episode spend post-BPCI. We also refer to this as the trend in spend. Without loss of generality, we suppose that 30% of hospitals receive treatment, and this is held fixed across scenarios for simplicity. Hospitals are ordered in increasing resistance to treatment, so that those to the left are most likely to participate and vice-versa.

Panel (a) presents the experimental ideal as a benchmark: no selection bias and constant treatment effects across hospitals. Under this scenario, all units would experience identical trends in spending absent treatment (ΔY^0), normalized to zero here for simplicity. This outcome is observed only for the non-participants, and is therefore shown as a dashed line for the participants. The potential treated trend in spending (ΔY^1) is assumed to be -5% for all units. The DD estimate, which is the difference between the observed mean trends of the treated and untreated units, is therefore -5%. In this ideal case, the DD estimate coincides with both the ATT and ATE, which cannot otherwise be computed directly from the data. This result also holds in the presence of treatment effect heterogeneity, as long as ΔY^1 is uncorrelated with participation status.

Panel (b) illustrates a scenario with favorable selection bias, the possibility that the treated units would have experienced greater spending reduction even in the absence of treatment. We continue to assume constant treatment effects across all units, -4% in this case. However, due to favorable selection, the DD estimate is biased for the ATT, overstating it.[41] On the other hand, participants could also be adversely selected, in which case the DD estimate would be biased toward zero and understate the ATT. Panel (c) illustrates this possibility, and is the mirror image of Panel (b).

Finally, Panel (d) considers the possibility of treatment effect heterogeneity and sorting on treatment gains. We assume away selection bias for simplicity (ΔY^0 is identical across all units). This ensures that the DD estimator recovers the ATT (both are -4.7%). However, due to favorable sorting, the DD and ATT exceed the ATE, which is computed across all treated and untreated units, and equals -4%. While we did not illustrate the possibility of reverse or negative sorting on treatment gains, it is straightforward to visualize it as the mirror image of Panel (d). In this case, the DD and ATT would understate the average effect on the full population. If both selection bias and sorting on gains are present, then it

[41] The DD estimate is -4.5%. $DD = E(\Delta Y^1|D=1) - E(\Delta Y^0|D=0) = -4.85 - (-0.35) = -4.5$. However, due to constant treatment effects, the ATT and ATE are both -4%. $ATT = E(\Delta Y^1 - \Delta Y^0|D=1) = -4.85 - (-0.85) = -4$.

becomes difficult to predict the sign of the bias between the DD estimate and the population treatment effects.

A.2 Sample 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 institional 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 home health agency fee-for-service claims (we have access to home health agency claims from 2007–2016). The claims data include beneficiary demographics (age, race, gender), enrollment information, death dates, and zip code of residence as well as a record of medical services. Information about medical servies 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 payment by Medicare, or involve a service for which the beneficiary left against medical advice. We further remove a small number claims where the admission date occurs after discharge date or date of the beneficiary's death.

To define episodes, we first identify candidate index hospitalizations at acute care hospitals paid under the inpatient prospective payment system (IPPS). To qualify as an index stay, we require that 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 longer than 365 days.
4. Does not involve a beneficiary enrolled in Medicare for end stage renal disease.
5. Is not a transfer from another inpatient hospital. In other words, when a transfer is involved, the index hospitalizations are the first hospital at which the beneficiary receives care.
6. Have a beneficiary and hospital that can both be assigned a hospital referral region (HRR) by zip code.

After defining candidate index hospitalizations, we 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 part 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 09/30/2016.

Our main analysis focuses on lower extremity joint replacement (LEJR) episodes defined by MS-DRG 469 or 470.[42] In placebo tests, we additionally analyze episodes for patients admitted for cholecystectomy and a set of other conditions with substantial BPCI participation. 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). 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 the diagnosis codes used by CMS for the comprehensive care for joint replacement (CJR) program.[43] In analyses of revisions to joint replacements, we define joint revision using DRG 466, 467, and 468.

In the main analytic samples for these conditions, we further require that beneficiaries be over 65 years old at the start of the episode. 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.

A.3 Construction of the instrumental variable

To capture baseline exposure to the 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 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 far before the intervention and to use the full set of orthopedic episodes to limit concern that the 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. The patient attributes include age, sex, race,

[42] For discharges occurring prior to 10/01/2007, LEJR was categorized as DRG 544.

[43] The diagnosis codes are available from CMS here.

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, 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 5.1 and displayed visually in figure 3, we find participation is much more likely for the set of hospitals with high predicted revenue and predicted post-discharge spending. Accordingly, 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.

A.4 Selection on unobservables

LEJR surgeries are elective in nature and the choice of hospital is often driven by the orthopedic surgeon's referral patterns and practice locations. Hence, hospitals could change their relationships with referring physicians in order to source healthier patients (eg., target more affluent neighborhoods). Hence, we test for changes in the mean patient distance to hospital. We compute distance between the 5-digit zipcodes of the patient and the hospital of the index surgery. Figure A.9 panels (e) and (f) present the reduced form event studies with distance in levels and logs, respectively.

The second exercise uses a different approach to test the same hypothesis– whether hospitals have changed their patient mix based on location and demographics. We estimate a patient preference for a BPCI hospital based on just their zip code information, as well as incorporating patient 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. We use these probabilities as the outcome in this analysis and examine if participating hospitals are attracting patients who otherwise would have less likelihood of being served by a BPCI hospital. We continue to find null effects, which are presented in Figure A.9 panels (g) and (h). Table A.6 Panel B presents the corresponding point estimates.

A.5 MTE estimation details

As described in Section 7, we estimate marginal treatment effects on a cross-sectional hospital-level sample by collapsing the patient-level data and using the difference between mean 2014-16 and 2009-13 log episode spend values as the outcome of interest. We use the corresponding hospital-level vector of covariates, ΔX_h. All remaining unobserved factors are denoted by ΔU_h. Equation 7 below presents the potential outcome in states $j = 0,\ 1$, corresponding to BPCI participation and non-participation, respectively.

$$\Delta Y_h^j = \mu^j(\Delta X_h) + \Delta U_h^j, \quad j = 0, 1, \tag{7}$$

where the $\mu^j(\cdot)$ are unspecified functions. We propose a latent selection model of how hospitals choose to participate in BPCI, based on observed and unobserved factors:

$$\begin{aligned} &d_h^* = Z_h'\delta - V_h, \\ &d_h = 1 \; if \; d_h^* \geq 0,\ d_h = 0 \quad \text{otherwise}, \end{aligned} \tag{8}$$

where $Z_h = (\Delta X, z)_h$ includes the covariates, X, and the binary instrument, z_h, which is excluded from the outcome equation 7 and satisfies Assumption 1. The use of a fixed coefficient, δ, in the selection model ensures the monotonicity condition is satisfied by construction. V_h enters negatively and is therefore interpreted as the unobserved resistance to participating in BPCI. Since the instrument z_h represents the financial incentive to participate, we interpret V_h as capturing unobserved non-financial factors that could affect a hospital's participation decision, such as the intrinsic motivation to innovate. Following the MTE literature, we transform equation 8 into the quantiles of the distribution of V_h:

$$d_h = 1 \implies Z_h'\delta - V_h \geq 0 \implies Z_h'\delta \geq V_h \implies \Phi(Z_h'\delta) \geq \Phi(V_h), \tag{9}$$

where Φ is the cumulative distribution function of V_h. We interpret $\Phi(Z_h'\delta)$ as the propensity score, the probability that a hospital with observed characteristics, Z_h, participates in the program, and denote it as $P(Z_h)$. $\Phi(V_h)$ represents the quantiles of unobserved resistance to treatment, and is denoted as U_{D_h}, a uniformly distributed variable. Equation 9 implies that hospitals are on the margin of participation when their propensity score $P(Z_h)$ equals U_{D_h}, their resistance to treatment. In the following equations we suppress the subscript h for expositional ease. We define the marginal treatment effect for a hospital with covariates x and propensity p as:

(10)
$$\begin{aligned} MTE(\Delta X = x, U_D = p) &= \mathbb{E}[\Delta Y^1 - \Delta Y^0 | \Delta X = x, U_D = p] \\ &= \mu^1(\Delta X) - \mu^0(\Delta X) + \mathbb{E}(U^1 - U^0 | \Delta X = x, U_D = p) \\ &= \mu^1(\Delta X) - \mu^0(\Delta X) + k(x, p), \end{aligned}$$

where $k(x,p) = k^1(x,p) - k^0(x,p)$. MTE$(x,p)$ is the average treatment effect for hospitals on the margin of treatment. So far, the model has relied on assumptions 1 and 2 used in the LATE analysis. We now make two parametric assumptions to enable the estimation of treatment effects for hospitals outside the complier group.

ASSUMPTION 3 (Additive separability):

$$\mathbb{E}(U^j | \Delta X = x, U_D = p) = \mathbb{E}(U^j | U_D = p) = k^j(p).$$

Assumption 3 implies that the unobserved component of the MTE does not depend on X. Hence, the MTE is additively separable in ΔX and U_D. This considerably eases data requirements to estimate the MTE curve since the slope of the curve in U_D does not depend on X. The MTE literature typically has made this assumption (Cornelissen et al., 2018). Additive separability is standard in conventional IV analysis as well, and implicit in our 2SLS model in equation 6. Accordingly, the MTE can be rewritten as

(11)
$$MTE(\Delta X = x, U_D = p) = \underbrace{\mu^1(\Delta X) - \mu^0(\Delta X)}_{\text{Observed}} + \underbrace{k^1(p) - k^0(p)}_{\text{Unobserved}}.$$

ASSUMPTION 4 (Linearity): Our primary models assume linear parametric functional forms for both the observed and unobserved components. We assume $\mu(\cdot) = \Delta X'\beta$. We follow Andresen (2018) in choosing the parametric forms for $k^j(p)$.[44]

We first estimate the selection model in Equation (8) using a probit model and obtain $\hat{p} = \Phi(Z'\hat{\delta})$ for each hospital. We then estimate the stacked outcome equation (12) below

(12)
$$\Delta Y = \Delta X'\beta^0 + K^j(p) + d\big[\Delta X(\beta^1 - \beta^0) + K^j(p)\big] + \epsilon$$
$$\text{where } K^j(p) = \begin{cases} K^1(p) & \text{if } d = 1 \\ K^0(p) & \text{if } d = 0, \end{cases}$$

[44] Specifically, $k^1(p) = \pi_1(p - \frac{1}{2})$ and $k^0(p) = \pi_0(p - \frac{1}{2})$, where π_1 and π_0 are parameters to be estimated. The corresponding selection control functions are $K^1(p) = \pi_1\left(\frac{p-1}{2}\right)$ and $K^0(p) = \pi_0\left(\frac{p}{2}\right)$.

Additional figures and tables

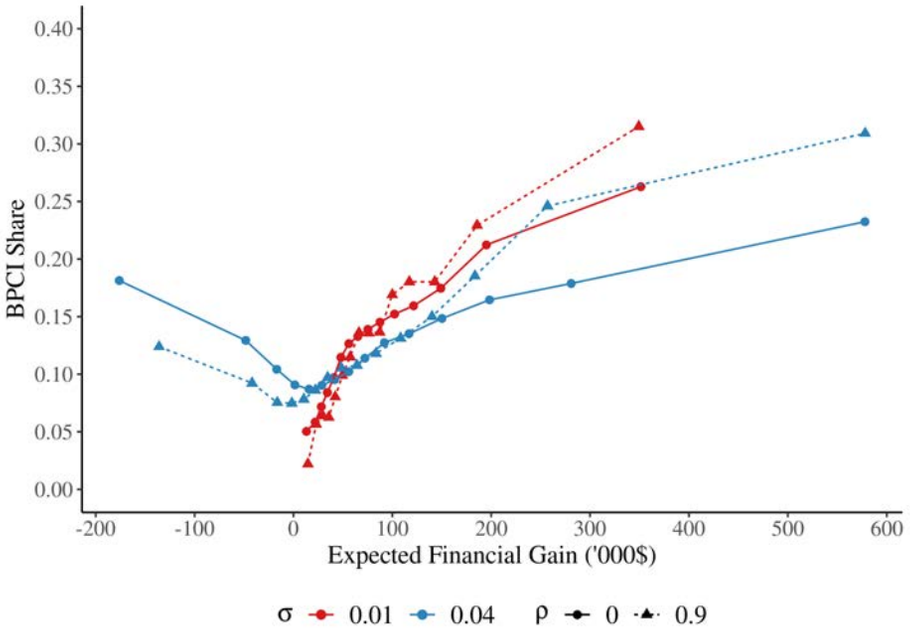


Figure A.1: Simulated financial gain and BPCI participation

Note: This figure presents the relationship between the expected annual gain from BPCI (in thousands of 2016 dollars) and the share of hospitals that participate. We simulate the expected gain. We first approximate the target price used by CMS. The target price for each hospital-year is computed by multiplying a hospital's pre-period spending by the ratio of national average spending in year t to national average spending in the pre-period (to account for secular spending trends). Then, CMS applies a 2% discount to compute the final target price. To reflect the information a hospital has when deciding to join, we use national spending in 2010–2012 to compute a linear trend in episode spending. We then use a hospital's average pre-period spending and the trend to compute un-discounted target prices for each hospital-year, p^u_{ht}. We also subtract a 2% discount at the end to get the discounted target price, $p^d_{ht} = 0.98 \times p^u_{ht}$. To get a hospital's expected spending per episode, we start from the undiscounted target price and generate a random hospital specific percent savings parameter γ_h. We assume $\gamma_h \sim \mathcal{N}(0.05, \sigma^2)$ to reflect previous ATT estimates of the BPCI program. We consider the possibility of low and high uncertainty in spending reduction, by assuming $\sigma = 0.01$ and $\sigma = 0.04$, respectively, indicated by the red and blue curves. To reflect the fact that higher spending hospitals may generate greater savings, we simulate γ_h to have an empirical correlation ρ with hospitals' pre-period observed spending. The lines with circle markers assume modest correlation of a hospital's savings with its pre-period spending ($\rho = 0$). The lines with triangular markers assume high correlation of a hospital's savings with its pre-period spending ($\rho = 0.9$). Expected spending per episode is $s_{ht} = (1 - \gamma_h) \times p^u_{ht}$. To compute annual expected gain for each hospital, we assume annual volume is fixed at the average 2010–2012 level, v_h, so that expected gain can be computed as $\frac{1}{3}\sum_{t=2014}^{2016} v_h(p^d_{ht} - s_{ht})$. Appendix Table A.1 presents point estimates for regressions of BPCI participation on expected financial gain.

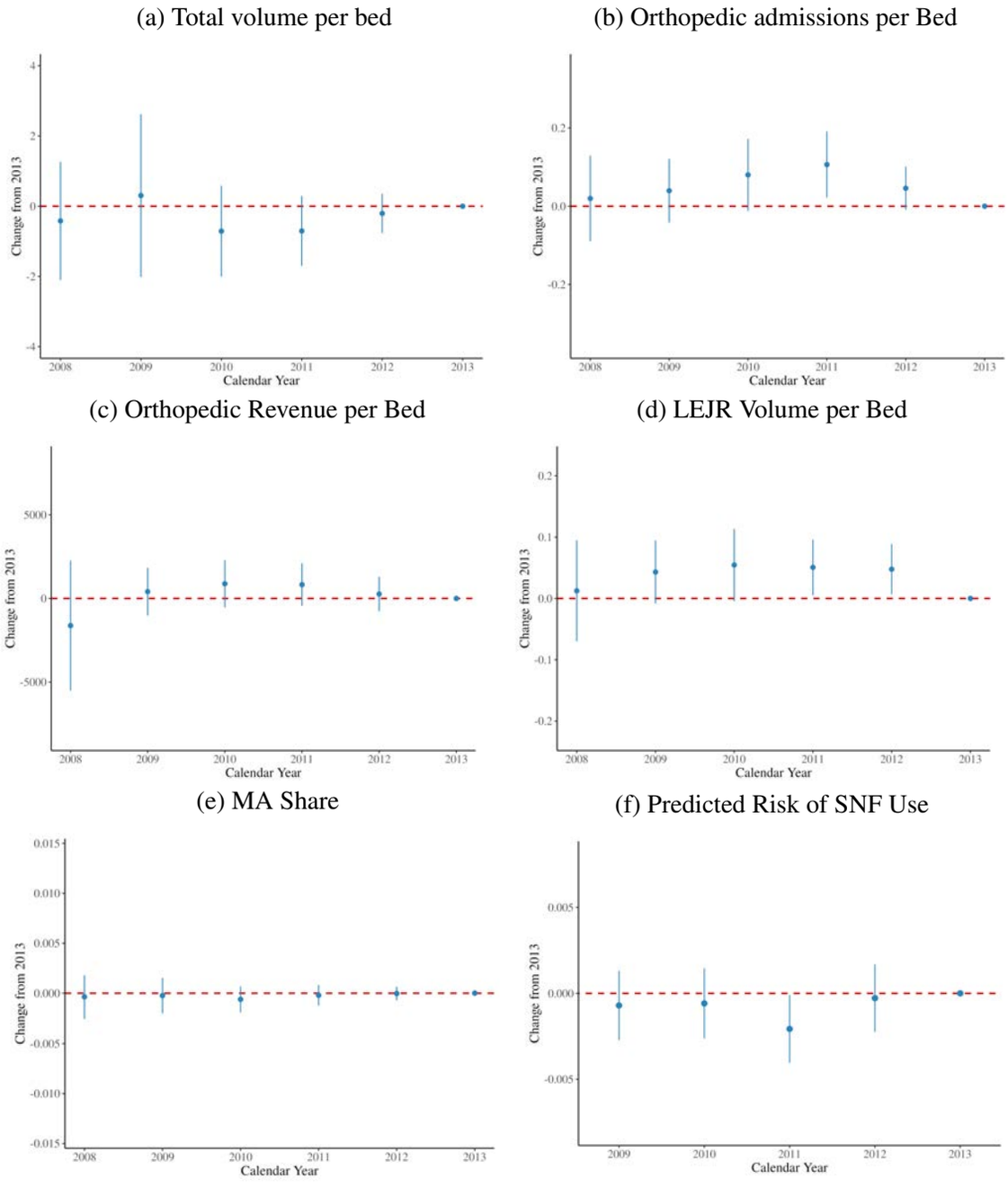
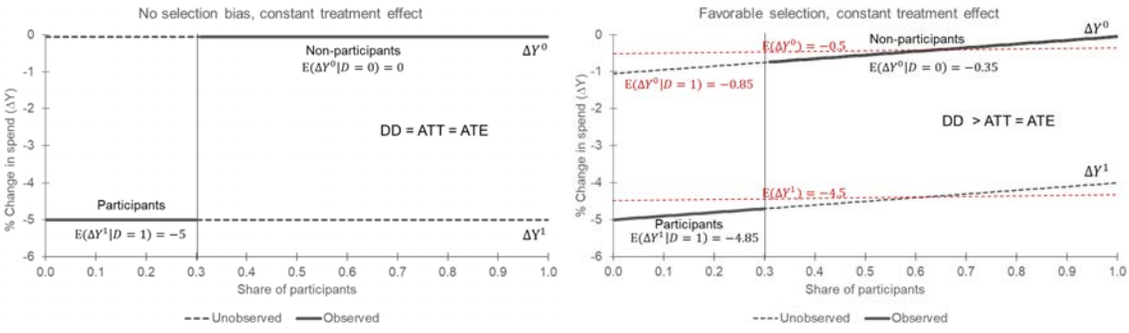


Figure A.2: Trends prior to BPCI (reduced form)

Note: This figure presents the estimated dynamic trends in time-varying hospital and patient characteristics prior to BPCI for those with high potential financial gain . The coefficients are obtained by estimating a variant of Equation 4 where participation, d_h, is replaced by the instrument, z_h. Total volume in panel (a) is taken from annual 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 LEJR volume. Panel (e) presents the trends in MA penetration, computed as described in Section 4.2. 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) No selection or sorting on gains

(b) Favorable selection, no sorting



(c) Unfavorable selection, no sorting

(d) Sorting on gains, no selection

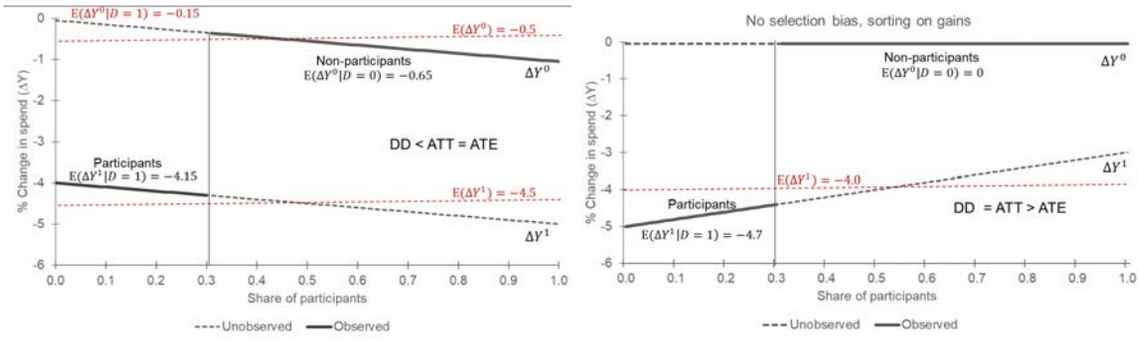
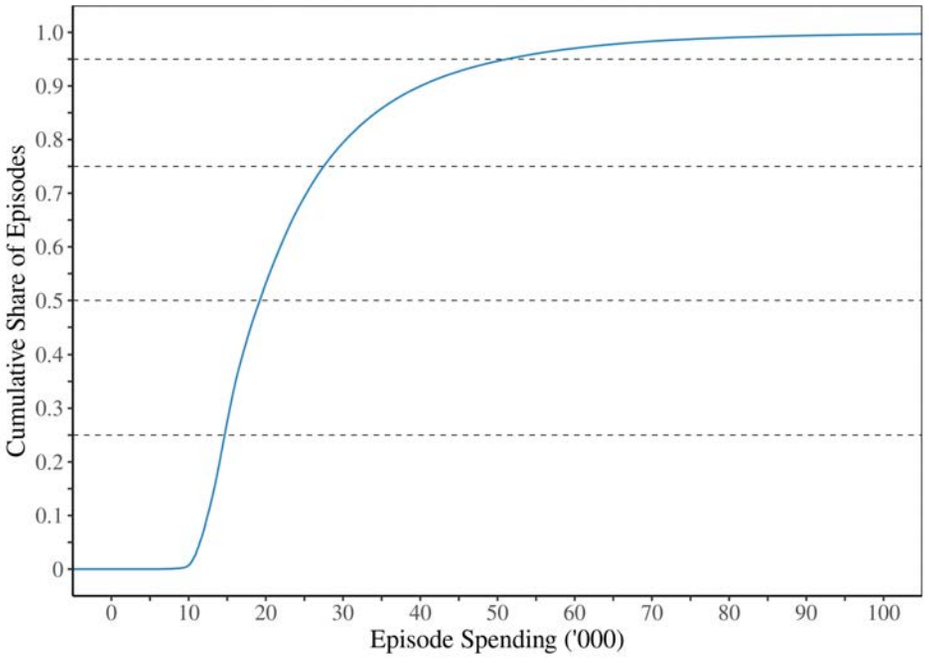


Figure A.3: Difference in differences and population treatment effects

Note: This figure presents different scenarios that could occur with regards to the presence of selection bias and/or sorting on treatment gains in a voluntary participation program, as described in Sections 3 and A.1. ΔY^0 and ΔY^1 denote the trend in the outcome, Y, in the absence and presence of treatment, respectively. DD denotes the difference-in-differences estimator, $DD = E(\Delta Y^1|D=1) - E(\Delta Y^0|D=0) = E(\Delta Y|D=1) - E(\Delta Y|D=0)$. ATE and ATT denote the average treatment effect in the population and on the treated, respectively. Since $ATT = E(\Delta Y^1 - \Delta Y^0|D=1)$ and $ATE = E(\Delta Y^1 - \Delta Y^0)$, these cannot be computed using observed data alone unlike the DD estimator. Panel (a) illustrates the case of experiments where both selection and sorting are eliminated due to explicit randomization into treatment. The remaining panels illustrate scenarios with one of the two effects: selection but no sorting (panels b and c) or sorting on gains without selection (panel d).

(a) Episode spend



(b) Index surgery length of stay

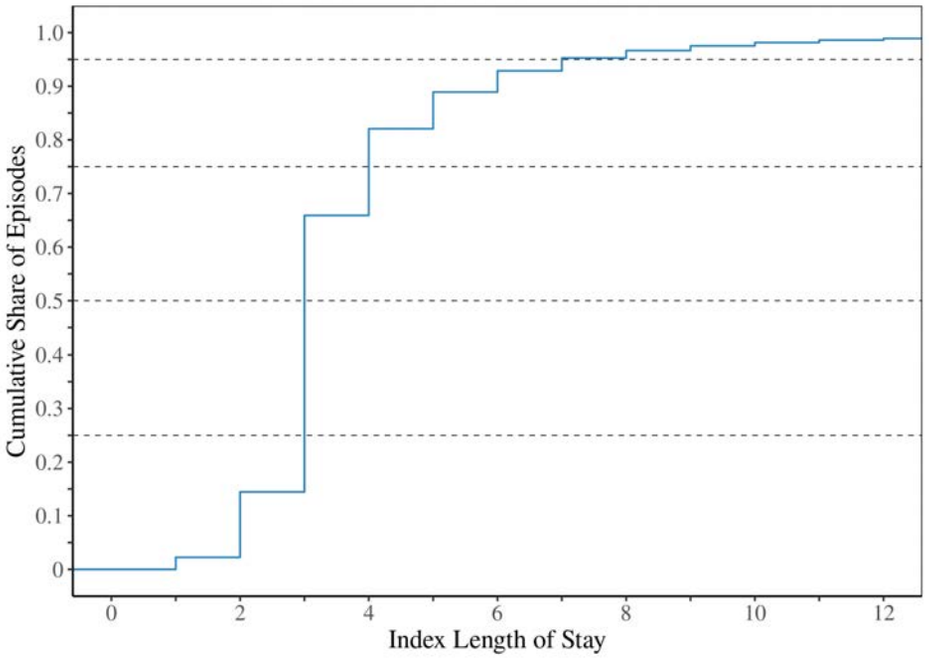


Figure A.4: Empirical CDFs in the pre-BPCI period

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

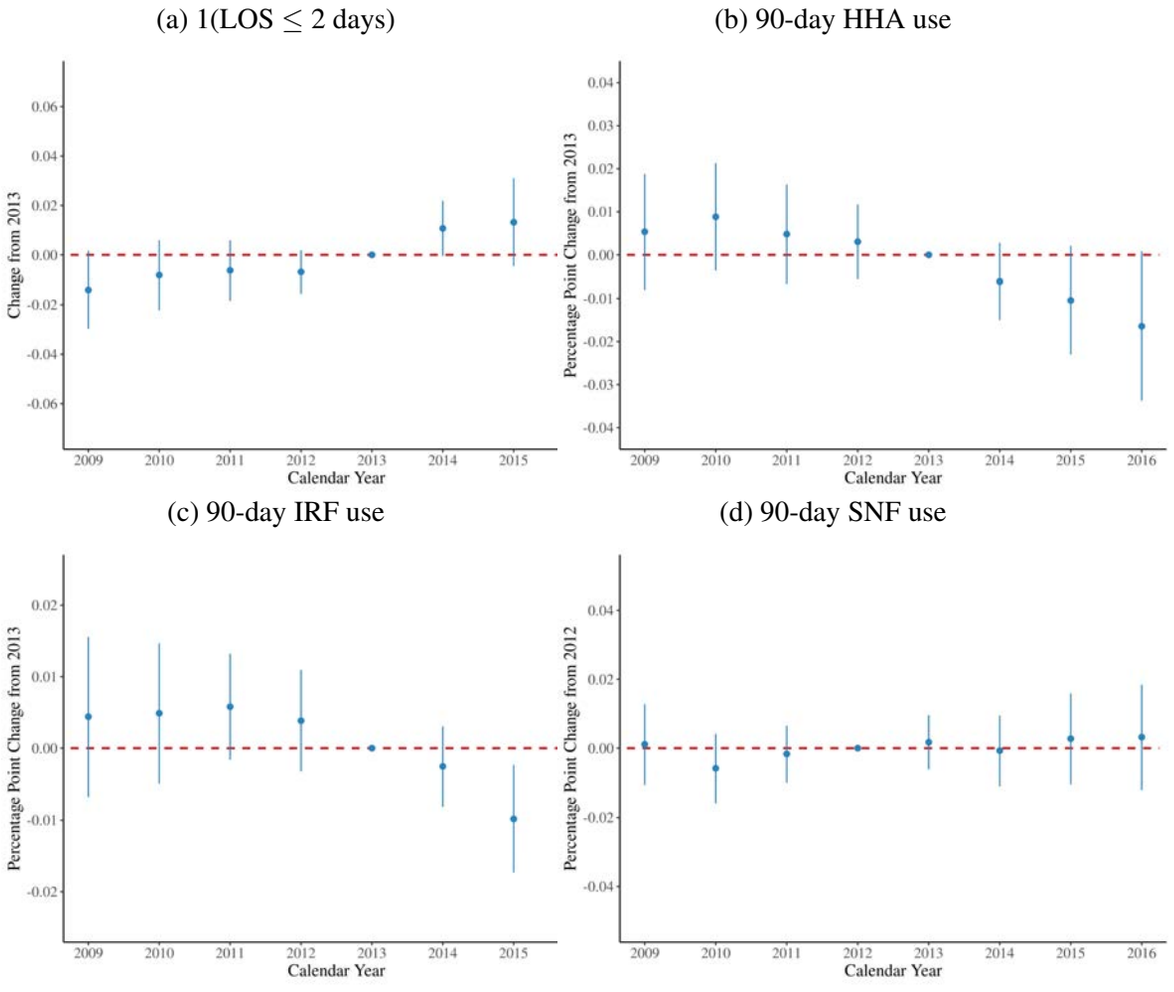


Figure A.5: Effects on utilization of care

Note: This figure presents results on the dynamic reduced form effects on index stay length and the use of post-acute care. 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 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.4b.

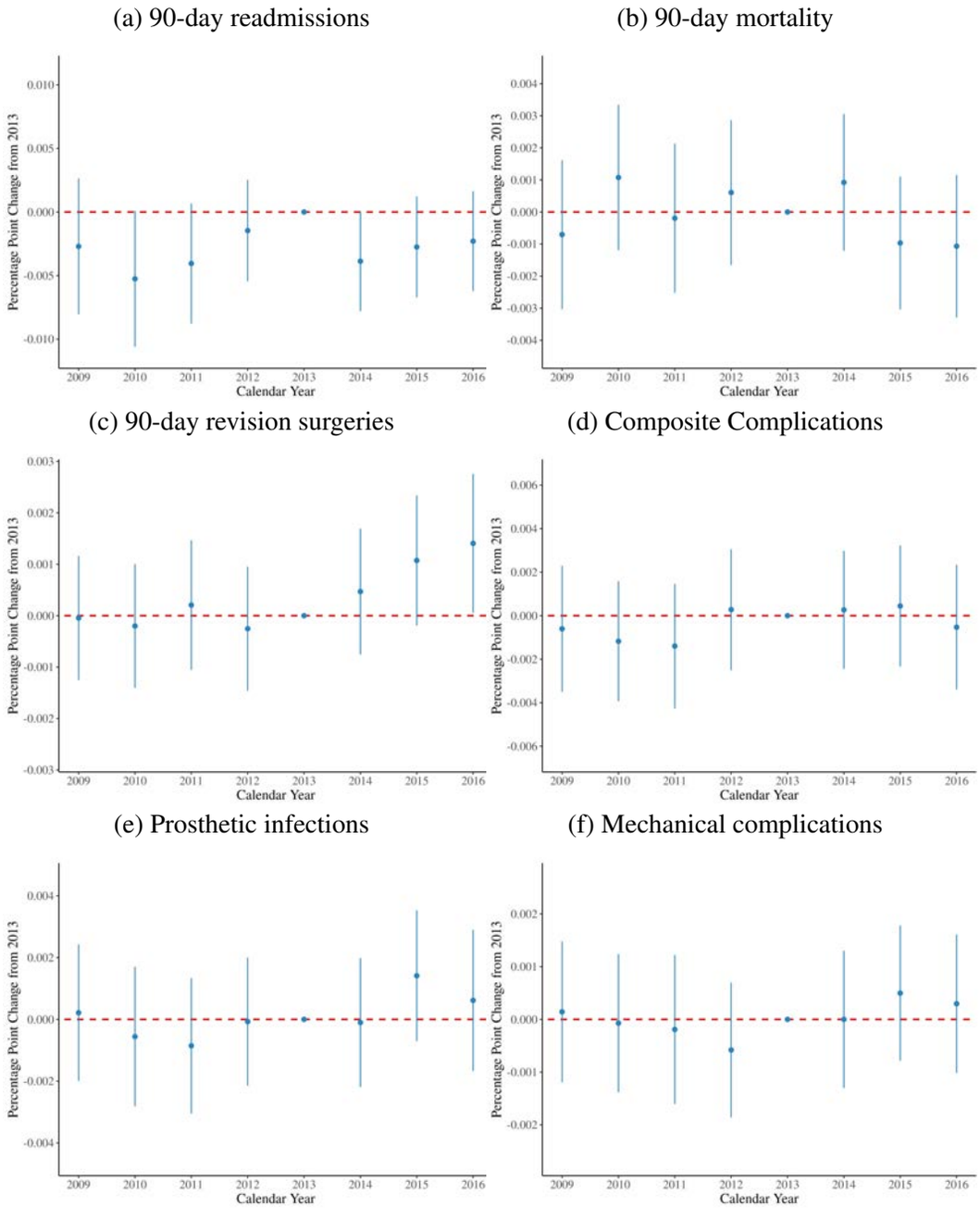


Figure A.6: Quality of care

Note: This figure presents the estimated reduced form dynamic effects on miscellaneous quality of care outcomes for LEJR patients. Readmissions, presented in panel (a), are defined to include revision surgeries. The models include hospital and market by year fixed effects and patient controls. Standard errors are clustered by hospital.

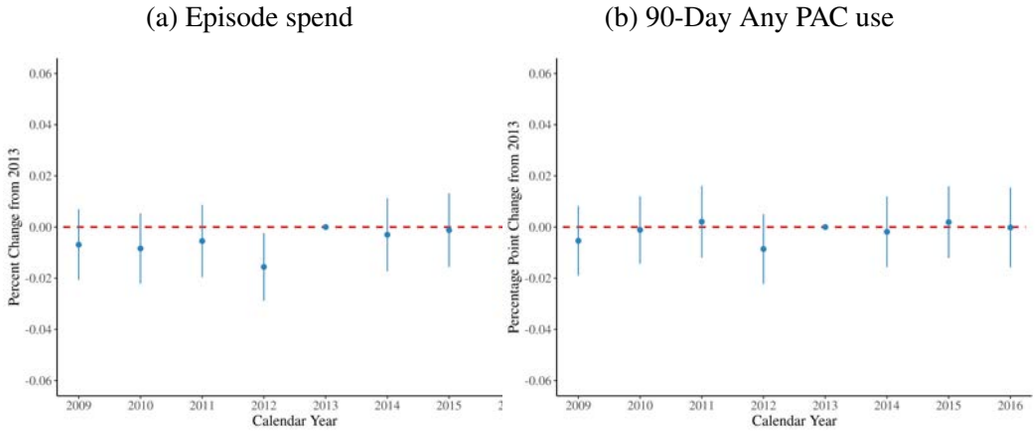


Figure A.7: Effect on cholecystectomy spending

Note: This figure presents the estimated dynamic reduced form effects on 90-day episode spend and PAC use for cholecystectomy episodes at hospitals participating in LEJR bundles. The models include patient controls, hospital, and HRR by year fixed effects. Standard errors are clustered by hospital.

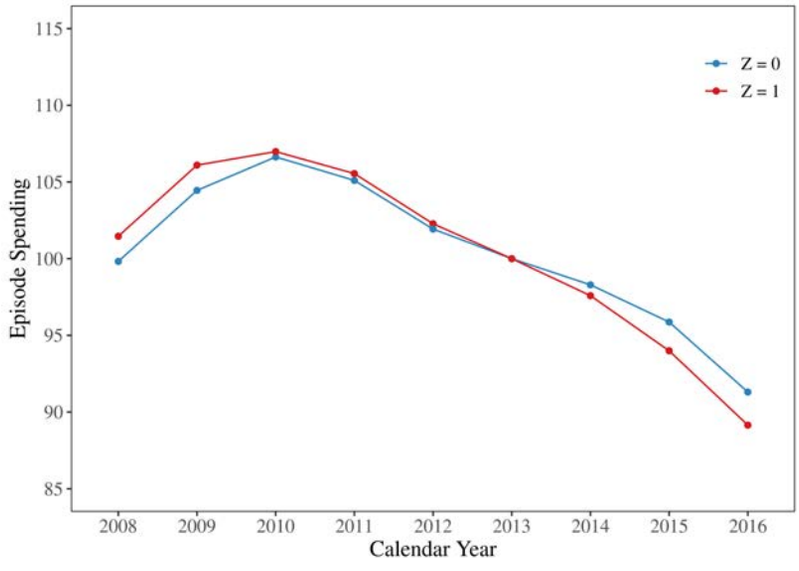


Figure A.8: Unadjusted trends in episode spend

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.

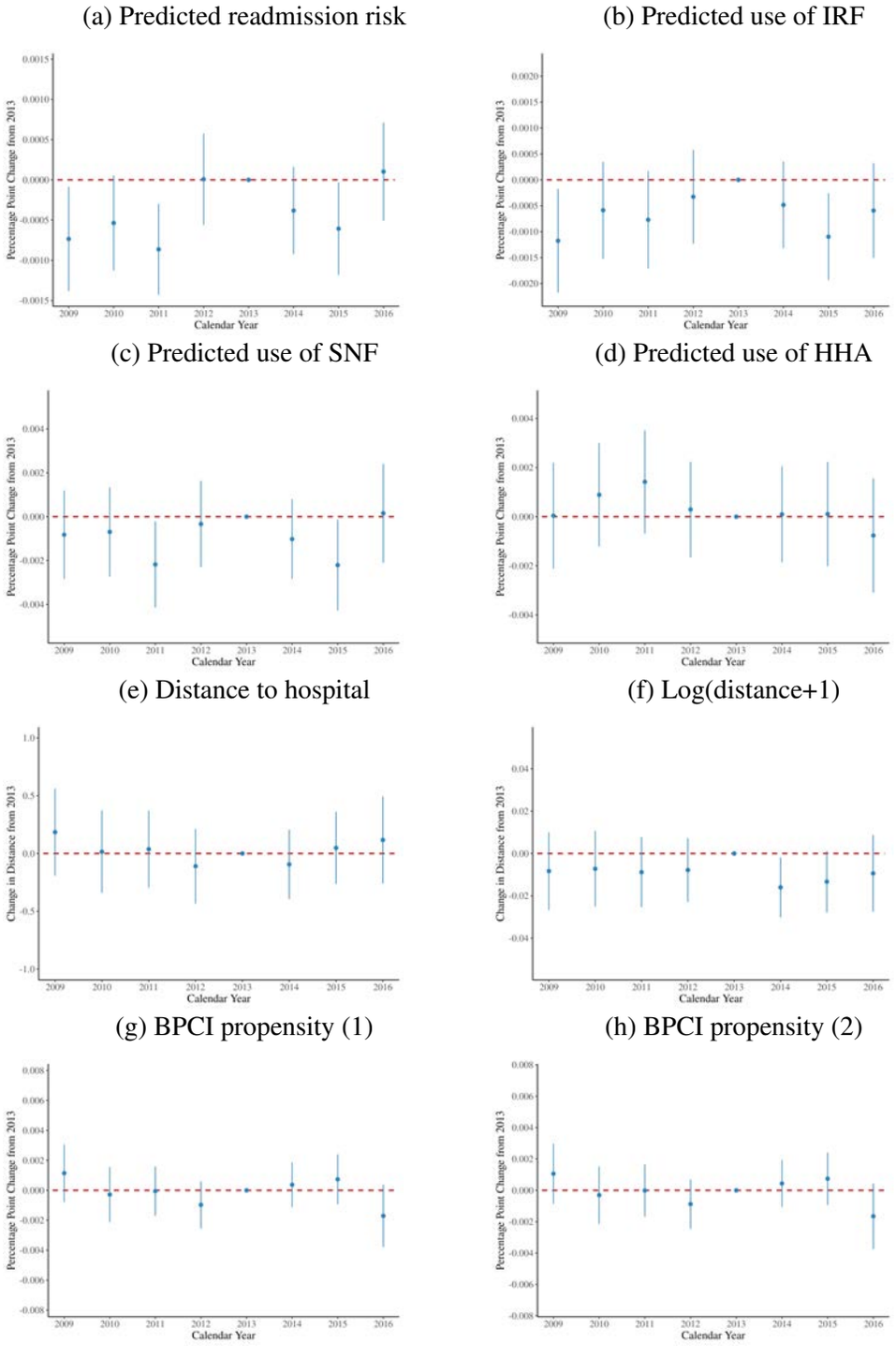


Figure A.9: Patient selection

Note: This figure presents the estimated reduced form dynamic effects on various measures of risk for LEJR patients, as described in Section 6.4. Panels (a)–(d) and (e)–(h) present results for selection on observed and unobserved risk, respectively. Predicted risk scores are generated using demographics and historical utilization and spending. Models for outcomes in panels (a)–(d) do not include patient controls, while those in panels (e)–(h) include patient controls. BPCI propensity is predicted using just 3-digit zipcode for panel (g), while in panel (h) we also use patient characteristics. Section A.4 describes the propensity measures in more detail.

Table A.1: Simulated financial gain and BPCI participation

	$\mathbf{1}(BPCI)$			
	(1)	(2)	(3)	(4)
A. Linear				
Expected Gain	0.052***	0.063***	0.013***	0.031***
	(0.005)	(0.004)	(0.004)	(0.003)
B. Quadratic				
Expected Gain	0.067***	0.090***	0.010	0.036***
	(0.010)	(0.003)	(0.006)	(0.003)
Expected Gain^2	−0.002	−0.002***	0.001	−0.0003**
	(0.002)	(0.0003)	(0.001)	(0.0001)
$p_{25}(Gains)$	0.358	0.373	0.054	0.015
$p_{75}(Gains)$	1.159	1.120	1.404	1.311
$Pr(BPCI)\|_{p_{25}(Gains)}$	0.097	0.087	0.121	0.102
$Pr(BPCI)\|_{p_{75}(Gains)}$	0.149	0.152	0.135	0.148
Semi-Elasticity	0.052	0.065	0.014	0.046
ρ	0.0	0.9	0.0	0.9
σ	0.01	0.01	0.04	0.04
μ	0.05	0.05	0.05	0.05

Note: This table presents point estimates from linear regressions of BPCI participation on expected gain (in $100,000 units using 2016 dollars) from the simulation exercise detailed in Figure A.1. Each column presents the results obtained using a different permutation of expected financial gain, determined by the parameter values ρ (correlation between baseline episode spend and saving) and σ (variance in spend reduction). $p_{25}(Gains)$ and $p_{75}(Gains)$ denote the 25th and 75th percentiles, respectively, of expected financial gain across hospitals. $Pr(BPCI)|_{p_{25}(Gains)}$ and $Pr(BPCI)|_{p_{75}(Gains)}$ denote the corresponding probabilities of BPCI participation for hospitals at the 25th and 75th percentiles, respectively, of financial gain. The semi-elasticity is the difference in probability of participation for hospitals at the 25th and 75th percentiles of expected financial gain.

Table A.2: 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 difference between average 2009–2013 and 2014–2016 log episode spending, matching the outcome used in Table 3 cols. 5 and 6. All regressions control for the patient characteristics included in the baseline specification. 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 99th percentile of each instrument, respectively. The instruments are described in detail in Section 6.1. Columns (4) and (5) use measures of expected financial gain described in Appendix Figure A.1, constructed using 2010–2012 data. 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 A.3: 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 5, 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. 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 A.4: Correlation of episode spending across clinical bundles

	LEJR	CHF	COPD	PNA	SEPSIS	UEJR	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	–	–
UEJR	0.616	0.561	0.554	0.533	0.537	1	–
UTI	0.731	0.844	0.836	0.875	0.838	0.506	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	–	–
UEJR	0.450	0.365	0.348	0.342	0.318	1	–
UTI	0.592	0.777	0.756	0.820	0.737	0.309	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.5: 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.6: Patient selection

Panel A	(1) Pred. Risk Readmission	(2) Pred. Use IRF	(3) Pred. Use SNF	(4) Pred. Use HHA
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. 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 A.4. Within each panel, the top and bottom rows present the DD and LATE coefficients, respectively. All regression models include patient 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 A.7: Complier versus participant hospitals

	Full Sample	Compliers		Treated	
Sub-Group	Proportion of Sample	Proportion of Sample	Relative to full	Proportion of Sample	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).

Table A.8: Sensitivity of MTE Results to using alternate instruments

	First stage f-stat	ATE	ATT	ATUT
	(1)	(2)	(3)	(4)
Baseline	25.3	-0.135***	-0.107***	-0.142***
		(0.032)	(0.027)	(0.036)
Observed Ortho. $>$ Med.	15.4	-0.152***	-0.121***	-0.16***
		(0.04)	(0.034)	(0.045)
Observed LEJR $>$ Med.	12.0	-0.207***	-0.187***	-0.213***
		(0.046)	(0.037)	(0.052)
Expected Gain $>$ Med.	32.7	-0.185***	-0.156***	-0.194***
		(0.037)	(0.026)	(0.045)
Expected Gain ($)	70.7	-0.181***	-0.136***	-0.195***
		(0.041)	(0.04)	(0.046)

Note: This table presents alternate estimates of population treatment effects, testing sensitivity to using instrument variants constructed under different assumptions. We consider the same alternate instruments used in Table A.2 and discussed in Section 6.1. The MTE estimations follow the same approach used for the baseline model described in Sections 7 and A.5. Column 1 presents the test statistic from the F-test of the first stage. The average treatment effect (ATE), average treatment effect on the treated (ATT), and average treatment effect on the untreated (ATUT) are presented in columns (2), (3), and (4), respectively. The first row presents results using the baseline instrument – an indicator for above median baseline *predicted orthopedic* post-discharge spending and revenue. The second row presents results using an above median indicator based on the corresponding *observed* values of orthopedic post-discharge spending and revenue. The third row presents results using an above median indicator based on the observed *LEJR* post-discharge spending and revenue. The fourth and fifth rows use measures of expected financial gain computed using the assumptions and approach described in Appendix Figure A.1. The fourth row presents results using an indicator for above median expected financial gain. The fifth row uses the expected financial gain value in dollars.

Table A.9: Selection and treatment effect heterogeneity

	Counterfactual change in spend (1)	Estimated treatment effect (2)
Unconditional Variance	30.04	12.72
Residual Variance with Additional Controls:		
Hospital Characteristics		
Hospital Bed Count	30.03 (100%)	12.7 (99.8%)
Ownership (For-Profit, Gov.)	29.97 (99.8%)	12.53 (98.4%)
Total Admissions	30.04 (100%)	12.63 (99.2%)
FTEs	30.04 (100%)	12.65 (99.4%)
FTEs per Bed	30.04 (100%)	12.7 (99.8%)
Medicare Share of Admissions	30.04 (100%)	12.69 (99.7%)
Medicaid Share of Admissions	30.04 (100%)	12.69 (99.7%)
Major Teaching	29.98 (99.8%)	12.63 (99.3%)
Any Teaching	29.97 (99.8%)	12.64 (99.3%)
Any Residents	29.86 (99.4%)	12.57 (98.8%)
Health System Size	30.02 (99.9%)	12.68 (99.6%)
Technologies	30.04 (100%)	12.59 (99%)
ACO Entry by 2016	29.98 (99.8%)	12.56 (98.7%)
Hosp. Spending > Nat. Median	29.9 (99.5%)	12.69 (99.7%)
Hosp. Spending > HRR Median	29.92 (99.6%)	12.72 (100%)
Pre-Period MA Trend	29.99 (99.8%)	12.62 (99.2%)
Pre-Period LEJR Vol. Trend	29.92 (99.6%)	12.72 (99.9%)
Pre-Period Ortho. Revenue Trend	29.94 (99.7%)	12.71 (99.9%)
Pre-Period Trend in Risk of SNF Use	26.94 (89.7%)	12.72 (100%)
Pre-Period Patient Age Trend	28.08 (93.5%)	12.72 (100%)
Pre-Period Trend in Readmission Risk	26.31 (87.6%)	12.72 (100%)
Pre-Period Trend in Mortality Risk	26.86 (89.4%)	12.72 (100%)
Market Characteristics		
Rural	30.04 (100%)	12.57 (98.8%)
HRR Spending > Median	30.04 (100%)	12.68 (99.6%)
All Characteristics		
Combination	25.46 (84.8%)	11.99 (94.2%)
HRR Fixed Effects	24.36 (81.1%)	9.83 (77.3%)
Combination with HRR FEs	20.78 (69.2%)	9.43 (74.1%)

Note: This table reports the variance in the estimated counterfactual change in episode spend (Col. 1) and the estimated treatment effect (Col. 2). These values are computed as described in Andresen (2018), building on the estimates from the MTE analysis. In each column, the top row presents the unconditional variance; subsequent rows present the residual variance and its ratio to the unconditional variance (expressed as a percentage in parentheses).