

Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Barnett ML, Wilcock A, McWilliams JM, et al. Two-year evaluation of mandatory bundled payments for joint replacement. *N Engl J Med* 2019;380:252-62. DOI: 10.1056/NEJMsa1809010

(PDF updated April 18, 2019)

Supplemental Appendix for “Impact of Mandatory Bundled Payments for Lower Extremity Joint Replacement on Spending and Outcomes in Medicare”

Investigators 2

Abbreviations 2

Appendix Methods..... 3

A. CJR Randomization Approach and Episode Weighting..... 3

B. Detailed Definitions of Outcomes..... 5

C. Risk Score Model Specification..... 8

D. Difference-in-Difference Model Specifications..... 9

E. Estimating Net Savings to Medicare..... 11

F. Testing Pre-Period Trends..... 11

Figure S1: Cohort Flow Diagram for Study Sample, 2015-2017 14

Table S1: Patient Characteristics of Continuously vs. Non-Continuously Enrolled Beneficiaries in Period Before LEJR Implementation, 2015..... 15

Table S2: Explanation of Measures of Institutional vs. Total Spending..... 16

Table S3: Differences in Pre-Period Trends in Outcomes in Intervention vs. Control MSAs, Primary Analysis and Alternate Model Specification 17

Table S4: Hospital Characteristics in Treatment and Control Groups in Period Before LEJR Implementation, 2015 20

Table S5: Comparison of Treatment and Control Metropolitan Statistical Areas (MSAs) 22

Table S6: Comparison of Differential Change in Institutional Spending Over Time After CJR Implementation 23

Table S7: Differential Change in LEJR Volume Before and After CJR Implementation 24

Table S8: Difference-in-Differences Sensitivity Analyses with Alternate Model Specifications for Primary and Secondary Outcomes 26

Table S9: Differential Change in Average Patient Risk Score Before and After CJR Implementation 28

Table S10: Differential Change in Patient and Procedure Characteristics Before and After CJR Implementation 29

Table S11: Results for Selected Outcomes Controlling for Separate Groups of Patient Characteristics..... 31

Investigators

Michael L. Barnett MD MS*, Andrew Wilcock PhD*, J. Michael McWilliams MD PhD, Arnold M. Epstein MD MA, Karen E. Joynt Maddox MD MPH, E. John Orav PhD, David C. Grabowski PhD, Ateev Mehrotra MD MPH

Department of Health Policy and Management, Harvard TH Chan School of Public Health, Boston MA (MLB, AE, EJO); Department of Health Care Policy, Harvard Medical School, Boston, MA (AW, JMM, DG, AM); Department of Medicine, Brigham and Women's Hospital, Boston, MA (MLB, JMM, AME, EJO); Department of Internal Medicine, Washington University School of Medicine in St. Louis, St. Louis, MO (KJM)

*These authors contributed equally to this manuscript.

Abbreviations

BPCI	Bundled Payments for Care Improvement
CJR	Comprehensive Care for Joint Replacement Model
CMS	Center for Medicare and Medicaid Services
DRG	Diagnosis related group
ED	Emergency department
HHA	Home health agency
ICD-9 or ICD-10	International Statistical Classification of Diseases and Related Health Problems, revision 9 or 10
IRF	Inpatient rehabilitation facility
LEJR	Lower extremity joint replacement
MSA	Metropolitan statistical area
PAC	Post-acute care
SNF	Skilled nursing facility

Appendix Methods

A. CJR Randomization Approach and Episode Weighting

In the Comprehensive Care for Joint Replacement (CJR) model, the unit of randomization is the MSA, with 75 “treatment” MSAs (8 of which “dropped out” from the intervention prior to implementation due to rule changes, see below) exposed to bundled payments beginning in April 1, 2016, and 121 “control” MSAs with no payment changes.

The Center for Medicare and Medicaid Services (CMS) identified 196 MSAs eligible for randomization based on the following inclusion criteria:

1. >400 lower extremity joint replacement (LEJR) episodes from 7/1/2013 to 6/30/2014
2. > 400 non-BPCI LEJR episodes from 7/1/2013 to 6/30/2014
3. <50% of LEJR episodes included in any BPCI model
4. >50% of LEJR episodes paid through the inpatient prospective payment system (e.g. excluding Maryland, critical access hospitals, etc.)

To select a treatment group from these 196 eligible MSAs, CMS divided the MSAs into 8 strata defined by quartiles of pre-period episode spending (defined from 7/1/2013 to 6/30/2014) and above vs. below median population (**Appendix Methods Table 1**). The probability of an MSA being selected for the treatment group depended on pre-period average LEJR 90-day episode spending, varying from 30% to 45% as shown in **Appendix Methods Table 1**.

Appendix Methods Table 1:

MSA Selection by Payment Quartile and Above/Below Median Population Criteria, as Defined by CMS

	Payment in lowest quarter	Payment in 2nd lowest quarter	Payment in 3rd lowest quarter	Payment in highest quarter	Total eligible MSAs
Selection Proportion	30%	35%	40%	45%
Less Than Median Population (Group #)	(1)	(2)	(3)	(4)
Number Eligible MSAs per Proposed Rule (80 FR 41198)	33	19	22	24	98
Proportion x Number	9.9	6.65	8.8	10.8	
Number initially selected from group	10	7	9	11	37
Number finally selected from group	8	6	8	11	33
More Than Median Population (Group #)	(5)	(6)	(7)	(8)	
Number Eligible MSAs per NPRM	16	30	27	25	98
Proportion x Number	4.8	10.5	10.8	11.25	
Number initially selected from group	5	11	11	11	38
Number finally selected from group	5	10	9	10	34
Total Eligible MSAs per Proposed Rule (80 FR 41198)	49	49	49	49	196
Number initially selected	15	18	20	22	75
Number finally selected from group	13	16	17	21	67

Source: CJR Final Rule, Federal Register 80 FR 73298 (<https://www.federalregister.gov/d/2015-29438/page-73298>)

During this initial process of randomization 75 MSAs were selected to participate in CJR announced in July 2015. However, in the last Final Rule announced in November 2015, the treatment group was reduced from 75 MSAs to 67 MSAs. The 8 MSAs were dropped because there was an increase in BPCI participation and therefore were not eligible via either inclusion criteria #2 or #3 listed above.

One approach to manage this selective reduction of treatment MSAs is to drop the same MSAs from the control group. However, CMS was unable to share with us the control MSAs which would be dropped using the same rules. Because this change disrupts initial randomization, we defined the treatment group as the initially randomized 75 MSAs with an “intent-to-treat” approach. The 8 dropped MSAs represent 7% of the total LEJR episodes in the treatment group in the post-intervention period used in the main manuscript.

Accounting for Variable Treatment Assignment Probability

As described above, MSAs had different probabilities of being assigned to the treatment group depending on their quartile of average LEJR spending from July 1, 2013 to June 30, 2014. This will create baseline imbalance between treatment and control groups in unadjusted comparisons. Also, to the degree that MSAs were high spending due to random chance, regression to the mean would bias our savings estimates. To account for the varying probabilities of treatment assignment, we produced standardizing weights to weight episodes, hospitals and MSAs such that the probability of the treatment or control MSAs being selected are equal within each stratum (i.e. analogous to “direct standardization”). The weights were chosen to match each stratum to the treatment/control probability of the entire sample.

The assignment probabilities and weights used are summarized in **Appendix Methods Table 2**. Among the 196 MSAs eligible for CJR, 38.3% (75 MSAs) were initially randomized to be in the treatment group and 61.7% (121 MSAs) were in the control group. This probability varied by MSA randomization stratum: for example, 30.3% of MSAs in Stratum 1 below were in the treatment group, whereas 44.0% were in the treatment group for Stratum 8. The treatment and control weights were derived to equalize the probability for each stratum to match the whole sample, with 38.3% of MSAs in the treatment group and 61.7% in the control group.

Appendix Methods Table 2: Derivation of Standardizing Weights for Treatment and Control Episodes

Stratum	Payment Quartile	Above/Below Median Population	Initial Treatment MSAs	Control MSAs	Treatment Probability	Control Probability	Treatment Weight (Overall Prob/Stratum Prob)	Control Weight (Overall Prob/Stratum Prob)
Overall	-	-	75	121	0.383	0.617	-	-
1	1 (low)	Below	10	23	0.303	0.697	1.263	0.886
2	1 (low)	Above	5	11	0.313	0.688	1.224	0.898
3	2	Below	7	12	0.368	0.632	1.039	0.977
4	2	Above	11	19	0.367	0.633	1.044	0.975
5	3	Below	9	13	0.409	0.591	0.935	1.045
6	3	Above	11	16	0.407	0.593	0.939	1.042
7	4 (high)	Below	11	13	0.458	0.542	0.835	1.140
8	4 (high)	Above	11	14	0.440	0.560	0.870	1.102

B. Detailed Definitions of Outcomes

The descriptions below are taken from the pre-specified updated protocol posted in the Open Science Framework (https://osf.io/2pdtf/?view_only=2d45938c018b4f9c9975f68598c26e96).

Primary Outcome #1: Facility/Institutional Spending per Episode

- We will use the full unadjusted payments (spending) to institutional providers (including Medicare, patient and primary payer portions) that include common CMS adjustments such as wage index
 - Rationale: we are interested in the pragmatic national impact of the program on spending, regardless of regional and service-specific adjustments that Medicare applies
 - As secondary outcomes, we will also examine total costs
 - Any denied payments will be excluded from spending outcomes
- The payments for any services that cross over the end of the LEJR episode (e.g. SNF admission or HHA episode) will be pro-rated based on the percentage of days occurring within the episode
 - For example, if a HHA service begins 86 days post discharge (i.e. 5 days left in the episode) and lasts 20 days, 5/20 or 25% of the HHA payment will be attributed to the LEJR episode.
- In the episode reconciliation process, CMS removes LEJR “unrelated” costs from calculated episode costs (see CMS website for [more documentation](#), <https://innovation.cms.gov/Files/worksheets/ccjr-exclusions.xlsx>; Accessed June 20, 2018). We are not applying these exclusions to capture a broader picture of total institutional spending per episode.

Primary Outcome #2: TKA/THR complications rate

- Definition: we will capture all complication events as defined in NQF Measure #1550
- We used Measure #1550 because it is endorsed by the NQF, employed by CMS for quality reporting, and it is used by CJR to set quality thresholds for receiving bonuses
- While we will capture complications as described by this measure, we will define the sample differently. We will include all episodes for the primary analysis as defined above in Section C.2.
 - We chose a different sample because the measure defines its sample based on the presence of ICD-9 procedure codes 81.51 (total hip replacement) and 81.54 (total knee replacement), not DRGs 469 and 470 as above.
 - The NQF measure also makes a number of exclusions to focus on elective joint replacements. The most relevant is exclusion is hip fracture patients. In our primary outcomes we did not make this exclusion.

- As a secondary outcome, we will additionally examine complication rates using the exact sample definition in the NQF defined measure
- This measure, as defined by CMS and used in CJR, captures a composite of the following outcomes:
 - Within 7 days of admission
 - Myocardial infarction
 - Pneumonia
 - Sepsis/shock
 - Within 30 days of admission
 - Surgical site bleeding
 - Pulmonary embolism
 - All-cause mortality
 - Within 90 days of admission
 - Implant mechanical complications
 - Joint/wound infection

Primary Outcome #3: Top quartile of patient risk score, estimated as predicted LEJR episode spending

- The purpose of this outcome is to capture patient selection by hospitals subsequent to CJR implementation in a single, integrated measure. The concern is that to reduce spending per episode, hospitals will selectively perform on LEJR on patients who are healthier and therefore less costly.
- Will examine the proportion of LEJR patients in the highest quartile of patient risk as defined below in both the treatment vs. control MSAs
- Using data from before CJR implementation (2013-2014) for a 20% sample of beneficiaries, we will estimate a linear regression model with the dependent variable of predicted total episode spending
 - We will use the 20% sample to be able to include estimated Part B (“carrier”) costs, which are only available for a 20% sample.
 - Estimates from this model will be used to generate predicted spending for each episode in our study sample
- Patient covariates in the model are listed below in part C below

Secondary Outcomes

Spending and Utilization

- 1) Total LEJR episode spending. Our data sample has only “Carrier” (Part B) claims for a 20% sample. We will include facility/institutional and Part B claims in this measure.
- 2) 90-day post-discharge spending in mutually exclusive categories (100% sample unless otherwise specified)
 - Index hospitalization
 - Repeat inpatient, ED and observation spending in 90-days
 - Post-acute care spending: SNF, IRF, and HHA
 - Other Medicare “Carrier” Part B spending (20% sample)
- 3) Number of LEJR episodes per 1,000 beneficiaries

- LEJR episodes per 1,000 beneficiaries at the MSA level
 - Denominator defined as all beneficiaries in an MSA-quarter with continuous FFS Medicare coverage for that quarter
 - Also examined LEJR episodes at the individual beneficiary level (See **Table S4**)
- 4) 90-day post hospital discharge all-cause hospital readmission rates
 - Binary outcome, 1 = any inpatient admission in the 90-day window
 - Excluding certain planned admissions as defined in the CMS hospital-wide readmission measure (NQF #1789)
 - 5) 90-day post hospital discharge all-cause observation visit rates without admission
 - Binary outcome, 1 = any observation stay in the 90-day window without admission
 - 6) 90-day post hospital discharge all-cause emergency department rates without admission
 - Binary outcome, 1 = any emergency department visit in the 90-day window without admission or observation stay
 - 7) 90-day post hospital discharge all cause hospital visit of any kind (emergency department visit, observation stay or inpatient admission)
 - Binary outcome, 1 = any hospital visit in the 90-day window
 - 8) 90-day all-cause mortality
 - 9) LEJR complications rate as defined in primary outcome #2 with CMS-defined exclusion criteria (as opposed to using full study sample defined above)
 - 10) Proportion of LEJR admissions discharged to home vs. different PAC settings:
 - A. Proportion discharged to any institutional PAC discharge (SNF, IRF, LTAC)
 - B. Proportion discharged home
 - C. Proportion discharged to home health agencies (HHA)
 - 11) Average length of stay among those admitted to institutional PAC settings during full episode among those with at least 1 day of institution PAC use

Patient Selection

Examine the proportion of patients in the treatment and control groups in the pre vs. post-periods with the following characteristics:

- 1) Sex
- 2) Age: 65-79, 80-89, 90+
- 3) Race
- 4) Urban/rural residence
- 5) Medicaid eligibility
- 6) Disability as original reason for enrollment
- 7) Prior inpatient admission in prior 12 months
- 8) Any institutional PAC (SNF, IRF, LTAC) admission in prior 12 months
- 9) *Post-hoc additions:*
 - a. Diagnosis related group 469 or 470
 - b. Presence of hip fracture
 - c. Type of procedure (defined by ICD-9 procedure code)

Post Hoc Outcomes

These outcomes were developed in response to our findings after initial analysis or reviewer comments and therefore were not pre-specified.

- 1) Average number of episodes among those admitted to home health agencies (HHA) during full episode among those with at least 1 day of HHA use
 - Examined this to understand why the increase in HHA use did not result in a change in total HHA spending
- 2) Average length of stay among those admitted to institutional PAC settings during full episode among those with at least 1 day of institution PAC use, excluding IRFs.
 - Examined both length of stay in any institution PAC setting and non-IRF settings because IRFs are not paid based on length of stay (except in outlier cases)

C. Risk Score Model Specification

To define the patient risk score, we estimated predicted LEJR episode spending using coefficients from a linear regression model fitted on a 20% sample of 2013-2014 Medicare administrative claims data, before the study period. Episodes were identified with same inclusion and exclusion criteria as study sample, defined in the Methods section of the main manuscript.

Model specification was the following:

$$TotalSpending_i = \beta_0 + X_i\tau + \varepsilon_i$$

- $TotalSpending_i$ is patient i 's total episode spending
- β_0 is a constant
- ε_i is random error
- X_i are beneficiary demographics, including:
 - Age indicators (<65, 65 to 69, 70 to 74, 75 to 79, 80 to 84, 85 to 89, and 90+)
 - Male indicator
 - Race indicators (white, black, Asian, other, Hispanic)
 - Medicaid eligibility status (dual eligible or not)
 - Original entitlement reason indicators (age 65+, disability or end-stage renal disease)
 - Metro residence indicator set equal to 1 if the bene's Zip code is located within Rural-Urban Commuting Area (RUCAs) 1-3 (i.e., metropolitan area), and 0 otherwise
 - For 0.16% of beneficiaries with missing RUCA, we used a missing RUCA indicator set equal to one if the bene's Zip code did not have RUCA, and 0 otherwise

- Prior inpatient use indicator equal to 1 if the patient had any admission to a short term or critical access hospital in the 12 months prior to their LEJR admission, and 0 otherwise
- Prior PAC use indicator was set to 1 if the patient had any admission to an institutional PAC setting including SNF, IRF and long term care hospital in the 12 months prior to their LEJR admission, and 0 otherwise
- 27 chronic condition indicators each set equal to 1 if the year of the (earliest) diagnosis detected by date was at least one year prior to the year of the LEJR index stay, and 0 otherwise
 - Conditions included: Alzheimer’s disease, Alzheimer’s disease and related disorders or senile dementia, anemia, asthma, atrial fibrillation, benign prostatic hyperplasia, breast cancer, cataract, chronic kidney disease, chronic obstructive pulmonary disease, colorectal cancer, depression, diabetes, endometrial cancer, glaucoma, heart failure, hip or pelvic fracture, hyperlipidemia, hypertension, hypothyroidism, ischemic heart disease, lung cancer, osteoporosis, prostate cancer, acute myocardial infarction, rheumatoid arthritis, and stroke or transient ischemic attack.

D. Difference-in-Difference Model Specifications

Primary and Secondary Outcomes

All difference-in-differences models for the primary and secondary outcomes (including our binary outcomes) used linear regression with the MSA sampling weights described above in section A (except for procedure volume which used logistic regression, see Appendix Table 3). We used linear probability models to be able to incorporate hospital random effects for 1756 hospitals and MSA random effects.

In our pre-specified protocol we used proposed using hospital fixed effects and clustered standard errors at the MSA level. Based on input from statistical reviewers, we changed the model specification to hospital and MSA random effects. In a post-hoc change before we submitted the manuscript, we did not limit eligible LEJR episodes to one per beneficiary (i.e. multiple LEJR episodes per beneficiary in the study period were allowed). We also specified in this version that we would not include DRG as a covariate in the models. Details on the rationale for the timeline and the rationale for these changes are provided in the analytic protocol available in the supplemental material. As a sensitivity analysis, we provide below the results of a model using hospital fixed effects, using multiple LEJR episodes per beneficiary, and using DRG as a covariate.

Model specification was the following

$$Outcome_{ihq} = \beta_0 + \delta_h + \mu_h + \alpha_q + Treatment_h + \beta_1 Post_q * Treatment_h + X_i \tau + \varepsilon_{ihq}$$

- $Outcome_{ihq}$ is patient i 's outcome value in hospital h in quarter q

- For spending, length of stay and patient risk outcomes this is a continuous value corresponding with the relevant outcome
- For utilization and quality outcomes such as readmission or institutional PAC use, this corresponds to a binary outcome
- For the proportion of top quartile risk patients, this corresponds to a binary outcome
- β_0 is a constant
- δ_h is a hospital random effect
- μ_h is an MSA random effect
- α_q are quarter effects (9 quarter indicators, as Q1 and Q2 of 2016 were excluded)
- $Post_q * Treatment_h$ is equal to 1 for hospitals located in a treatment MSA during the post period (Q3 of 2016 through Q3 of 2017) and 0 otherwise
 - There is no main effects for $Post$ included because this term is incorporated into quarter fixed effects
- ε_{ihq} is the error
- X_i are beneficiary demographics:
 - **Note:** models using patient risk (manuscript Table 1, Appendix Table 5) or patient characteristics (manuscript Table 4) as outcomes **do not** include patient characteristics since these were used to generate the patient risk score.
 - Index episode indicators:
 - Indicator for fracture as defined in NQF Measure #1550
 - Indicators for procedures: total knee, total hip, partial, or none as defined in NQF Measure #1550
 - <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html>
 - Age indicators (<65, 65 to 69, 70 to 74, 75 to 79, 80 to 84, 85 to 89, and 90+)
 - Male indicator
 - Race indicators (white, black, Asian, other, Hispanic)
 - Medicaid eligibility status (dual eligible or not)
 - Original entitlement reason indicators (age 65+, disability or end-stage renal disease)
 - Metro residence indicator set equal to 1 if the bene's Zip code is located within Rural-Urban Commuting Area (RUCAs) 1-3 (i.e., metropolitan area), and 0 otherwise
 - For 0.16% of beneficiaries with missing RUCA, we used a missing RUCA indicator set equal to one if the bene's Zip code did not have RUCA, and 0 otherwise
 - 27 chronic condition indicators each set equal to 1 if the year of the (earliest) diagnosis detected by date was at least one year prior to the year of the LEJR index stay, and 0 otherwise
 - Conditions included: Alzheimer's disease, Alzheimer's disease and related disorders or senile dementia, anemia, asthma, atrial fibrillation, benign prostatic hyperplasia, breast cancer, cataract, chronic kidney disease, chronic obstructive pulmonary disease, colorectal cancer, depression, diabetes, endometrial cancer, glaucoma, heart failure, hip or pelvic fracture, hyperlipidemia, hypertension,

hypothyroidism, ischemic heart disease, lung cancer, osteoporosis, prostate cancer, acute myocardial infarction, rheumatoid arthritis, and stroke or transient ischemic attack.

E. Estimating Net Savings to Medicare

To estimate the net savings attributable to CJR for Medicare, we incorporated publicly available data released by Medicare in late 2018 on reconciliation payments made in the first 2 performance periods of CJR (2016 and 2017). These data are available at <https://innovation.cms.gov/initiatives/cjr> (accessed December 13, 2018).

The publicly released data shows that for CJR performance year 1 (2016), there were net payments of \$35,541,774 for 47,426 episodes, while for performance year 2 (2017) there were \$93,393,307 net payments for 100,497 episodes. This equates to an average reconciliation payment of \$872 per episode across the first two performance years. We subtracted this average per LEJR episode payment from the \$1,084 total differential savings we estimated per LEJR episode, yielding \$212 net savings per episode. Using the baseline total pre-period spending per LEJR episode of \$30,504 for the treatment group (Table 3 in the main manuscript), this corresponds to an average net 0.7% relative decrease in spending per episode.

We recognize that the sample of episodes we used in our analyses and the episodes used when reporting reconciliation payments are slightly different given we used additional exclusion criteria. Further, the performance year 2 numbers available at the time of publication were preliminary and therefore the \$213 estimate may change once the final reconciliation payments for year 2 are released.

One question is whether the net savings to Medicare increased in Year 2. While we found that the differential reduction in spending increased over time, the reconciliation payments also increased. The reconciliation payments will also change in Year 3 of the program because the benchmark will change from a target price based primarily on hospital-specific historical spending (2/3 weight in years 1 and 2, with 1/3 weight on regional historical spending) to a target price based primarily on historical spending in the entire region (2/3 regional spending, 1/3 hospital spending). In years 4 and 5 of the program, the target price will be based entirely on regional spending alone.

F. Testing Pre-Period Trends

To estimate differences in pre-period trends in our primary and secondary outcomes between treatment and control MSAs, we used data from the pre-treatment period of 2015.

In a post-hoc sensitivity analysis, we also tested for pre-period differences in a longer time period of 2011-2015 but only using a 20% random sample of enrollees (the data we had available for this time period). As noted above (**Appendix Methods D**), we conducted a sensitivity analysis using different model specifications. We present all of these pre-trend differences in **Table S3** below.

We used the following model specification for each outcome:

$$Outcome_{ihq} = \beta_0 + \delta_h + \mu_h + \alpha_y + Time_y + Treatment_h + \beta_1 Time_y * Treatment_h + X_i \tau + \varepsilon_{ihy}$$

- $Outcome_{ihq}$ is patient i 's (with LEJR in hospital h in year [or quarter] y) outcome value
 - For spending, length of stay and patient risk outcomes this is a continuous value corresponding with the relevant outcome
 - For utilization and quality outcomes such as readmission or institutional PAC use, this corresponds to a binary outcome
 - For the proportion of top quartile risk patients, this corresponds to a binary outcome
- β_0 is a constant
- δ_h is a hospital random effect
- μ_h is an MSA random effect
- α_y are quarter or year effects (either 4 quarter effects for each quarter in 2015 or 5 year effects from 2011-2015)
- $Time_y * Treatment_h$ is an interaction term between a linear term for quarter/year and a binary indicator for whether a hospital was in a treatment vs. control MSA
- ε_{ihy} is the error
- X_i are beneficiary demographics as defined in section D above

Pre-Period Trend Results

In 2015, the pre-period used for the difference-in-difference models in the manuscript, trends in outcomes were not significantly different between treatment and control MSAs using either model specification (**Table S3**).

From 2011-2015, pre-period trends were not significantly different for most outcomes, however, there were differential changes in institutional and total spending in treatment vs. control MSAs. For example, there was a differential increase of roughly \$90 in the treatment MSAs in institutional spending under both model specifications (**Table S3**). These differences were not statistically significant using the fixed effects alternate model specification but were statistically significant in the random effects model.

We have chosen not to emphasize these differences in the manuscript for several reasons. First, as we mention above, it is not the pre-period directly used in our statistical models, which is 2015 alone. Second, it is the result of a post-hoc analysis and we have not corrected for multiple testing. Also, the difference is relatively small in magnitude (institutional spending is roughly \$25 thousand per episode), and a statistically significant result was not seen using the time period pre-specified in our protocol or different model specifications. Further, if we assume that this pre-period difference extrapolates into the intervention period, it would imply that the estimate of savings in our manuscript could be underestimated by ~\$130. Therefore, this change does not substantively alter our interpretation of our findings that the CJR program led to modest reduction in spending.

Figure S1: Cohort Flow Diagram for Study Sample, 2015-2017

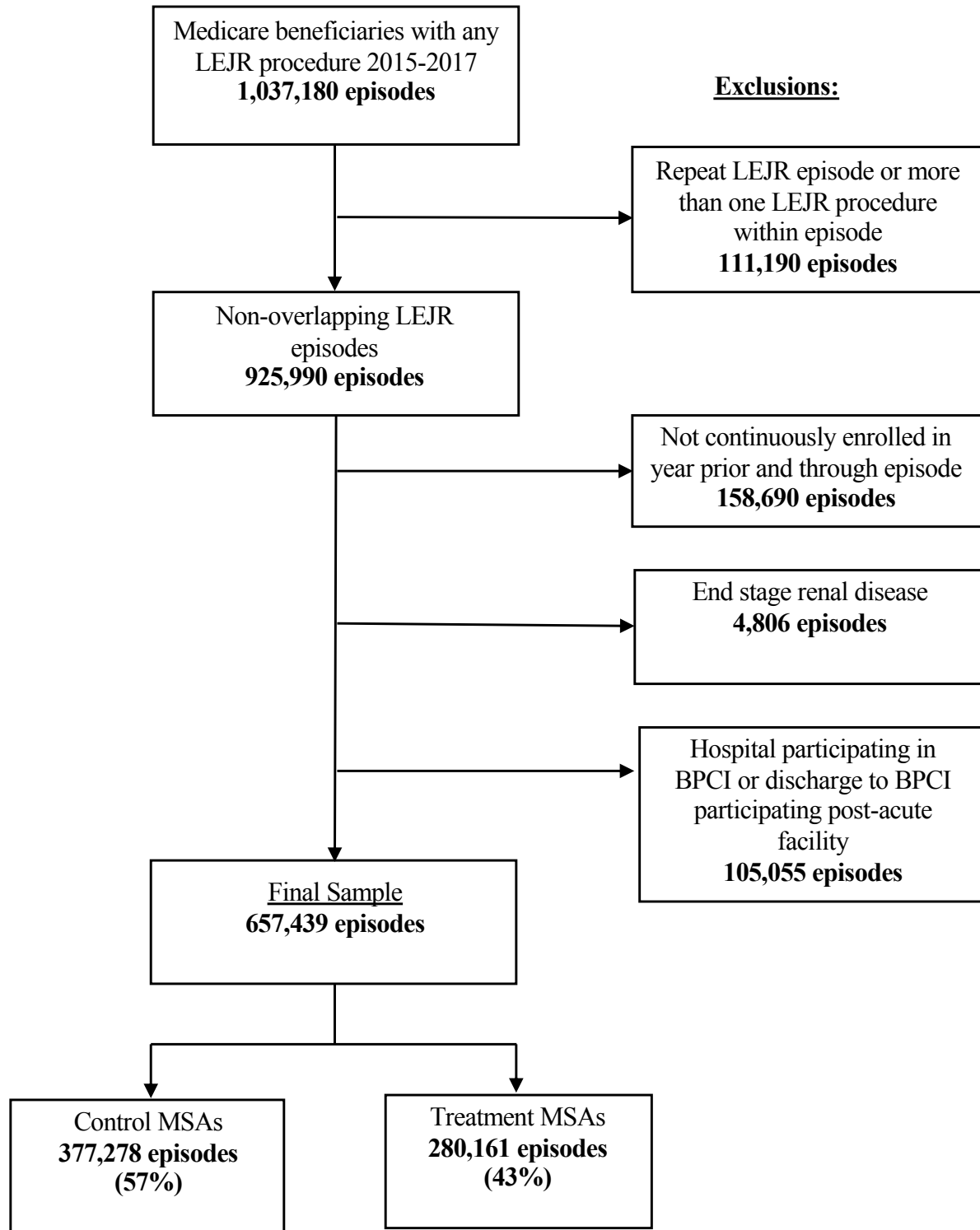


Table S1: Patient Characteristics of Continuously vs. Non-Continuously Enrolled Beneficiaries in Period Before LEJR Implementation, 2015

Patient and Episode Characteristics	Continuously Enrolled Beneficiaries (Primary Analysis)	Non-Continuously Enrolled Beneficiaries (Excluded)
Patients (weighted N)	233,490	139,288
Age (mean)	74.4	68.6
Male (%)	36.5%	40.6%
Race/Ethnicity (%)		
Non-Hispanic White	90.3%	85.5%
Non-Hispanic Black	5.4%	7.4%
Asian	0.8%	1.2%
Other Race/Ethnicity	2.5%	4.5%
Hispanic	0.9%	1.5%
Original Reason for Entitlement (%)		
Age >65	84.5%	83.7%
Disability	15.5%	16.2%
ESRD ¹	0.1%	0.1%
Medicaid Eligible	10.3%	10.7%
Urban Residence ²	83.4%	85.0%
Number of chronic conditions (mean) ³	7.0	2.1

Abbreviations: diagnosis related group (DRG), lower extremity joint replacement (LEJR), disproportionate share hospital population (DSH). Hospital characteristics were measured using the Medicare Provider of Services and Impact files.

¹Patients with ESRD were excluded from the payment program. However, some patients initially qualified for Medicare due to ESRD but no longer classified as ESRD at the time of LEJR.

²Urban location defined using the Health Resources and Service Administration (HRSA) Rural-Urban Commuting Area (RUCA) code database (<http://depts.washington.edu/uwruca/index.php>). Urban was defined as a patient residing in a “metropolitan” ZIP code. Data was missing for 0.16% of episodes which were largely located in Puerto Rico.

³We assessed the presence of 27 conditions from the Chronic Condition Data Warehouse (CCW), which uses claims since 1999 to describe Medicare beneficiaries’ accumulated chronic disease burden (see **Appendix Methods C** for list of conditions).

Table S2: Explanation of Measures of Institutional vs. Total Spending

Primary Outcome: Institutional spending	Secondary outcomes: Total spending
<p>Payments to...</p> <ul style="list-style-type: none"> • hospitals for inpatient admission or readmission • home health, skilled nursing facilities, hospice and inpatient rehabilitation • hospitals for outpatient hospital services (office visits, radiology, laboratory testing) • vendors for durable medical equipment <p>Includes all Part A Medicare spending and some part B spending (e.g., home health services paid under Part B)</p>	<p>Includes primary outcome of institutional spending and payments to...</p> <ul style="list-style-type: none"> • physicians and other providers for inpatient and outpatient services • independent clinical laboratories • ambulance providers • free-standing ambulatory surgical centers <p>Adds remainder of Part B Medicare spending. This does not include Part D or pharmaceutical spending.</p>

Table S3: Differences in Pre-Period Trends in Outcomes in Intervention vs. Control MSAs, Primary Analysis and Alternate Model Specification

	Primary Analysis (Hospital + MSA Random Effects)				Alternate Specification (Hospital Fixed Effects + MSA Clustered Standard Errors)			
	Pre-Period Quarterly Trend, 2015 Alone	p-value	Pre-Period Annual Trend, 2011-2015	p-value	Pre-Period Quarterly Trend, 2015 Alone	p-value	Pre-Period Annual Trend, 2011-2015	p-value
Institutional Spending (\$)	21.6	0.75	90.9	0.02	29.3	0.64	88.2	0.11
Composite Rate of Complications (%)	0.001	0.17	0.0001	0.84	0.001	0.13	0.0001	0.78
Proportion of Episodes with Top Quartile of Patient Risk	-0.001	0.56	-0.0005	0.72	-0.001	0.67	-0.0002	0.87
Average Patient Risk Score	-5.26	0.85	1.24	0.96	0.37	0.99	9.1	0.73
Total Spending (20% sample) (\$)	86.1	0.43	93.5	0.03	120.2	0.37	91.7	0.10
Index LEJR Inpatient (\$)	-3.7	0.94	22.3	0.05	9.8	0.7	24	0.34
Other Inpatient (e.g. Readmissions) (\$)	21.8	0.32	-4.9	0.80	22	0.32	-15.5	0.45
Skilled Nursing Facility (\$)	18.8	0.52	55.0	0.03	15.1	0.67	58	0.2
Inpatient Rehabilitation Facility (\$)	-31.7	0.05	-8.4	0.54	-29.3	0.17	-10.4	0.71

Long Term Care Hospital (\$)	-6.2	0.42	14.1	0.04	-4.7	0.51	13.9	0.10
Home Health Agency (\$)	7.2	0.23	12.3	0.01	7.8	0.52	11.1	0.45
Hospice (\$)	0.1	0.96	0.9	0.66	0.2	0.94	1.2	0.56
Outpatient Facility (\$)	5.9	0.35	4.4	0.35	7.2	0.25	5.8	0.35
Durable Medical Equipment (\$)	1.4	0.46	0.2	0.89	1.2	0.67	0.08	0.97
Carrier (Professional Services, 20% sample) (\$)	4.6	0.80	3.5	0.65	12.8	0.43	5.8	0.62
90 Day Mortality (%)	0.0005	0.36	0.0004	0.32	0.0006	0.19	0.0003	0.35
Medicare Defined Complication Rate (No Fractures, %)	0.0004	0.50	-0.0004	0.50	0.0004	0.53	-0.0002	0.68
Discharged to Institutional Post Acute Care (PAC) (%)	0.002	0.16	0.003	0.04	0.002	0.45	0.003	0.26
Discharged to Home Health Agency (HHA) (%)	-0.0001	0.95	0.0002	0.88	0.0001	0.96	0.0001	0.96
Index LEJR Hospitalization Length of Stay, Days (mean)	0.0083	0.16	0.0054	0.29	0.011	0.11	0.006	0.57
Institutional PAC Length of Stay, Days (mean)	0.0050	0.96	0.18	0.01	-0.009	0.93	0.19	0.07

Non-IRF Institutional PAC Length of Stay, Days (mean)	-0.02	0.82	0.17	0.03	-0.22	0.85	0.16	0.14
HHA Number of Episodes (mean)	-0.0629	0.08	-0.0179	0.56	-0.084	0.18	-0.025	0.64
All Cause Inpatient Readmission	-0.0007	0.51	0.051	0.11	-0.0005	0.64	0.065	0.05
Emergency Department (ED) Visit without Admission	0.0005	0.71	-0.0002	0.81	0.0006	0.19	-0.0002	0.68
Observation Stay without Admission	-0.0004	0.44	-0.0001	0.79	-0.0004	0.42	0.00003	0.95
Any ED/Observation/Inpatient Visit	-0.0010	0.48	0.0018	0.08	-0.0007	0.64	0.0016	0.27

To estimate pre-period trends in our primary and secondary outcomes, we used data from 2015 alone as well as 2011-2015. Using 2011-2015 was a post-hoc change to our pre-specified approach using just 2015 for pre-period trends. The table shows the results of pre-period trend tests for both time periods both for our primary analysis reported in the paper (using hospital + MSA random effects), as well as for an alternate model specification (hospital fixed effects and MSA clustered standard errors, including multiple LEJR procedures per beneficiary).

*For the patient risk score outcomes, we fitted the risk score models using data from 2 years prior to the earliest year in the pre-period models, so that the data used for the model did not overlap with the pre-period for assessing trends. Therefore, for the 2011-2015 period, we fitted the predicted spending model using 2009-2010 data, while for the 2015 period, we fitted the predicted spending model using 2013-2014 data. See Appendix Methods section E for further details on model specification.

Table S4: Hospital Characteristics in Treatment and Control Groups in Period Before LEJR Implementation, 2015

Hospital Characteristics	Treatment MSAs	Control MSAs
Hospitals (unweighted N)	803	962
Hospitals (weighted N)	780	981
Control Status		
Non-profit	63.3%	63.1%
Profit	23.3%	26.1%
Government	13.3%	10.8%
Number of Beds		
<100	14.1%	20.5%
100-249	34.2%	34.5%
>250	51.7%	45.0%
Teaching Status		
Major	19.2%	18.4%
Minor	18.7%	16.2%
Non-Teaching	62.1%	65.4%
Census Region		
North	17.2%	15.5%
South	20.9%	26.5%
Midwest	30.1%	38.8%
West	31.8%	17.1%
Puerto Rico	0.0%	2.1%
LEJR Procedure Volume over 2015-2017		
Q1 (<64)	24.56%	23.37%
Q2 (64-203)	27.50%	22.86%
Q3 (204-500)	24.22%	26.56%
Q4 (501-4420)	23.71%	27.22%
DSH Percent*		
Q1 (<0.16)	24.4%	25.1%
Q2 (0.16-0.25)	24.5%	25.5%
Q3 (0.25-0.35)	22.4%	27.2%
Q4 (>0.35)	28.5%	21.8%

Abbreviations: diagnosis related group (DRG), lower extremity joint replacement (LEJR), disproportionate share hospital population (DSH). Hospital characteristics were measured using the Medicare Provider of Services and Impact files.

* The Disproportionate Share Hospital adjustment percentage (DSH percentage) of each hospital is a federally defined measure that determines additional payments for uncompensated care, is largely based on the percentage of admissions at a hospital that are for Medicaid patients and low-income Medicare patients. We divided hospitals into quartiles based on the DSH percentage in the 2015 Impact file. 1.1% of hospitals were missing DSH percentages and are not included in the quartiles, but were included in all analyses.

Table S5: Comparison of Treatment and Control Metropolitan Statistical Areas (MSAs)

	Treatment MSAs (n=75)	Control MSAs (n=121)
County Characteristics		
Counties (unweighted N)	220	295
Counties (weighted N)	216	298
County Population (mean)	458,326	385,136
% Medicare Advantage Penetration	32.2	30.6
SNF Total Beds in county (mean)	2078	1754

Comparison of counties on population, Medicare Advantage penetration, and availability of SNF total beds uses sampling weights (described in Appendix Methods A)

Table S6: Comparison of Differential Change in Institutional Spending Over Time After CJR Implementation

Post-Implementation Time Period	Adjusted Difference-in-Differences Estimate for Institutional Spending (\$, 95% CI) ^a	Difference (95% CI) vs. 1 st or 2 nd 6-month period after CJR implementation	
		Vs. April-September 2016	Vs. October 2016 - March 2017
1 st period: April 2016 - September 2016	-541 (-754, -328)	--	--
2 nd period: October 2016 – March 2017	-832 (-1044, -621)	-291 (-535, -47)	--
3 rd period: April 2017 - September 2017	-860 (-1075, -645)	-319 (-566, -71)	-27 (-272, 217)

Abbreviations: Comprehensive Care for Joint Replacement initiative (CJR), confidence interval (CI)

Adjusted estimates show the differential change in institutional spending between treatment and control groups after vs. before CJR implementation for the time period indicated in each row. The pre-implementation period is January-December 2015. Note that the first row includes the period from April-June 2016, which is excluded in the main pre-specified analysis in **Table 2** and **Table 4**. All estimates adjust for hospital and MSA random effects, as well as patient and episode characteristics as described in the Methods and Appendix Methods.

Table S7: Differential Change in LEJR Volume Before and After CJR Implementation

Number of LEJR Procedures per 1,000 Beneficiaries	Treatment			Control			Adjusted Difference-in-Differences Estimate	95% CI
	Pre-Period Episode Average	Post-Period Episode Average	Unadjusted Difference	Pre-Period Episode Average	Post-Period Episode Average	Unadjusted Difference		
Individual-Level Model (Procedures per 9-Month Period)	9.95	10.5	0.55	10.5	11.1	0.64	-0.004	-0.03, 0.02
MSA-Level Model (Procedures per 3-Month Period [Quarter])	3.94	4.00	0.06	3.91	3.97	0.06	-0.006	-0.08, 0.07

This Table shows the unadjusted and adjusted changes in number of LEJR procedures per 1,000 beneficiaries in the treatment and control groups. Neither of the adjusted results are statistically significant. We are not reporting p-values because these are secondary outcomes.

For the individual-level model, we performed a beneficiary-period level analysis using logistic regression, where the outcome was whether any continuously-enrolled beneficiary in a treatment or control MSA had any LEJR procedure in a 9-month time period in the pre-period (January-September 2015) or post-period (January-September 2017). These time periods were chosen to have balanced representation of seasonality (same 9 months in both periods). Our post-period in the main analysis (15 months) is longer than the pre-period (12 months). We therefore chose the same nine-month window in a year to avoid biasing our estimates. The adjusted difference-in-differences estimate controls for hospital fixed effects, as well as patient and episode characteristics as described in the Methods and Appendix Methods.

For the MSA-level model, we performed an MSA-quarter level analysis using linear regression, where the outcome was the rate of LEJR procedures per 1,000 beneficiaries enrolled in Medicare in that MSA for that quarter. The adjusted difference-in-differences

estimate controls for MSA fixed effects, as well as all patient characteristics in the individual-level model aggregated at MSA-level averages.

Table S8: Difference-in-Differences Sensitivity Analyses with Alternate Model Specifications for Primary and Secondary Outcomes

	(1) Alternate Specification Linear Model + Hospital Fixed Effects and MSA Clustered Standard Errors (95% CI)	(2) Alternate Specification Linear Model + MSA Random Effects (95% CI)	(3) Alternative Model (GLM) GLM with Log Link + MSA Random Effects (95% CI)	(4) Alternative Model (Logit) Logistic Model + MSA Random Effects (95% CI)
Primary Outcomes				
Institutional Spending (\$)	-796 (-1,126, -465)	-940 (-1107, -771)	-907 (-1068, -747)	--
Composite Rate of Complications	0.06% (-0.15, 0.26)	-0.05% (-0.25, 0.15)	--	-0.09% (-0.35, 0.17)
Proportion of Episodes in Top Quartile of Patient Risk	-0.33% (-0.83, 0.17)	-0.42% (-0.87, 0.04)	--	-0.35% (-0.84, 0.14)
Secondary Outcomes				
Total Spending (20% sample)	-1,036 (-1,440, -632)	-1,253 (-1578, -927)	-1196 (-1515, 879)	--
Discharged to Institutional PAC (%)	-2.70% (-4.0, -1.3)	-2.97% (-3.4, -2.5)	--	-4.90% (-5.59, -4.3)
Discharged to Home Health Agency (HHA) (%)	2.40% (-0.6, 5.4)	3.30% (2.8, 3.8)	--	4.00% (3.5, 4.6)
Institutional PAC Length of Stay, Days (mean)	-1.7 (-2.3, -1.1)	-1.9 (-2.1, -1.6)	-2.1 (-2.4, -1.8)	--
Non-IRF Institutional PAC Length of Stay, Days (mean)	-2.1 (-2.8, -1.4)	-2.3 (-2.6, -2.0)	-2.4 (-2.5, -2.3)	--
HHA Number of Episodes (mean)	-0.9 (-1.6, -0.4)	-0.8 (-0.9, -0.7)	-0.9 (-1.0, -0.8)	--
Medicare Defined Complication Rate (No Fractures, %)	-0.04% (-0.2, 0.1)	-0.07% (-0.25, 0.11)	--	-0.08% (-0.31, 0.14)

90 Day Mortality (%)	0.00% (-0.1, 0.1)	-0.05% (-0.20, 0.10)	--	-0.10% (-0.29, 0.93)
All Cause Inpatient Readmission	-0.50% (-1.0, 0.03)	-0.76% (-1.1, -0.5)	--	-0.80% (-1.2, -0.5)
Emergency Department (ED) Visit without Admission	0.20% (-0.1, 0.6)	0.28% (-0.09, 0.65)	--	0.28% (-0.10, 0.66)
Observation Stay without Admission	0.04% (-0.1, 0.1)	0.05% (-0.11, 0.21)	--	0.04% (-0.11, 0.20)
Any ED/Observation/Inpatient Visit	-0.30% (-1.0, 0.4)	-0.51% (-0.95,-0.07)	--	-0.56% (-1.0, -0.1)

This Table shows the results of the difference-in-difference analyses presented in **Tables 2** and **4** using alternate regression models: using hospital fixed effects with robust standard errors clustered at the MSA level, generalized linear models with a log link for spending and length of stay outcomes, and logistic regression models for binary utilization and quality outcomes. For the non-linear models, hospital random effects were not included due to poor model convergence fitting >1700 hospital random effects. Because of this, their results may differ from the primary results reported in **Tables 2** and **4** due to this difference in specification alone. For this reason, we included a set of results using the same model as presented in **Tables 2** and **4**, but omitting hospital random effects. The comparison between this model and the non-linear models shows the sensitivity of our results to using models with different functional forms.

Column (1) shows models as specified in our pre-analysis protocol (see **Appendix Methods D** above) using hospital fixed effects with robust standard errors clustered at the MSA level. This model included multiple LEJR procedures per beneficiary as specified in our pre-analysis plan.

Column (3) shows results using generalized linear models with a log link and proportional to mean variance function for continuous spending and count outcomes. These models control for all covariates in the primary analysis in **Tables 2** and **4** and MSA random effects. Compare these results to column (2) to examine the difference between results from the non-linear vs. the linear model.

Column (4) results using logistic regression to model the difference-in-difference outcome for binary outcomes. These models control for all covariates in the primary analysis in **Tables 2** and **4** and MSA random effects. Compare these results to column (2) to examine the difference between results from the non-linear vs. the linear model.

Table S9: Differential Change in Average Patient Risk Score Before and After CJR Implementation

	Treatment			Control			Adjusted Difference-in-Differences Estimate	
	Pre-Period Episode Average	Post-Period Episode Average	Unadjusted Difference	Pre-Period Episode Average	Post-Period Episode Average	Unadjusted Difference	95% CI	
Patient Risk Score (Predicted 2013-2014 \$, mean)	30,288	29,945	-342	29,768	29,472	-296	-22	-106, 62

This Table shows the unadjusted and adjusted changes in patient risk score, as defined in predicted spending, in the treatment and control groups. This is the output from the patient risk model described in the Appendix Methods, section C. The adjusted difference-in-differences estimate controls for hospital and MSA random effects, and does **not** control for patient or episode characteristics since these were used to generate the risk score itself. We repeated this analysis using a predicted spending model including hospital fixed effects, which had no impact on the results.

Table S10: Differential Change in Patient and Procedure Characteristics Before and After CJR Implementation

	Treatment			Control			Adjusted Difference-in-Differences Estimate	
	Pre-Period Episode Average	Post-Period Episode Average	Unadjusted Difference	Pre-Period Episode Average	Post-Period Episode Average	Unadjusted Difference	95% CI	
Diagnosis related group								
DRG 469 (LEJR with major complication/comorbidity)	5.6%	5.5%	-0.1%	5.1%	4.9%	-0.3%	0.3%	0.1, 0.6
DRG 470 (LEJR without major complication or comorbidity)	94.4%	94.5%	0.1%	94.9%	95.1%	0.3%	-0.3%	-0.56, -0.07
Fracture	16.2%	15.3%	-0.9%	15.0%	14.3%	-0.7%	0.3%	-0.04, 0.72
Procedure Type								
Total Knee	54.2%	53.6%	-0.6%	56.2%	55.2%	-1.0%	0.3%	-0.23, 0.86
Total Hip	31.4%	32.4%	1.0%	30.9%	32.1%	1.2%	-0.5%	-1.06, -0.04
Partial Hip/Knee	13.7%	12.7%	-1.1%	12.4%	11.7%	-0.8%	0.2%	-0.2, 0.51
Unknown	0.6%	1.3%	0.7%	0.5%	1.0%	0.6%	0.1%	-0.03, 0.17
Age								
<65	7.9%	6.8%	-1.2%	7.9%	7.2%	-0.6%	-0.5%	-0.82, -0.24
65-69	22.7%	23.4%	0.7%	23.1%	23.3%	0.2%	0.4%	-0.02, 0.91
70-74	24.5%	25.7%	1.2%	25.2%	26.4%	1.2%	-0.1%	-0.56, 0.4
75-79	19.6%	20.0%	0.3%	19.7%	19.8%	0.1%	0.2%	-0.2, 0.68
80-84	13.3%	12.8%	-0.5%	13.0%	12.7%	-0.3%	-0.2%	-0.55, 0.18
85-89	7.7%	7.2%	-0.5%	7.4%	7.0%	-0.4%	0.0%	-0.33, 0.25
90+	4.2%	4.1%	-0.1%	3.8%	3.6%	-0.2%	0.1%	-0.08, 0.34
Male	35.9%	36.1%	0.2%	36.0%	36.5%	0.4%	-0.1%	-0.68, 0.38
Race/Ethnicity								
Non-Hispanic White	89.8%	89.6%	-0.2%	90.8%	90.6%	-0.3%	0.0%	-0.29, 0.35
Non-Hispanic Black	5.9%	5.3%	-0.6%	5.6%	5.4%	-0.2%	-0.2%	-0.45, 0.04
Asian	1.0%	1.1%	0.2%	0.6%	0.6%	0.0%	0.1%	-0.01, 0.19
Other Race/Ethnicity	2.2%	2.9%	0.6%	2.2%	2.6%	0.4%	0.2%	0, 0.34
Hispanic	1.1%	1.1%	0.0%	0.7%	0.8%	0.1%	-0.1%	-0.2, 0.01
Urban Residence ¹	85.0%	85.7%	0.7%	82.4%	82.1%	-0.3%	0.3%	-0.07, 0.67
Original Reason for Entitlement								
Age >65	84.2%	85.4%	1.2%	84.1%	84.6%	0.5%	0.6%	0.18, 0.97

	Disability	15.7%	14.5%	-1.2%	15.8%	15.3%	-0.5%	-0.6%	-0.97, -0.18
	ESRD ²	0.1%	0.1%	0.0%	0.1%	0.1%	0.0%	0.0%	-0.04, 0.03
	Medicaid Eligible	11.4%	10.5%	-0.9%	10.3%	9.9%	-0.4%	-0.4%	-0.75, -0.1
	Prior Utilization								
	Prior Inpatient	20.4%	15.3%	-5.1%	19.7%	14.8%	-4.9%	0.0%	-0.38, 0.45
	Prior PAC	7.7%	5.3%	-2.3%	7.2%	5.0%	-2.2%	-0.1%	-0.34, 0.19

Adjusted estimates show the differential change in the proportion of patient characteristics between treatment and control groups after vs. before CJR implementation. All estimates adjust for hospital and MSA random effects. The differential change estimates can be interpreted as the average within-hospital change attributable to CJR implementation.

¹Urban location defined using the Health Resources and Service Administration (HRSA) Rural-Urban Commuting Area (RUCA) code database (<http://depts.washington.edu/uwruca/index.php>). Urban was defined as a patient residing in a “metropolitan” ZIP code. Data was missing for 0.16% of episodes which were largely located in Puerto Rico.

²Patients with ESRD were excluded from the payment program. However, some patients initially qualified for Medicare due to ESRD but no longer classified as ESRD at the time of LEJR

Table S11: Results for Selected Outcomes Controlling for Separate Groups of Patient Characteristics

	(1)	(2)	(3)	(4)
	Hospital FE, Quarter FE	Add Demographics	Add CCWs	Add Index LEJR Characteristics
Institutional Spending (\$)				
Coefficient on treatment X post	-767.5	-745.3	-748.0	-812.3
SE	96.7	92.3	90.2	86.1
p value	<0.001	<0.001	<0.001	<0.001
AIC	11999243	11949438	11924685	11874078
Total Spending (\$, 20% Sample)				
Coefficient on treatment X post	-972.5	-969.5	-1002.3	-1084.4
SE	197.3	184.9	178.6	165.7
p value	<0.001	<0.001	<0.001	<0.001
AIC	2490355	2475647	2467905	2451084
Any Complication (%)				
Coefficient on treatment X post	0.08	0.07	0.06	-0.04
SE	0.11	0.11	0.11	0.10
p value	0.50	0.53	0.62	0.67
AIC	-186681	-199916	-207779	-272507

*** p<0.001, ** p<0.01, * p<0.05

Abbreviations: Akaike Information Criterion (AIC), random effects (RE), lower extremity joint replacement (LEJR), standard error (SE)

This Table shows the effect on two primary outcomes (institutional spending and complications) and one key secondary outcome (total spending, 20% sample) of successively adding patient characteristics to our models. The “coefficient on treatment X post” row shows the main difference-in-difference outcome of interest. The AIC gives an estimate of the relative quality of each model, with

lower values being equivalent to models which fit the data more effectively. It is not possible to compare the AIC across models with different outcomes, therefore the AIC for each of the three outcomes have completely different absolute baseline values. Within each outcome, however, the AIC gives an estimate of the improvement in fit with adding more variables.

Column (1) shows the difference-in-difference models adjusted for hospital, MSA random effects and quarter fixed effects.

In column (2), we add patient demographics, which includes all variables except the 27 chronic condition indicators described above in Appendix Methods section D. Notably, this includes the characteristics in Table 4 with differential changes in the treatment group, such as entitlement to Medicare due to disability.

In column (3), we add the 27 indicators for chronic conditions described above in Appendix Methods section D.

The results in column (4) include LEJR episode characteristics: the presence of hip fracture and procedure type. These are the results reported in the main manuscript **Tables 2 and 4**.