Cardiovascular Outcomes and All-cause Mortality Following Measurement of Endogenous Testosterone Levels

Kasper Adelborg, MD, PhD[a,b,*], Thomas Bøjer Rasmussen, MSc[a], Helene Nørrelund, MD, PhD, DMSc[a], J. Bradley Layton, PhD[c,d], Henrik Toft Sørensen, MD, PhD, DMSc[a], and Christian Fynbo Christiansen, MD, PhD[a]

Although reduced testosterone levels are common in aging populations, the clinical consequences remain to be further explored. We examined whether low total testosterone levels are associated with stroke (ischemic and hemorrhagic), myocardial infarction (MI), venous thromboembolism (VTE), and all-cause mortality in adult men. We conducted a cohort study in the Central Denmark Region (2000 to 2015). We included all men with a first-ever laboratory testosterone result and computed the 5-year risks of cardiovascular outcomes and all-cause mortality. Propensity score-weighted hazard ratios were computed, comparing persons with normal versus low testosterone levels. Individuals were censored at testosterone treatment during follow-up (3%). We identified 4,771 men with low testosterone levels and 13,467 with normal levels. Persons with low testosterone levels were older (median ages, 55 years vs 50 years) and had more co-morbidities than men with normal testosterone levels. Persons with low testosterone had higher 5-year risks of stroke (2.4% vs 1.5%), MI (1.5% vs 1.2%), VTE (1.4% vs 0.9%), and all-cause mortality (17.8% vs 6.8%) than persons with normal testosterone levels. After propensity score-weighting, the associations with cardiovascular outcomes reached unity. The 5-year hazard ratios were 1.14 (95% confidence intervals [CIs] 0.87 to 1.49) for stroke, 0.95 (95% CI 0.70 to 1.30) for MI, 1.10 (95% CI 0.78 to 1.55) for VTE, whereas it was 1.48 (95% CI 1.32 to 1.64) for all-cause mortality. In conclusion, low testosterone level was a strong predictor for cardiovascular outcomes and all-cause mortality in unadjusted models, however only the association between low testosterone and all-cause mortality persisted after adjustment for age and co-morbidity. **(Am J Cardiol 2019;123:1757−1764)**

Some recent studies suggest that testosterone therapy is associated with an increased risk of stroke, myocardial infarction (MI), and death.[1−5] As a consequence, testosterone has come under scrutiny by the US Food and Drug Administration.[6] Health Canada[7] also strongly warns about potential cardiovascular side effects in testosterone users, although the body of literature still lacks conclusive evidence. One explanation for the observed associations may be confounding by indication, as low levels of endogenous testosterone levels *per se* might be linked to stroke, MI, and death.[8−14] Additional analyses to clarify this association are therefore needed. In some countries, including Denmark, use of testosterone therapy has been limited, making it possible to examine the effect of endogenous testosterone levels on risk of cardiovascular outcomes. We therefore examined the risk of stroke, MI, venous thromboembolism (VTE), and all-cause mortality in men with normal versus low testosterone levels in a Danish cohort study.

Methods

This study was conducted in the Central Denmark Region from January 1, 2000 to November 1, 2015. This Region has a population of 1.3 million inhabitants (24% of the Danish population).[15] Denmark has a tax-supported health care system that guarantees unfettered access to medical care for all residents, as well as partial reimbursement of prescribed drugs.

The study population consisted of all male inhabitants with a first-ever testosterone measurement in the study period, where the person was 18 years or older at the time of the measurement. Data on testosterone measurements were obtained from the Clinical Laboratory Information System Research Database (LABKA), using Nomenclature for Properties and Units codes.[16] The database includes laboratory results from all hospitals, outpatient clinics, and general practices in the Central Denmark Region since 2000. Only men whose measurement had a nonmissing result were used and only individuals who lived in the

[a]Department of Clinical Epidemiology, Aarhus University Hospital, Aarhus, Denmark; [b]Department of Clinical Biochemistry, Aarhus University Hospital, Aarhus, Denmark; [c]Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina; and [d]RTI Health Solutions, North Carolina. Manuscript received December 14, 2018; revised manuscript received and accepted February 22, 2019.

Funding: This work was supported by the Program for Clinical Research Infrastructure (PROCRIN) established by the Lundbeck Foundation and the Novo Nordisk Foundation.

See page 1763 for disclosure information.

***Corresponding author**: Tel: +45 20 34 17 24; fax: +45 87167215.
E-mail address: kade@clin.au.dk (K. Adelborg).

0002-9149/
https://doi.org/10.1016/j.amjcard.2019.02.042

Central Denmark Region at the time of the measurement were eligible for inclusion. Total testosterone laboratory methods included a variety of immunoassays and liquid chromatography-tandem mass spectrometry.

Testosterone levels were categorized according to age-specific reference values for normal and low levels (Supplementary Table 1). If a man's first-ever measurement had a result classified as high, the patient was excluded from the study. Patients with prostate cancer before the test, identified from the Danish National Patient Registry (DNPR), also were excluded, because most have low testosterone levels due to antiandrogen therapy, and we were unable to identify a sufficient number of comparators with normal testosterone levels for these patients. The DNPR is an ongoing population-based registry. It has collected data on admission and discharge dates as well as diagnoses from all nonpsychiatric hospitals since 1977 and on emergency room and outpatient clinic visits since 1995. Each hospital discharge or outpatient visit is recorded in the DNPR with one primary diagnosis and one or more secondary diagnoses coded according to the *International Classification of Diseases, Eighth Revision* between 1977 and 1993 and *Tenth Revision* thereafter. We also excluded individuals treated with exogenous testosterone and antiandrogen therapy within 90 days before the first testosterone test results. The index date was the date of the first-ever testosterone result.

Outcomes included 1-year and 5-year risk of first-time stroke, MI, and VTE identified in the DNPR, based on primary and secondary diagnoses from inpatient and outpatient hospital contacts.[17] Emergency room diagnoses were not considered due to the assumed low positive predictive value of diagnoses in this setting since they are initial working diagnoses.[17] Secondarily, we also examined all-cause mortality. These data were obtained from the Danish Civil Registration System, which provides daily updates on vital statistics, including dates of emigration and death.[18]

Data on the most recent albumin level, follicle-stimulating hormone status, and luteinizing hormone status (all categorized into missing, low, normal, and high) within the previous year or at the index date were retrieved from the LABKA database. Using all inpatient and outpatient clinic diagnoses, data on several co-morbidities and potential causes of low testosterone within 10 years before the index date were retrieved from the DNPR.[17]

We also retrieved information on various filled prescriptions within 90 days before the index date from the Aarhus University Prescription Database, using the Anatomical Therapeutic Chemical Classification System.[19,20] The Aarhus University Prescription Database contains complete information on all prescriptions redeemed in the Central Denmark Region since 1998. All Nomenclature for Properties and Units, Anatomical Therapeutic Chemical, and *International Classification of Diseases* codes are provided in Supplementary Table 2.

For each outcome-specific analysis, persons were followed from the date of their first-ever testosterone laboratory test until the date of the outcome, death (unless the outcome of interest was all-cause mortality), date of emigration, 1 or 5 years of follow-up depending on the analysis, or 31 December 2015, whichever occurred first. For each outcome, patients with a previous event were excluded from the analysis (e.g., when VTE was the outcome, patients with previous VTE were excluded). In addition, individuals were censored at testosterone treatment during follow-up. We described individuals with low and normal testosterone levels according to the covariates listed above and presented these data only for individuals included in the all-cause mortality analysis. In addition, the number of individuals initiating testosterone treatment at hospitals or through redemption of a prescription during follow-up was tabulated. We calculated incidence rates of the outcomes per 100 person-years and calculated 1-year and 5-year risks of the outcomes, comparing men with low versus normal testosterone levels. We also plotted cumulative risks, accounting for death as a competing risk.

Each patient's propensity score were estimated with generalized boosted models using the covariates shown in Table 1 and then transformed the propensity score into inverse probability of treatment weights (IPTW),[21] which permits estimation of an average treatment effect in the treated population.[22] We computed hazard ratios (HRs) with 95% confidence intervals (CIs) using Cox regression analysis before and after propensity score weighting, to compare individuals with low testosterone levels to those with normal levels. To assess the balance of covariates after propensity score weighting we estimated the standardized difference for all covariates included in the propensity score.[22] Furthermore, we estimated and compared the empirical cumulative distribution function for each of the continuous covariates.[22] Because age may modify the effect of testosterone level, we stratified the analyses by age groups.

We examined proportionality of hazards assumption using log(-log) plots, and the assumption was found to be appropriate. All analyses were conducted using SAS version 9.4 (SAS Institute, Cary, North Carolina). The study was approved by the Danish Data Protection Agency (record number: 2013-41-1924). According to Danish law, use of registry data does not require informed consent from patients.

Results

We identified 13,467 individuals with a normal testosterone level and 4,771 individuals with a low level (Table 1 and Figure 1). The cohort of persons with low testosterone was older than the cohort of persons with a normal level (median age, 55 years vs 50 years, respectively). Persons with low testosterone had a substantial higher prevalence of co-morbidities than persons with normal testosterone levels. The prevalence of hypogonadism, hypopituitarism, Klinefelter's syndrome, Down's syndrome, testicular torsion, varicocele, cryptorchidism, and orchitis were all below 1% in the study population as a whole. After propensity score weighting, the standardized difference of all covariates were less than 0.1. Furthermore, the empirical cumulative distribution functions of continuous variables was almost identical after weighting. Both indicate that covariate balance was achieved after weighting. During the first 5 years of follow-up, 485 individuals (~3% of the study population) initiated testosterone treatment.

Persons with low testosterone had a higher 1-year risk of stroke (1% vs 0.5%), MI (0.7% vs 0.4%), VTE (0.7% vs 0.3%), and all-cause mortality (9% vs 2%) than persons

Table 1
Characteristics of individuals with normal and low testosterone levels, Denmark, 2000 to 2015

	Unweighted cohorts		Propensity score weighted cohorts	
Variable	Low testosterone	Normal testosterone	Low testosterone	Normal testosterone
Number of men	4,771	13,467	4,771.0	4,470.7
Median age (25th−75th percentiles)	55.4 (38.3−69.2)	50.4 (33.8−63.4)	55.4 (38.3−69.2)	54.8 (37.9−68.5)
Albumin level				
Low	794 (17%)	734 (5.5%)	794.0 (17%)	652.1 (15%)
Normal	2,072 (43%)	5,772 (43%)	2,072.0 (43%)	1,985.2 (44%)
High	371 (7.8%)	1,031 (7.7%)	371.0 (7.8%)	336.4 (7.5%)
Missing	1,534 (32%)	5,930 (44%)	1,534.0 (32%)	1,496.9 (34%)
Co-morbidities				
Myocardial infarction	216 (4.5%)	301 (2.2%)	216.0 (4.5%)	172.4 (3.9%)
Congestive heart failure	292 (6.1%)	295 (2.2%)	292.0 (6.1%)	233.5 (5.2%)
Peripheral vascular disease	270 (5.7%)	354 (2.6%)	270.0 (5.7%)	214.9 (4.8%)
Cerebrovascular disease	400 (8.4%)	602 (4.5%)	400.0 (8.4%)	338.6 (7.6%)
Dementia	48 (1.0%)	45 (0.3%)	48.0 (1.0%)	29.3 (0.7%)
Chronic pulmonary disease	502 (11%)	780 (5.8%)	502.0 (11%)	436.4 (9.8%)
Connective tissue disease	193 (4.0%)	459 (3.4%)	193.0 (4.0%)	178.8 (4.0%)
Ulcer disease	171 (3.6%)	236 (1.8%)	171.0 (3.6%)	127.9 (2.9%)
Mild liver disease	103 (2.2%)	184 (1.4%)	103.0 (2.2%)	94.5 (2.1%)
Diabetes without end-organ damage	522 (11%)	803 (6.0%)	522.0 (11%)	456.0 (10%)
Hemiplegia	31 (0.6%)	40 (0.3%)	31.0 (0.6%)	20.1 (0.4%)
Moderate to severe renal disease	219 (4.6%)	214 (1.6%)	219.0 (4.6%)	162.4 (3.6%)
Diabetes with end-organ damage	295 (6.2%)	430 (3.2%)	295.0 (6.2%)	247.7 (5.5%)
Moderate to severe liver disease	36 (0.8%)	49 (0.4%)	36.0 (0.8%)	30.4 (0.7%)
AIDS	9 (0.2%)	41 (0.3%)	9.0 (0.2%)	10.2 (0.2%)
Hypogonadism	15 (0.3%)	11 (0.1%)	15.0 (0.3%)	10.2 (0.2%)
Hypopituitarism	37 (0.8%)	35 (0.3%)	37.0 (0.8%)	23.8 (0.5%)
Klinefelter's syndrome	8 (0.2%)	11 (0.1%)	8.0 (0.2%)	6.3 (0.1%)
Down's syndrome	7 (0.1%)	3 (0.0%)	7.0 (0.1%)	4.3 (0.1%)
Testicular torsion	5 (0.1%)	21 (0.2%)	5.0 (0.1%)	4.8 (0.1%)
Varicocele	8 (0.2%)	30 (0.2%)	8.0 (0.2%)	8.6 (0.2%)
Cryptorchidism	25 (0.5%)	53 (0.4%)	25.0 (0.5%)	22.6 (0.5%)
Orchitis	54 (1.1%)	134 (1.0%)	54.0 (1.1%)	45.1 (1.0%)
Chronic kidney disease	222 (4.7%)	244 (1.8%)	222.0 (4.7%)	183.9 (4.1%)
Myxedema	46 (1.0%)	72 (0.5%)	46.0 (1.0%)	32.7 (0.7%)
Obesity*	240 (5.0%)	255 (1.9%)	240.0 (5.0%)	196.7 (4.4%)
Alcoholism	331 (6.9%)	718 (5.3%)	331.0 (6.9%)	297.1 (6.6%)
Hypertension	785 (17%)	1,234 (9.2%)	785.0 (17%)	693.4 (16%)
Any cancer (except prostate cancer)	586 (12%)	1,195 (8.9%)	586.0 (12%)	532.6 (12%)
Illicit drug abuse	30 (0.6%)	38 (0.3%)	30.0 (0.6%)	17.6 (0.4%)
Comedications				
ACE/ARB	966 (20%)	1,785 (13%)	966.0 (20%)	863.4 (19%)
Beta-blockers	595 (13%)	959 (7.1%)	595.0 (13%)	525.2 (12%)
Statins	881 (19%)	1,405 (10%)	881.0 (19%)	779.2 (17%)
Low-dose aspirin	711 (15%)	1,121 (8.3%)	711.0 (15%)	612.2 (14%)
Clopidogrel	102 (2.1%)	157 (1.2%)	102.0 (2.1%)	98.0 (2.2%)
Vitamin K antagonists	186 (3.9%)	344 (2.6%)	186.0 (3.9%)	177.5 (4.0%)
Diuretics	818 (17%)	941 (7.0%)	818.0 (17%)	684.8 (15%)
NSAID	679 (14%)	1,396 (10%)	679.0 (14%)	589.3 (13%)
Opiods	785 (17%)	1,132 (8.4%)	785.0 (17%)	666.9 (15%)
Antidepressants	688 (14%)	1,153 (8.6%)	688.0 (14%)	583.1 (13.0%)
Antipsychotics	199 (4.2%)	308 (2.3%)	199.0 (4.2%)	163.9 (3.7%)
Erectile dysfunction drugs	29 (0.6%)	151 (1.1%)	29.0 (0.6%)	30.2 (0.7%)

ACE/ARB, angiotensin-converting enzyme/angiotensin II receptor blockers; AIDS, acquired immune deficiency syndrome; NSAID, nonsteroidal anti-inflammatory drugs.

* Defined as hospital-based diagnoses of obesity. Data are counts (%), unless otherwise stated. The characteristics were tabulated for the individuals where all-cause mortality was the outcome of interest.

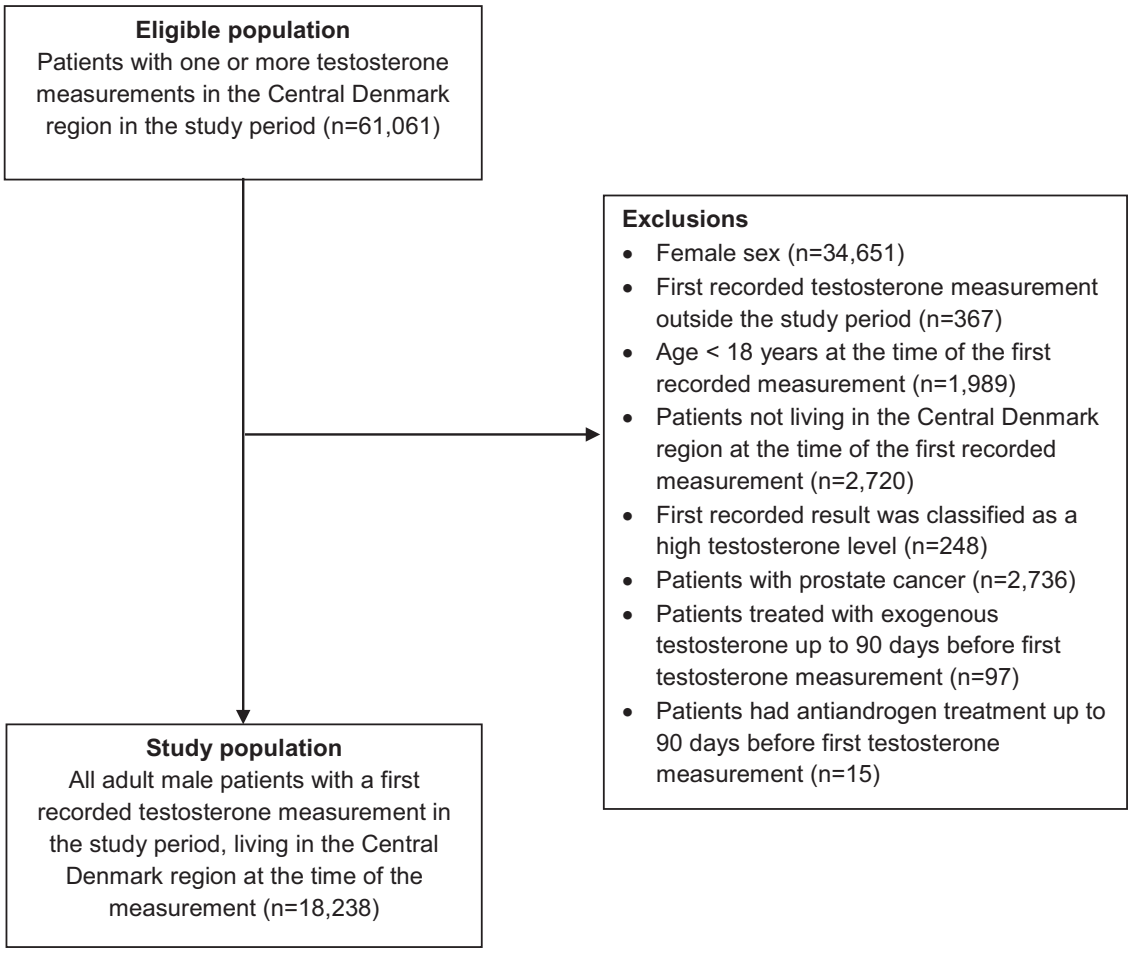


Figure 1. Flow chart of the study population.

with normal testosterone levels (Table 2 and Figure 2). Correspondingly, the 1-year unadjusted HRs was increased for stroke, MI, VTE, and all-cause mortality. Using 5 years of follow-up period, the unadjusted HRs were attenuated but remained elevated for all outcomes.

After accounting for measured confounders using IPTW, the cumulative incidence of stroke, MI, and VTE were comparable for persons with low and normal testosterone, whereas the cumulative incidence of all-cause mortality remained higher for persons with low testosterone levels (Figure 2). After applying IPTW, the 1-year HRs were 1.33 (95% CI 0.84 to 2.09) for stroke, 1.47 (95% CI 0.91 to 2.38) for MI, 1.10 (95% CI 0.65 to 1.85) for VTE, and 2.08 (95% CI 1.72 to 2.52) for all-cause mortality. The 5-year HRs for cardiovascular outcomes reached unity. For all-cause mortality, the association was attenuated but persisted (HR 1.48, 95% CI 1.32 to 1.64).

In analyses stratified by age, the associations were broadly consistent across all age groups with few exceptions (Supplementary Tables 3-6).

Discusssion

In this cohort study, in unadjusted models, a low testosterone level was a strong predictor for increased risk of stroke, MI, VTE, and all-cause mortality, especially in the first year, but also during 5 years. However, the increased risks of cardiovascular outcomes were largely explained by increased age and co-morbidity levels in persons with a low testosterone level. Thus, the associations were greatly attenuated after accounting for differences in these variables.

Previous studies have examined the association between endogenous testosterone level, mortality, and cardiovascular outcomes.[8,11,14] However, many were limited by low numbers of events,[13,23,24] reported only surrogate end points for cardiovascular outcomes (e.g., degree of aortic atherosclerosis),[23–25] and did not assess individual cardiovascular outcomes or included data on VTE.[8,11] Our analysis thus complements the literature by providing data on the association between low testosterone levels and several cardiovascular outcomes, accounting for several potential confounders, within a uniformly organized health care system, with complete individual-level linkage of data in various registries.

A previous meta-analysis of 19 studies examined the association between endogenous testosterone and atherosclerosis, stroke, MI, ischemic heart disease, death from coronary artery disease, and all-cause mortality.[11] In total, 18 studies had data on total testosterone level, with follow-up ranging between 3 and 15 years. A weak protective

Table 2
Risk of stroke, myocardial infarction, venous thromboembolism, and all-cause mortality in men with normal and low testosterone levels

Outcome by testosterone level	0–1 year of follow-up				0–5 years of follow-up			
	No. at risk/No. of events	Incidence rate* (95% CI)	Unadjusted hazard ratio (95% CI)	Hazard ratio after IPTW weighting (95% CI)	No. at risk/No. of events	Incidence rate* (95% CI)	Unadjusted hazard ratio (95% CI)	Hazard ratio after IPTW weighting (95% CI)
Stroke								
Normal	13,034/65	0.55 (0.43; 0.70)	1.00 (ref)	1.00 (ref)	13,034/199	0.47 (0.41;0.54)	1.00 (ref)	1.00 (ref)
Low	4,469/45	1.17 (0.87; 1.57)	2.13 (1.46; 3.11)	1.33 (0.84;2.09)	4,469/105	0.79 (0.65;0.95)	1.67 (1.32;2.12)	1.14 (0.87; 1.49)
Myocardial infarction								
Normal	13,026/50	0.42 (0.32; 0.55)	1.00 (ref)	1.00 (ref)	13,026/154	0.36 (0.31; 0.42)	1.00 (ref)	1.00 (ref)
Low	4,460/29	0.75 (0.52; 1.08)	1.77 (1.12; 2.80)	1.47 (0.91; 2.38)	4,460/65	0.49 (0.38; 0.62)	1.33 (1.00; 1.78)	0.95 (0.70; 1.30)
Venous thromboembolism								
Normal	13,266/42	0.35 (0.26; 0.47)	1.00 (ref)	1.00 (ref)	13,266/113	0.26 (0.22; 0.31)	1.00 (ref)	1.00 (ref)
Low	4,656/33	0.82 (0.59; 1.16)	2.37 (1.50; 3.73)	1.10 (0.65; 1.85)	4,656/65	0.47 (0.37; 0.60)	1.78 (1.31; 2.41)	1.10 (0.78; 1.55)
All-cause mortality								
Normal	13,467/263	2.14 (1.89; 2.41)	1.00 (ref)	1.00 (ref)	13,467/919	2.08 (1.95; 2.22)	1.00 (ref)	1.00 (ref)
Low	4,771/421	10.25 (9.31; 11.27)	4.75 (4.07; 5.54)	2.08 (1.72; 2.52)	4,771/848	5.95 (5.56; 6.36)	2.83 (2.58; 3.11)	1.48 (1.32; 1.64)

CI, confidence interval; IPTW, inverse probability of treatment weighting.
* Per 100 person-years. For nonfatal outcomes, individuals with previous events were excluded, for example, when stroke was the outcome, individuals with previous stroke were excluded to assess first-time events.

effect of a one-standard-deviation increase in total testosterone (overall risk ratio = 0.89, 95% CI 0.83 to 0.96) was reported, with a stronger association in men above age 70.[11] Another meta-analysis of 12 studies found that low endogenous testosterone was associated with increased risk of all-cause mortality (overall relative risk = 1.35, 95% CI 1.13 to 1.62), and cardiovascular mortality (overall relative risk = 1.25, 95% CI 0.97 to 1.60).[8] Consistent with these findings, a recent meta-analysis also found that low testosterone was a predictor for cardiovascular morbidity and mortality, in both unadjusted and fully adjusted models.[14]

Our analyses suggested that the increased risk of cardiovascular outcomes associated with low testosterone were driven mainly by age and co-morbidity, both of which themselves can contribute to reduced testosterone levels. As the CIs of the effect estimates for 1-year cardiovascular outcomes after applying IPTW were relatively wide, we cannot exclude entirely an association between testosterone level and some cardiovascular outcomes. However, this does not necessarily imply a causal link, as our findings could be susceptible to residual confounding (e.g., we lacked data on disease severity such as cancer stage and/or unmeasured confounding (e.g., physical activity, smoking, and alcohol abuse).

The strength of present study lies in its population-based design. As well, earlier studies found high positive predictive values of diagnoses in the DNPR of MI (~97%), ischemic stroke (~97%), and VTE (~88%), and somewhat lower positive predictive values for hemorrhagic stroke (~65% to 75%).[17,26] Our study also has some limitations. First, we had no valid information on what time of day the sample was drawn, which is known to affect the level of testosterone.[27] Nonetheless, timing of testosterone blood sampling may be independent of subsequent testosterone level, suggesting nondifferential misclassification, which could have biased the results against the null. Data were almost entirely missing on free testosterone levels (99%) in the LABKA registry, as all analyses for this laboratory test were performed in another Danish region during 2000 to 2009, and thus the test results were not available in the LABKA registry. It is also likely that free testosterone levels as well as luteinizing hormone and follicle-stimulating hormone are rarely measured in the primary health care sector as part of the initial diagnostic work-up. Therefore, these results were not available at baseline, but may have been present at a later stage, for example after referral to a specialist outpatient hospital clinic. Testosterone is a non-routine laboratory test, and will only be performed in those with a suspected reason to draw it. Laboratory tests were not standardized, which we were unable to account for in the analyses. However, a study found overall good correlations several immunoassays and the liquid chromatography-tandem mass spectrometry method.[28] Thus this issue is likely to be of minor importance.

Our study may have implications for clinical practice and future research. First, the prevalence of conditions related to hypogonadism for example, testicular torsion, were low, suggesting that almost all of the hypogonadism observed in routine clinical care is age-and co-morbidity-related. Second, persons with low testosterone levels have a higher absolute risk of dying and a higher absolute risk of

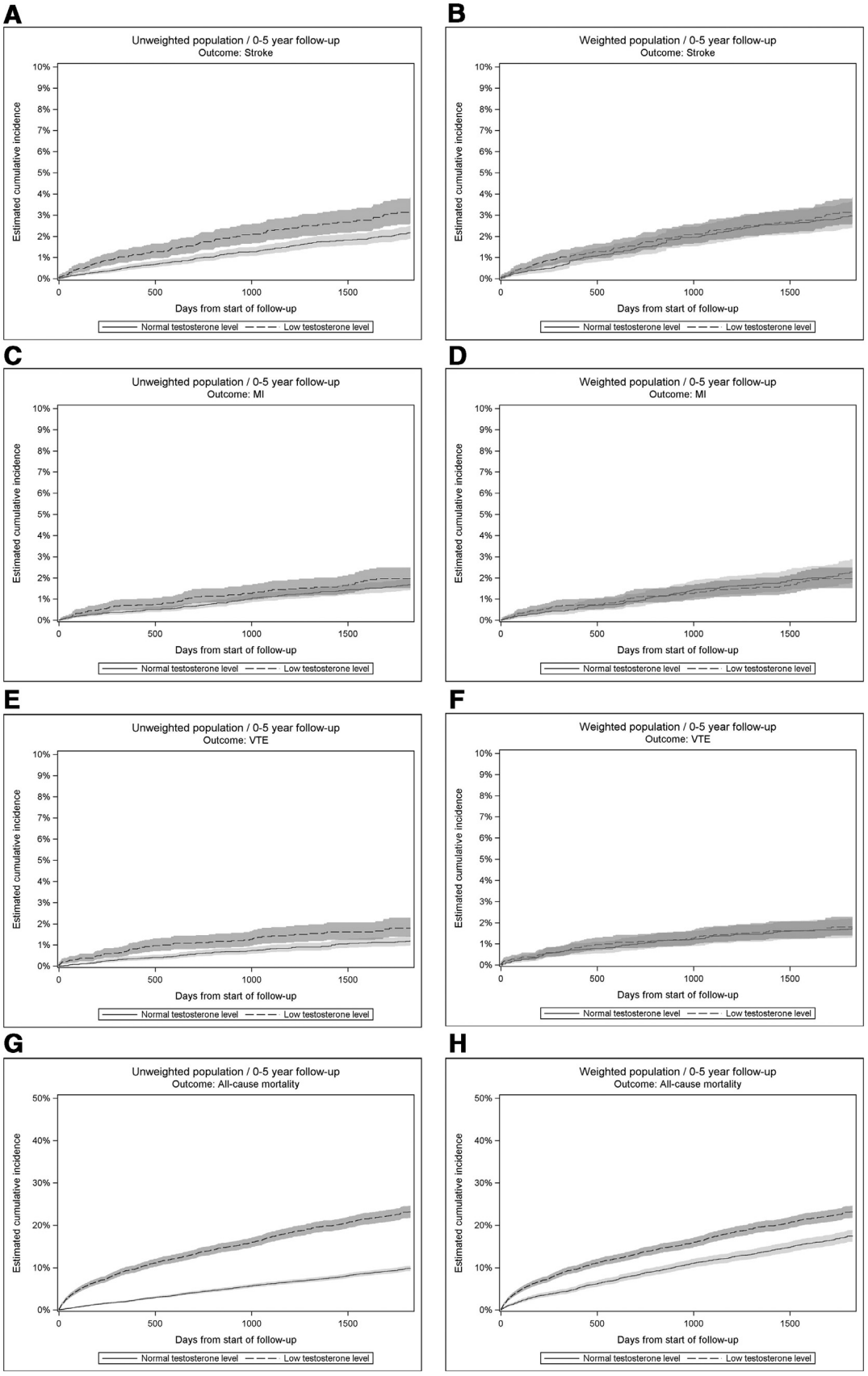


Figure 2. Cumulative incidence curves for stroke (*A-B*), myocardial infarction (MI) (*C-D*); venous thromboembolism (VTE) (*E-F*), and all-cause mortality (*G-H*) in men with low and normal testosterone levels, graphically illustrating risks in unweighted and weighted cohorts. The gray shaded areas represent 95% confidence intervals.

cardiovascular outcomes than individuals with normal testosterone levels. This suggests that the endogenous testosterone level is a potential marker of poor health, although our results do not suggest that low testosterone is an independent risk factor. Third, our study highlights the importance of taking into account confounding by low testosterone level in future pharmacoepidemiological studies on the safety and benefit of testosterone therapy.

In this cohort study, men with low total testosterone levels experienced more stroke, MI, VTE and had higher all-cause mortality than men with testosterone levels in the normal range. However, the associations between low testosterone levels and cardiovascular outcomes were mainly attributable to higher age and level of co-morbidity.

Author Contribution

All authors conceived the idea and designed the study. T. B.R performed the statistical analyses. All authors interpreted the data and reviewed the literature. K.A drafted the first manuscript. All authors critically reviewed the manuscript and approved the final version for submission. C.F.C has the overall responsibility of the accuracy of the data and the manuscript.

Ethics Approval

As this study did not involve patient contacts or any interventions, it was not necessary to obtain permission from the Danish Scientific Ethical Committee.

Disclosures

JBL is an employee of RTI International, an independent, non-profit research organization that performs contract work on behalf of governmental agencies and pharmaceutical companies. The remaining co-authors have no conflicts of interests.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.1016/j.amjcard.2019.02.042.

1. Onasanya O, Iyer G, Lucas E, Lin D, Singh S, Alexander GC. Association between exogenous testosterone and cardiovascular events: an overview of systematic reviews. *Lancet Diabetes Endocrinol* 2016;4:943–956.
2. Finkle WD, Greenland S, Ridgeway GK, Adams JL, Frasco MA, Cook MB, Fraumeni JF Jr, Hoover RN. Increased risk of non-fatal myocardial infarction following testosterone therapy prescription in men. *PLoS One* 2014;9:e85805.
3. Vigen R, O'Donnell CI, Baron AE, Grunwald GK, Maddox TM, Bradley SM, Barqawi A, Woning G, Wierman ME, Plomondon ME, Rumsfeld JS, Ho PM. Association of testosterone therapy with mortality, myocardial infarction, and stroke in men with low testosterone levels. *JAMA* 2013;310:1829–1836.
4. Xu L, Freeman G, Cowling BJ, Schooling CM. Testosterone therapy and cardiovascular events among men: a systematic review and meta-analysis of placebo-controlled randomized trials. *BMC Med* 2013;11:108.
5. Layton JB, Meier CR, Sharpless JL, Sturmer T, Jick SS, Brookhart MA. Comparative safety of testosterone dosage forms. *JAMA Intern Med* 2015;175:1187–1196.
6. U.S. Food and Drug Administration. FDA Drug Safety Communication: FDA cautions about using testosterone products for low testosterone due to aging; requires labeling change to inform of possible increased risk of heart attack and stroke with use, 2015. http://www.fda.gov/Drugs/DrugSafety/ucm436259.htm. (Accessed 2 September 2018)
7. Health Canada. 2014 Information update - possible cardiovascular problems associated with testosterone products. http://healthycanadians.gc.ca/recall-alert-rappel-avis/hc-sc/2014/40587a-eng.php (Accessed August 3, 2018)
8. Araujo AB, Dixon JM, Suarez EA, Murad MH, Guey LT, Wittert GA. Clinical review: Endogenous testosterone and mortality in men: a systematic review and meta-analysis. *J Clin Endocrinol Metab* 2011;96:3007–3019.
9. Haring R, Volzke H, Steveling A, Krebs A, Felix SB, Schofl C, Dorr M, Nauck M, Wallaschofski H. Low serum testosterone levels are associated with increased risk of mortality in a population-based cohort of men aged 20-79. *Eur Heart J* 2010;31:1494–1501.
10. Khaw KT, Dowsett M, Folkerd E, Bingham S, Wareham N, Luben R, Welch A, Day N. Endogenous testosterone and mortality due to all causes, cardiovascular disease, and cancer in men: European prospective investigation into cancer in Norfolk (EPIC-Norfolk) Prospective Population Study. *Circulation* 2007;116:2694–2701.
11. Ruige JB, Mahmoud AM, De Bacquer D, Kaufman JM. Endogenous testosterone and cardiovascular disease in healthy men: a meta-analysis. *Heart* 2011;97:870–875.
12. Shores MM, Matsumoto AM, Sloan KL, Kivlahan DR. Low serum testosterone and mortality in male veterans. *Arch Intern Med* 2006;166:1660–1665.
13. Yeap BB, Hyde Z, Almeida OP, Norman PE, Chubb SA, Jamrozik K, Flicker L, Hankey GJ. Lower testosterone levels predict incident stroke and transient ischemic attack in older men. *J Clin Endocrinol Metab* 2009;94:2353–2359.
14. Corona G, Rastrelli G, Di Pasquale G, Sforza A, Mannucci E, Maggi M. Endogenous testosterone levels and cardiovascular risk: meta-analysis of observational studies. *J Sex Med* 2018;15:1260–1271.
15. Danish Regions. Statistics. http://www.regioner.dk/om+regionerne/statistik+opdateret+dec+2014 (Accessed 1 June, 2018).
16. Grann AF, Erichsen R, Nielsen AG, Froslev T, Thomsen RW. Existing data sources for clinical epidemiology: the clinical laboratory information system (LABKA) research database at Aarhus University, Denmark. *Clin Epidemiol* 2011;3:133–138.
17. Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol* 2015;7:449–490.
18. Schmidt M, Pedersen L, Sorensen HT. The Danish Civil Registration System as a tool in epidemiology. *Eur J Epidemiol* 2014;29:541–549.
19. Johannesdottir SA, Horvath-Puho E, Ehrenstein V, Schmidt M, Pedersen L, Sorensen HT. Existing data sources for clinical epidemiology: the Danish National Database of Reimbursed Prescriptions. *Clin Epidemiol* 2012;4:303–313.
20. Ehrenstein V, Antonsen S, Pedersen L. Existing data sources for clinical epidemiology: Aarhus University Prescription Database. *Clin Epidemiol* 2010;2:273–279.
21. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods* 2004;9:403–425.
22. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* 2015;34:3661–3679.
23. Hak AE, Witteman JC, de Jong FH, Geerlings MI, Hofman A, Pols HA. Low levels of endogenous androgens increase the risk of atherosclerosis in elderly men: the Rotterdam study. *J Clin Endocrinol Metab* 2002;87:3632–3639.
24. Price JF, Lee AJ, Fowkes FG. Steroid sex hormones and peripheral arterial disease in the Edinburgh Artery Study. *Steroids* 1997;62:789–794.
25. Muller M, van den Beld AW, Bots ML, Grobbee DE, Lamberts SW, van der Schouw YT. Endogenous sex hormones and progression of carotid atherosclerosis in elderly men. *Circulation* 2004;109:2074–2079.

26. Sundboll J, Adelborg K, Munch T, Froslev T, Sorensen HT, Botker HE, Schmidt M. Positive predictive value of cardiovascular diagnoses in the Danish National Patient Registry: a validation study. *BMJ Open* 2016;6:e012832.
27. Brambilla DJ, Matsumoto AM, Araujo AB, McKinlay JB. The effect of diurnal variation on clinical measurement of serum testosterone and other sex hormone levels in men. *J Clin Endocrinol Metab* 2009;94:907–913.
28. Wang C, Catlin DH, Demers LM, Starcevic B, Swerdloff RS. Measurement of total serum testosterone in adult men: comparison of current laboratory methods versus liquid chromatography-tandem mass spectrometry. *J Clin Endocrinol Metab* 2004;89:534–543.

Supplemental Materials

Cardiovascular Outcomes and All-cause Mortality Associated with Endogenous Testosterone Levels

Kasper Adelborg, MD, PhD; Thomas Bøjer Rasmussen, MSc; Helene Nørrelund, MD, PhD, DMSc; J Bradley Layton, PhD; Henrik Toft Sørensen, MD, PhD, DMSc; Christian Fynbo Christiansen, MD, PhD

Contents

Supplementary Table 1. Age-specific normal references for laboratory measurements.

Total serum testosterone (nmol/L)	
13 – 20 years	1.0-39
>20 – 30 years	11-34
>30 – 40 years	10-33
>40 – 50 years	9.7-32
>50 – 60 years	9.3-32
>60 – 70 years	8.9-31
>70 – 80 years	8.6-31
>80 years	8.4-31
Follicle-stimulating hormone (IU/l)	
All ages	1.2-15.8
Luteinizing hormone (IU/l)	
All ages	1.7-8.6

Supplementary Table 2. Codes used in the study.

	NPU-code	ATC-code	ICD-8	ICD-10
Total testosterone	NPU03543, ASS00208	N/A	N/A	N/A
Free testosterone	NPU03549, NPU18879	N/A	N/A	N/A
Follicle-stimulating hormone	NPU02072, NPU04014, NPU21567	N/A	N/A	N/A
Luteinizing hormone	NPU02618, NPU04015	N/A	N/A	N/A
Albumin	NPU19673, ASS00224, DNK05449, NPU01132, AAA00774 14110, DNK05001	N/A	N/A	N/A
Comorbidities				
Myocardial infarction	N/A	N/A	410	I21-I23
Congestive heart failure	N/A	N/A	427.09, 427.10, 427.11, 427.19, 428.99, 782.49	I50, I11.0, I13.0, I13.2
Peripheral vascular disease	N/A	N/A	440, 441, 442, 443, 444, 445	I70, I71, I72, I73, I74, I77
Cerebrovascular disease	N/A	N/A	430-438	I60-I69, G45, G46
Dementia	N/A	N/A	290.09-290.19, 293.09	F00-F03, F05.1, G30
Chronic pulmonary disease	N/A	N/A	490-493, 515-518	J40-J47, J60-J67, J68.4, J70.1, J70.3, J84.1, J92.0, J96.1, J98.2, J98.3
Connective tissue disease	N/A	N/A	712, 716, 734, 446, 135.99	M05, M06, M08, M09, M30, M31, M32, M33, M34, M35, M36, D86
Ulcer disease	N/A	N/A	530.91, 530.98, 531-534	K22.1, K25-K28

Mild liver disease	N/A	N/A	XXXX	B18, K70.0-K70.3, K70.9, K71, K73, K74, K76.0
Diabetes without end-organ damage	N/A	N/A	249.00, 249.06, 249.07, 249.09, 250.00, 250.06, 250.07, 250.09	E10.0, E10.1, E10.9, E11.0, E11.1, E11.9
Hemiplegia	N/A	N/A	344	G81, G82
Moderate to severe renal disease	N/A	N/A	403, 404, 580-583, 584, 590.09, 593.19, 753.10-753.19, 792	I12, I13, N00-N05, N07, N11, N14, N17-N19, Q61
Diabetes with end-organ damage	N/A	N/A	249.01-249.05, 249.08, 250.01-250.05, 250.08	E10.2-E10.8, E11.2-E11.8
Moderate to severe liver disease	N/A	N/A	403, 404, 580-583, 584, 590.09, 593.19, 753.10-753.19, 792	I12, I13, N00-N05, N07, N11, N14, N17-N19, Q61
AIDS	N/A	N/A	079.83	B21-B24
Hypogonadism	N/A	N/A	25719	DE291, DE895
Hypopituitarism	N/A	N/A	25310, 25311,25318, 25319	DE230, DE231, DE893
Klinefelter's syndrome	N/A	N/A	75951	DQ980, DQ981, DQ982, DQ984
Down's syndrome	N/A	N/A	75939	DQ90
Testicular torsion	N/A	N/A	60770-60779	DN44
Varicocele	N/A	N/A	603	DI861
Cryptorchidism	N/A	N/A	75210-75211, 75219	DQ53
Orchitis	N/A	N/A	604	DN459
Chronic kidney disease	N/A	N/A	249.02, 250.02, 753.10-753.19, 582, 583, 584, 590.09, 593.20, 792	N03, N05-N08, N11.0, N14-N16; N18-N19, N26 -N27, N28.0, N39.1,Q61,E10.2, E11.2, E14.2, I12.0, I13.1, I13.2, I15.0, I15.1
Myxedema	N/A	N/A	244	DE00, DE03, DE890

Obesity	N/A	N/A	277	E66
Alcoholism	N/A	N/A	980, 291.09-291.99, 303.09-303.99, 57109-57111, 57710	F10 (except F10.0), G31.2, G62.1, G72.1, I 42.6, K29.2, K86.0, Z72.1 AND/OR ATC: N07BB
Hypertension	N/A	N/A	XXXX	DI10-DI15, I67.4
Any cancer (except prostate cancer)	N/A	N/A	140-209 (except 185)	C00-C85 (except C61), C88, C90, C91-C95, C96
Illicit drug abuse	N/A	N/A	971, 97090, 30459	F11-F16, F18-F19
Co-medications				
Testosterone	N/A	G03B	N/A	N/A
Antiandrogens	N/A	G03H	N/A	N/A
ACE/ARB	N/A	C09A, C09B, C09C, C09D	N/A	N/A
Beta blockers	N/A	C07	N/A	N/A
Statins	N/A	C10AA	N/A	N/A
Low-dose aspirin	N/A	B01AC06, N02BA01, N02BA51	N/A	N/A
Clopidogrel	N/A	B01AC04	N/A	N/A
Vitamin K antagonists	N/A	B01AA	N/A	N/A
Diuretics	N/A	MC03	N/A	N/A
NSAID	N/A	M01A	N/A	N/A
Opioids	N/A	N02A	N/A	N/A
Antidepressants	N/A	N06A	N/A	N/A
Antispychotics	N/A	N05A	N/A	N/A
Erectile dysfunction drugs	N/A	G04BE	N/A	N/A
Outcomes	N/A	N/A		
Stroke	N/A	N/A	433-434, 431	I61, I63-I64

Venous thromboembolism	N/A	N/A	450.99, 451.00	I80.1-I80.3, I26.0, I26.9
Myocardial infarction	N/A	N/A	410	I21-I23

Abbreviations: ACE/ARB: angiotensin-converting enzyme/angiotensin II receptor blockers; AIDS: acquired immune deficiency syndrome; ATC: Anatomical Therapeutic Chemical Classification; ICD: *International Classification of Diseases*; NPU: Nomenclature for Properties and Units; NSAIDs: nonsteroidal anti-inflammatory drugs.

;;

Supplementary Table 3. Risk of stroke in men with normal and low testosterone levels, by age groups.

	0-1 year of follow-up				0-5 years of follow-up			
Testosterone level	**No. at risk / No. of events**	**Incidence rate* (95% CI)**	**Unadjusted hazard ratio (95% CI)**	**Hazard ratio after IPTW weighting (95% CI)**	**No. at risk / No. of events**	**Incidence rate* (95% CI)**	**Unadjusted hazard ratio (95% CI)**	**Hazard ratio after IPTW weighting (95% CI)**
<55 years								
Normal	7,812 / 5	0.07 (0.03;0.17)	1.00 (ref)	1.00 (ref)	7,812 / 29	0.11 (0.07;0.16)	1.00 (ref)	1.00 (ref)
Low	2,329 / 8	0.38 (0.19;0.76)	5.48 (1.79;16.76)	4.33 (1.34;14.04)	2,329 / 14	0.18 (0.11;0.30)	1.65 (0.87;3.13)	1.59 (0.82;3.08)
55-64 years								
Normal	2,512 / 16	0.69 (0.42;1.13)	1.00 (ref)	1.00 (ref)	2,512 / 55	0.66 (0.51;0.86)	1.00 (ref)	1.00 (ref)
Low	788 / 7	1.01 (0.48;2.13)	1.46 (0.60;3.55)	1.19 (0.47;3.01)	788 / 23	0.95 (0.63;1.43)	1.43 (0.88;2.33)	1.16 (0.70;1.94)
65-74 years								
Normal	1,800 / 31	1.97 (1.38;2.80)	1.00 (ref)	1.00 (ref)	1,800 / 67	1.33 (1.05;1.69)	1.00 (ref)	1.00 (ref)
Low	674 / 14	2.50 (1.48;4.22)	1.26 (0.67;2.37)	1.11 (0.54;2.27)	674 / 37	2.11 (1.53;2.91)	1.57 (1.05;2.35)	1.34 (0.86;2.08)
+75 years								
Normal	910 / 13	1.66 (0.96;2.86)	1.00 (ref)	1.00 (ref)	910 / 48	2.08 (1.57;2.76)	1.00 (ref)	1.00 (ref)
Low	678 / 16	3.27 (2.00;5.34)	1.95 (0.95;4.01)	1.36 (0.60;3.05)	678 / 31	2.33 (1.64;3.31)	1.12 (0.71;1.74)	1.01 (0.60;1.69)

*Per 100 person-years

Abbreviations: CI: confidence interval; IPTW: inverse probability of treatment weighting

Supplementary Table 4. Risk of myocardial infarction in men with normal and low testosterone levels, by age groups.

	0-1 year of follow-up				0-5 years of follow-up			
Testosterone level	**No. at risk / No. of events**	**Incidence rate* (95% CI)**	**Unadjusted hazard ratio (95% CI)**	**Hazard ratio after IPTW weighting (95% CI)**	**No. at risk / No. of events**	**Incidence rate* (95% CI)**	**Unadjusted hazard ratio (95% CI)**	**Hazard ratio after IPTW weighting (95% CI)**
<55 years								
Normal	7,824 / 6	0.08 (0.04;0.18)	1.00 (ref)	1.00 (ref)	7,824 / 30	0.11 (0.08;0.16)	1.00 (ref)	1.00 (ref)
Low	2,335 / 2	0.09 (0.02;0.38)	1.13 (0.23;5.63)	0.89 (0.17;4.70)	2,335 / 15	0.19 (0.12;0.32)	1.25 (0.65;2.41)	1.32 (0.69;2.56)
55-64 years								
Normal	2,514 / 14	0.60 (0.36;1.02)	1.00 (ref)	1.00 (ref)	2,514 / 40	0.48 (0.35;0.65)	1.00 (ref)	1.00 (ref)
Low	793 / 5	0.72 (0.30;1.74)	1.17 (0.42;3.26)	0.96 (0.32;2.86)	793 / 9	0.37 (0.19;0.72)	0.60 (0.28;1.27)	0.71 (0.34;1.48)
65-74 years								
Normal	1,793 / 21	1.34 (0.87;2.05)	1.00 (ref)	1.00 (ref)	1,793 / 51	1.02 (0.78;1.35)	1.00 (ref)	1.00 (ref)
Low	670 / 6	1.07 (0.48;2.37)	0.80 (0.32;1.97)	0.75 (0.30;1.91)	670 / 14	0.78 (0.46;1.32)	0.61 (0.33;1.14)	0.62 (0.33;1.15)
+75 years								
Normal	895 / 9	1.17 (0.61;2.25)	1.00 (ref)	1.00 (ref)	895 / 33	1.45 (1.03;2.04)	1.00 (ref)	1.00 (ref)
Low	662 / 16	3.36 (2.06;5.48)	2.79 (1.23;6.35)	2.79 (1.17;6.65)	662 / 27	2.05 (1.41;2.99)	1.23 (0.70;2.17)	1.33 (0.77;2.29)

*Per 100 person-years

Abbreviations: CI: confidence interval; IPTW: inverse probability of treatment weighting

Supplementary Table 5. Risk of venous thromboembolism in men with normal and low testosterone levels, by age groups.

	0-1 year of follow-up				0-5 years of follow-up			
Testosterone level	**No. at risk / No. of events**	**Incidence rate* (95% CI)**	**Unadjusted hazard ratio (95% CI)**	**Hazard ratio after IPTW weighting (95% CI)**	**No. at risk / No. of events**	**Incidence rate* (95% CI)**	**Unadjusted hazard ratio (95% CI)**	**Hazard ratio after IPTW weighting (95% CI)**
<55 years								
Normal	7,825 / 7	0.10 (0.05;0.20)	1.00 (ref)	1.00 (ref)	7,825 / 30	0.11 (0.08;0.16)	1.00 (ref)	1.00 (ref)
Low	2,334 / 6	0.28 (0.13;0.63)	2.93 (0.99;8.70)	1.61 (0.47;5.45)	2,334 / 12	0.15 (0.09;0.27)	1.37 (0.70;2.67)	1.10 (0.54;2.27)
55-64 years								
Normal	2,563 / 11	0.46 (0.26;0.84)	1.00 (ref)	1.00 (ref)	2,563 / 31	0.36 (0.26;0.52)	1.00 (ref)	1.00 (ref)
Low	814 / 9	1.26 (0.66;2.43)	2.72 (1.13;6.54)	1.58 (0.61;4.07)	814 / 14	0.56 (0.33;0.94)	1.52 (0.81;2.86)	1.01 (0.51;2.02)
65-74 years								
Normal	1,883 / 9	0.54 (0.28;1.04)	1.00 (ref)	1.00 (ref)	1,883 / 25	0.47 (0.31;0.69)	1.00 (ref)	1.00 (ref)
Low	741 / 5	0.80 (0.33;1.93)	1.49 (0.50;4.43)	1.68 (0.55;5.15)	741 / 16	0.81 (0.50;1.32)	1.73 (0.92;3.24)	1.40 (0.69;2.86)
+75 years								
Normal	995 / 15	1.77 (1.07;2.93)	1.00 (ref)	1.00 (ref)	995 / 27	1.08 (0.74;1.57)	1.00 (ref)	1.00 (ref)
Low	767 / 13	2.34 (1.36;4.04)	1.30 (0.62;2.72)	0.87 (0.38;2.00)	767 / 23	1.52 (1.01;2.29)	1.35 (0.77;2.36)	1.10 (0.59;2.05)

*Per 100 person-years

Abbreviations: CI: confidence interval; IPTW: inverse probability of treatment weighting

Supplementary Table 6. Risk of all-cause mortality in men with normal and low testosterone levels by age groups.

	0-1 year of follow-up				0-5 years of follow-up			
Testosterone level	**No. at risk / No. of events**	**Incidence rate* (95% CI)**	**Unadjusted hazard ratio (95% CI)**	**Hazard ratio after IPTW weighting (95% CI)**	**No. at risk / No. of events**	**Incidence rate* (95% CI)**	**Unadjusted hazard ratio (95% CI)**	**Hazard ratio after IPTW weighting (95% CI)**
<55 years								
Normal	7,871 / 38	0.52 (0.38;0.72)	1.00 (ref)	1.00 (ref)	7,871 / 132	0.49 (0.41;0.58)	1.00 (ref)	1.00 (ref)
Low	2,360 / 34	1.59 (1.13;2.22)	3.04 (1.91;4.82)	2.18 (1.24;3.81)	2,360 / 83	1.04 (0.84;1.29)	2.14 (1.63;2.82)	1.55 (1.15;2.10)
55-64 years								
Normal	2,619 / 52	2.15 (1.64;2.82)	1.00 (ref)	1.00 (ref)	2,619 / 198	2.26 (1.96;2.60)	1.00 (ref)	1.00 (ref)
Low	841 / 59	8.00 (6.20;10.33)	3.71 (2.55;5.38)	2.20 (1.47;3.29)	841 / 135	5.21 (4.40;6.17)	2.30 (1.85;2.86)	1.60 (1.27;2.02)
65-74 years								
Normal	1,944 / 71	4.13 (3.28;5.22)	1.00 (ref)	1.00 (ref)	1,944 / 264	4.76 (4.22;5.37)	1.00 (ref)	1.00 (ref)
Low	773 / 93	14.30 (11.67;17.52)	3.44 (2.53;4.69)	2.01 (1.40;2.89)	773 / 212	10.19 (8.91;11.66)	2.14 (1.78;2.57)	1.61 (1.31;1.97)
+75 years								
Normal	1,033 / 102	11.47 (9.44;13.92)	1.00 (ref)	1.00 (ref)	1,033 / 325	12.30 (11.04;13.72)	1.00 (ref)	1.00 (ref)
Low	797 / 235	40.59 (35.72;46.13)	3.41 (2.71;4.30)	2.12 (1.61;2.79)	797 / 418	26.01 (23.63;28.63)	2.06 (1.78;2.38)	1.51 (1.28;1.77)

*Per 100 person-years

Abbreviations: CI: confidence interval; IPTW: inverse probability of treatment weighting

Comparative study of acute and mid-term complications with leadless and transvenous cardiac pacemakers

Daniel J. Cantillon, MD, FHRS,* Srinivas R. Dukkipati, MD, FHRS,† John H. Ip, MD,‡ Derek V. Exner, MD, FHRS,§ Imran K. Niazi, MD,‖ Rajesh S. Banker, MD,¶ Mayer Rashtian, MD, FHRS,# Kenneth Plunkitt, MD, FHRS,** Gery F. Tomassoni, MD, FHRS,†† Yelena Nabutovsky, MS,‡‡ Kevin J. Davis, BS,‡‡ Vivek Y. Reddy, MD†

*From the *Cleveland Clinic, Cleveland, Ohio, †Icahn School of Medicine at Mount Sinai, New York, New York, ‡Sparrow Clinical Research Institute, Lansing, Michigan, §Libin Cardiovascular Institute of Alberta, Calgary, Alberta, Canada, ‖Aurora Medical Group, Milwaukee, Wisconsin, ¶Premier Cardiology, Newport Beach, California, #Huntington Memorial Hospital, Pasadena, California, **Naples Community Hospital, Naples, Florida, ††Central Baptist Hospital, Lexington, Kentucky, and ‡‡Abbott, Sylmar, California.*

BACKGROUND Leadless cardiac pacemakers (LCPs) aim to mitigate lead- and pocket-related complications seen with transvenous pacemakers (TVPs).

OBJECTIVE The purpose of this study was to compare complications between the LCP cohort from the LEADLESS Pacemaker IDE Study (Leadless II) trial and a propensity score–matched real-world TVP cohort.

METHODS The multicenter LEADLESS II trial evaluated the safety and efficacy of the Nanostim LCP (Abbott, Abbott Park, IL) using structured follow-up, with serious adverse device effects independently adjudicated. TVP data were obtained from Truven Health MarketScan claims databases for patients implanted with single-chamber TVPs between April 1, 2010 and March 31, 2014 and more than 1 year of preimplant enrollment data. Comorbidities and complications were identified via *International Classification of Diseases, Ninth Revision* and Current Procedural Terminology codes. Short-term ($\leq$1 months) and mid-term (>1–18 months) complications were compared between the LCP cohort and a propensity score–matched subset of the TVP cohort.

RESULTS Among 718 patients with LCPs (mean age 75.6 $\pm$ 11.9 years; 62% men) and 1436 patients with TVPs (mean age 76.1 $\pm$ 12.3 years; 63% men), patients with LCPs experienced fewer complications (hazard ratio 0.44; 95% confidence interval 0.32–0.60; $P < .001$), including short-term (5.8% vs 9.4%; $P = .01$) and mid-term (0.56% vs 4.9%; $P < .001$) events. In the short-term time frame, patients with LCPs had more pericardial effusions (1.53% vs 0.35%; $P = .005$); similar rates of vascular events (1.11% vs 0.42%; $P = .085$), dislodgments (0.97% vs 1.39%; $P = .54$), and generator complications (0.70% vs 0.28%; $P = .17$); and no thoracic trauma compared to patients with TVPs (rate of thoracic trauma 3.27%). In short- and mid-term time frames, TVP events absent from the LCP group included lead-related, pocket-related, and infectious complications.

CONCLUSION Patients with LCPs experienced fewer overall short- and mid-term complications, including infectious and lead- and pocket-related events, but more pericardial effusions, which were uncommon but serious.

KEYWORDS Complications; Leadless; Comparative Study; Pacemakers; Transvenous

(Heart Rhythm 2018;15:1023–1030)

Introduction

Approximately 1 million transvenous pacemakers (TVP) are implanted annually worldwide.[1] Despite technological advances, the implantation technique involving a subcutaneous pulse generator and transvenous lead has remained unchanged and is the most common source of complications, occurring in up to 12% of device recipients.[2,3] Acute complications are related to implantation and include pneumothorax,

This study was funded by Abbott. **Address reprint requests and correspondence:** Dr Daniel J. Cantillon, Heart & Vascular Institute, Cleveland Clinic, 9500 Euclid Avenue J2-2, Cleveland, OH 44195. E-mail address: cantild@ccf.org.

1547-5271/$-see front matter https://doi.org/10.1016/j.hrthm.2018.04.022

hemothorax, cardiac perforation, pocket hematoma, and lead dislodgment.[4] Most long-term complications are associated with the pulse generator or lead and include pocket erosion, infection, lead fracture or insulation failure, tricuspid valve regurgitation, and venous thrombosis.[2,3,5–7]

Leadless cardiac pacemakers (LCPs) represent a new paradigm in cardiac pacing developed to mitigate complications by eliminating the need for a subcutaneous pocket and transvenous leads. These devices are small ($\sim$1 cm^3), entirely self-contained units that are delivered via a transfemoral venous catheter and affixed in the right ventricle using either an active (Nanostim, Abbott, Abbott Park, IL) or a passive (Micra, Medtronic, Minneapolis, MN) fixation mechanism.[8–13] The short-term safety and efficacy of these devices at 6 months have been established in nonrandomized comparisons to prespecified historical performance measures of TVPs.[8,9] Complications occurred in 4.0%–6.7% of patients, with cardiac perforation being the most common adverse event. While the quantity and type of complications were fewer and different from those reported with TVPs, comparison is limited by differences in patient comorbidities and study characteristics.

In this study, short-term and mid-term complications of the Nanostim LCP (Abbott, Abbott Park, IL) are compared with those of conventional single-chamber TVPs. The LCP safety data are obtained from the extended follow-up of the previously reported LEADLESS II IDE study.[8] Comparative safety data for TVPs are reported from a propensity score–matched cohort obtained from a large US real-world insurance claims database.

Methods

LCP study

The LEADLESS Pacemaker IDE Study (Leadless II) trial is a prospective, nonrandomized, multicenter clinical study conducted in the United States, Canada, and Australia. The trial design has been described in detail previously.[8] Patients with indications for permanent single-chamber ventricular pacing were implanted with a Nanostim LCP between February 1, 2014 and January 31, 2016. Full inclusion and exclusion criteria for the LEADLESS II trial are described in the Supplement. The LCP is a self-contained, active-fixation, rate-adaptive single-chamber pacemaker. The 42-mm-long, 5.99-mm-diameter device contains a helical screw-in fixation electrode at the distal end. A specially designed delivery catheter is used to percutaneously implant the LCP in the right ventricular apex or apical septum. Patients were evaluated before hospital discharge with device interrogation, chest radiography, and standard 12-lead electrocardiography. Subsequently, patients were followed at 2 weeks, 6 weeks, 3 months, 6 months, and every 6 months thereafter.

LCP safety data

All complications in the LEADLESS II trial were reported as part of the active clinical study follow-up and adhered to the International Standard Organization definition of a serious adverse device effect (SADE). A *SADE* is any untoward but not unanticipated medical occurrence that is related to the investigational device or procedure and that is classified as serious. A *"serious" event* is defined as any event that led to death or to a serious health deterioration that resulted in either a life-threatening illness or injury or a permanent impairment of a body structure or body function. It also includes events that led to an inpatient or prolonged hospitalization or medical or surgical intervention that was required to prevent the above-mentioned effects. All adverse events were adjudicated by an independent clinical events committee. SADEs were categorized into those related to cardiac perforation, vascular complications, device dislodgment, pacing threshold elevation, or other types of events. Complications were evaluated from implantation until 18 months or the time of withdrawal from the study, last available follow-up visit, or death.

TVP study

TVP data were extracted from the Truven Health MarketScan Research Databases, which contain more than 20 billion de-identified, person-specific health insurance claims from approximately 350 US private sector payers.[14] Data for this study were extracted from 2 MarketScan databases—the Commercial Claims and Encounters database and the Medicare Supplemental database—spanning the time period from April 1, 2010 to March 31, 2014. The Commercial Claims and Encounters database contains data from active employees, dependents, and early retirees covered by employer-sponsored health plans. The MarketScan database contains data from Medicare-eligible retirees with employer-sponsored Medicare Supplemental plans.

The study population included patients 18 years and older implanted with single-chamber pacemakers from any device manufacturer. Patients with pacemaker were identified as those having the *International Classification of Diseases, Ninth Revision* procedure code 37.81 (initial insertion of a single-chamber device, not specified as rate responsive) or 37.82 (initial insertion of a single-chamber device, rate responsive) or the Current Procedural Terminology code 33207 (insertion or replacement of a permanent pacemaker and lower-chamber electrodes). Patients with any implantable cardiac rhythm management device-related codes at any time before pacemaker implantation (Supplemental Table S1) were excluded from the analysis to eliminate non–de novo implants.

To characterize baseline comorbidities in the study population with TVPs, relevant inpatient and outpatient diagnostic and procedure codes were identified over the entire available time period before implantation. To ensure completeness of baseline data, patients with less than 1 year of MarketScan enrollment data were excluded from the analysis. Codes that indicated a history of atrial fibrillation, hypertension, diabetes, coronary artery disease, vascular disease, or tricuspid valve disease were included in the baseline characterization (comorbidity codes are listed in Supplemental Table S2).

TVP safety data

Pacemaker-related complications were identified for the TVP cohort using inpatient and outpatient billing codes recorded

from the day of implantation onward. Complications were compiled into the following categories (detailed in Supplemental Table S3): (1) infection, including endocarditis and other device-related infection; (2) thoracic trauma, including pneumothorax and hemothorax attributed to lead insertion; (3) pocket complication, including hematoma and pocket revision; (4) electrode dislodgment; (5) other lead complication requiring revision; (6) venous embolism or thrombosis; (7) cardiac perforation and its downstream clinical manifestations; and (8) generator complications. Generator explants were considered generator complications since they occurred within 30 days for acute and within 18 months for mid-term time frames, which are earlier than expected longevity of these devices.

To avoid overestimating complication rates, multiple codes from the same complication category that occurred on the same or consecutive dates were counted as a single event. In cases in which a pacemaker implantation and a complication occurred on the same date, the implantation was assumed to have preceded the complication. Thoracic trauma, cardiac perforation, and venous embolism/thrombosis occurring more than 1 month after implantation were not included in the TVP mid-term complication analysis, as they could not be definitively attributed to the pacemaker implant beyond the first month of implantation. Complications were evaluated from implantation until 18 months or the time of device upgrade, removal, or replacement or the patients' withdrawal from MarketScan.

Safety data comparison

Both LCP and TVP complications were classified as short- or mid-term relative to device implantation. Short-term complications occurred within 1 month, while mid-term complications occurred between 1 and 18 months after pacemaker implantation. In order to compare complication rates between LCP and TVP groups, a subset of patients with TVPs with similar baseline comorbidities to patients with LCPs was identified. Patients with TVPs were 2:1 propensity score matched to patients with LCPs using the nearest-neighbor method without replacement. The 2:1 ratio was the highest ratio for which resulting groups were well-matched on all baseline parameters. Propensity scores were computed on the basis of age, sex, and relevant baseline comorbidities including atrial fibrillation, coronary artery disease, diabetes, hyperlipidemia, hypertension, tricuspid valve disease, and peripheral vascular disease. The overall freedom from complications was evaluated in the matched cohort. In addition, the proportion of patients experiencing each prespecified short-term complication type was compared between patients with LCPs and patients with TVPs. In the mid-term time frame, rates of complications per patient-year were compared between the groups.

Statistical analysis

Continuous variables were compared using the Student *t* test. Categorical variables were compared using the χ^2 test. Complication rates were quantified by the number of patients with pacemaker who experienced at least 1 instance of a particular complication. Percentages were calculated relative to the total number of patients with pacemaker available within each time frame. Proportions were compared using the Fisher exact test, and event rates were compared using Poisson regression. Freedom from complications was computed using the Kaplan-Meier method and compared between patients with TVPs and patients with LCPs using the weighted Cox proportional hazards regression, adjusted for age, sex, and baseline comorbidities. All calculations were performed in R version 3.1.1, augmented with the following R packages: survival,[15] MatchIt,[16] and coxphw.[17]

Results

LCP cohort

The baseline clinical characteristics of the patient cohort enrolled in the multicenter LEADLESS II trial (n = 718) between February 2014 and January 2016 with a minimum follow-up of 180 days and a median follow-up of 323 days (interquartile range 197–489 days) are listed in Table 1. Single-chamber pacemaker indications in the LCP cohort were atrial fibrillation with atrioventricular block (n = 407 [56.7%]), sinus rhythm with high-grade atrioventricular block (n = 61 [8.5%]), and sinus bradycardia with infrequent pauses or syncope (n = 250 [34.8%]). *Acute implantation success*, defined as the patient leaving the implant procedure

Table 1 Baseline demographic characteristics of propensity score–matched patients

Characteristic	Patients with leadless pacemaker (n = 718)	Patients with transvenous pacemaker (n = 1436)	*P*
Age (y)	75.6 ± 11.9	76.1 ± 12.3	.39
Follow-up (d)	323 (197–489)	408 (167–547)	<.001
Sex: male	447 (62.3%)	905 (63.0%)	.77
Atrial fibrillation	425 (59.2%)	881 (61.4%)	.36
Coronary artery disease	262 (36.5%)	485 (33.8%)	.23
Diabetes mellitus	178 (24.8%)	335 (23.3%)	.49
Hyperlipidemia	475 (66.2%)	970 (67.5%)	.55
Hypertension	557 (77.6%)	1146 (79.8%)	.25
Tricuspid valve disease	150 (20.9%)	266 (18.5%)	.21
Peripheral vascular disease	91 (12.7%)	163 (11.4%)	.41

Values are presented as mean ± SD, as median (interquartile range), or as n (%).

with an implanted and functioning device, was achieved in 692 patients (96.4%) with the mean implantation time of 27.5 ± 17.0 minutes, which included 13.4 ± 9.3 minutes of fluoroscopy. Most of the failed implants were due to inability to deliver the LCP to the desired location in the right ventricle. Five of the failed implantations were due to pericardial effusion with tamponade and 1 (0.14%) due to pericardial effusion without tamponade. These events are included in the overall complication analysis. The pacemaker required repositioning more than 2 times in 25 patients (3.6%), and the mean hospital stay was 1.1 ± 1.0 days.

Short-term LCP complications occurred in 42 patients (5.8%), including 7 dislodgments (0.97%) requiring percutaneous retrieval, vascular-related events in 8 patients (1.11%), pericardial effusion with tamponade in 7 patients (0.97%), and pericardial effusion without tamponade in 4 patients (0.56%). Of the 8 vascular events, 2 (0.28%) required surgery; and of the 7 pericardial effusion with tamponade events, 3 (0.42%) required surgery. Pacing threshold elevation requiring percutaneous retrieval occurred in 5 patients (0.70%) within 1 month of implantation and in 1 patient (0.14%) after the first month. Overall, there were only 4 patients (0.56%) with a complication beyond 1 month. There were no reported dislodgments beyond 1 month. There were no infections in this patient cohort at short- and mid-term follow-up. The overall freedom from SADEs was 95.7% at implantation (95% confidence interval [CI] 94.2%–97.2%), 94.1% at 1 month (95% CI 92.4%–95.9%), and then remained at 93.5% (95% CI 91.7%–95.3%) starting at 100 days onward up to 18 months.

TVP cohort

The MarketScan database query yielded 120,556 patients with pacemaker, from whom we excluded 33,126 (27.4%) patients with less than 1 year of preimplant clinical data, 7442 (6.17%) patients with indeterminate pacemaker type, 7174 (5.95%) patients with evidence of preexisting devices, 113 (0.094%) patients less than 18 years of age, and 63,325 (52.5%) patients with dual-chamber devices. Ultimately, 9376 (7.78%) patients with single-chamber pacemaker (5323 [56.8%] men; mean age 80.4 ± 9.6 years; median follow-up 393 days [interquartile range 166–547 days]) were included in the unmatched analysis. The unmatched transvenous cohort was older and had fewer men and higher incidence of comorbidities including atrial fibrillation, coronary artery disease, diabetes, hyperlipidemia, hypertension, tricuspid valve disease, and peripheral vascular disease. A summary of complications in the unmatched TVP cohort and a comparison with LCPs are presented in Supplemental Figure S1.

Propensity score–matched analysis

After applying 1:2 propensity score matching to the 9376 patients with TVPs from the unmatched analysis, the 718 patients with LCPs were matched with 1435 (15.3% of unmatched cohort) patients with TVPs with clinical characteristics as listed in Table 1. As shown in Figure 1, there were

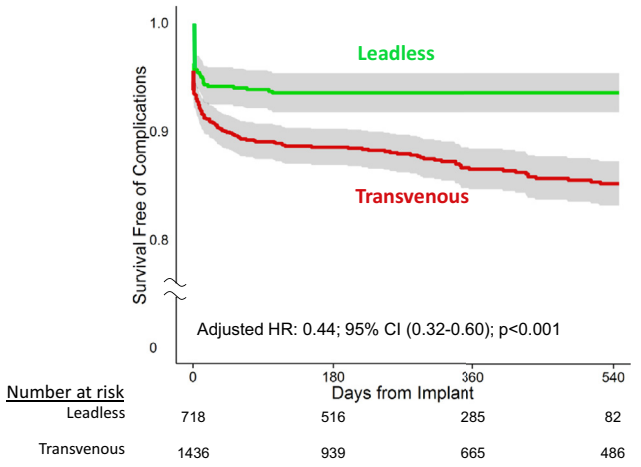


Figure 1 Kaplan-Meier curve (with 95% CI) illustrates that patients with LCPs were at a lower risk of experiencing a complication than were patients with TVPs. The Cox proportional hazards result is adjusted for age, sex, and baseline comorbidities. The starting point for the curves is the implantation of the device for both the LCP and TVP cohorts. CI = confidence interval; HR = hazard ratio; LCP = leadless pacemaker; TVP = transvenous pacemaker.

fewer overall complications in the leadless group when compared with the propensity score–matched transvenous group (adjusted hazard ratio 0.44; 95% CI 0.32–0.60; $P < .001$). This reduction persisted in all demographic and comorbidity subgroups (Figure 2).

Short-term complications were greatly reduced in the LCP cohort (42 [5.8%] vs 165 [9.4%]; $P = .0095$) despite a higher rate of pericardial effusions (11 [1.53%] vs 5 [0.35%]; $P = .0056$) in the leadless group. Of the 5 patients with TVPs with pericardial effusions, 4 (0.28%) were identified with the code for cardiac tamponade (423.3). There were no statistical differences between the leadless and transvenous groups with regard to rates of vascular complications (8 [1.11%] vs 6 [0.42%]; $P = .085$), electrode dislodgment (7 [0.97%] vs 20 [1.39%]; $P = .54$), and generator complications (5 [0.70%] vs 4 [0.28%]; $P = .17$). In the leadless group, there was a complete absence of lead-related complications, infections, and pocket complications, which were seen in 52 (3.62%), 25 (1.74%) and 6 (0.42%) TVP patients, respectively (Figure 3). In the LCP group, there was a single case of hemothorax associated with a perforation and cardiopulmonary resuscitation performed during the procedure, while in the TVP group, there were 47 (3.27%) occurrences of thoracic trauma. There were several complications in patients with LCPs that could not have been quantified in patients with TVPs because of limitations of insurance claims data. These included 5 instances of arrhythmia during implantation (0.70%), 2 acute migrations during implantation (0.28%), 1 angina event (0.14%), and 3 transient neurological events (0.42%). Of the 25 patients in the TVP cohort experiencing infection, 20 (1.39%) used an insurance code indicative of endocarditis while the other 5 (0.35%) used only the 996.61 code (infection and inflammatory reaction due to cardiac device, implant, and graft). Of the 6 patients experiencing pocket complications, none had an infection.

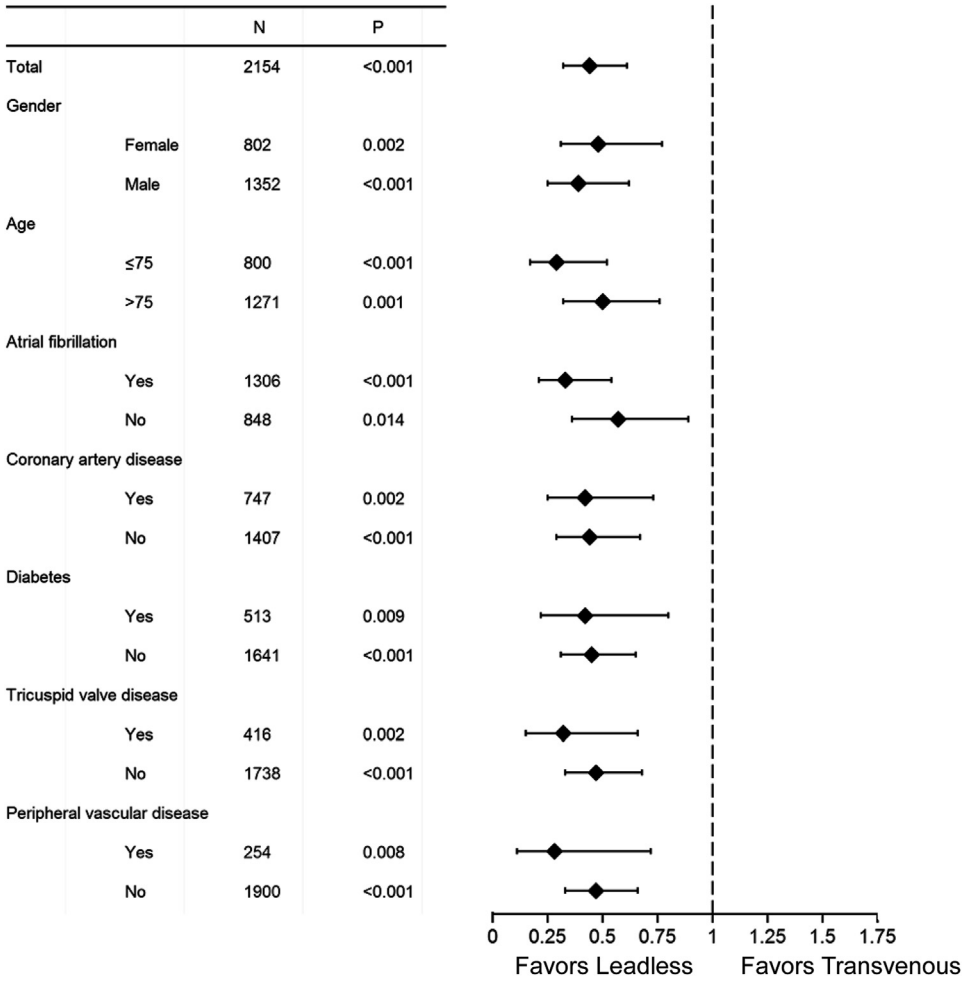


Figure 2 Plot presents adjusted hazard ratios and 95% confidence intervals for the risk of complication with the leadless pacemaker vs transvenous pacemaker in various patient subgroups.

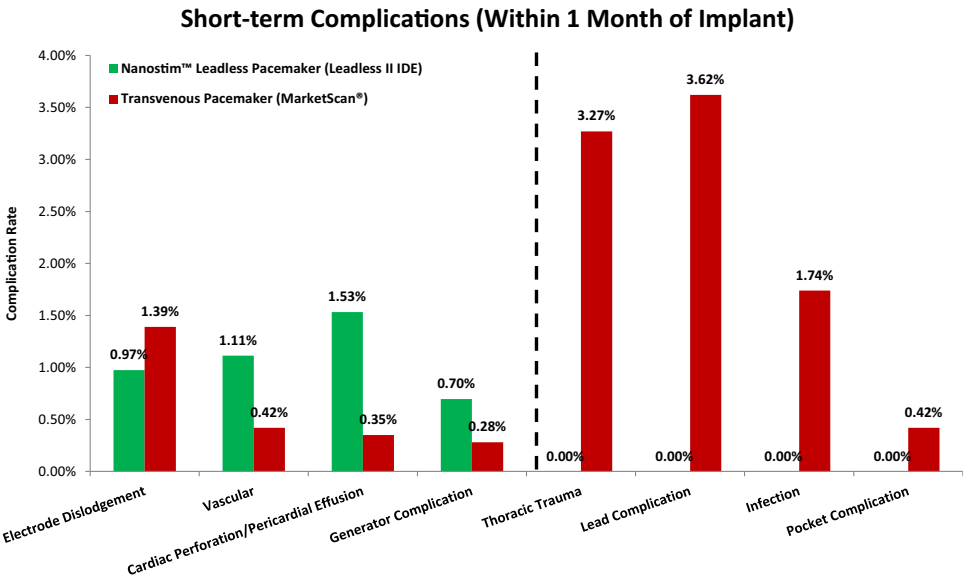


Figure 3 Short-term complication rates presented per category for patients with leadless pacemaker and patients with transvenous pacemaker. The exact rate is shown at the top of each bar.

Beyond 1 month, there were only 4 patients (0.56%) experiencing 4 complications in the leadless group (0.62 per 100 patient-years) vs 71 patients (4.94%) experiencing 127 complications in the TVP group (9.12 per 100 patient-years) ($P < .001$). In the leadless group, the mid-term complications included 1 instance of threshold elevation requiring revision (0.16 per 100 patient-years) and 1 temporary loss in pacing and sensing during ablation (0.16 per 100 patient-years) as compared to 5 (0.36 per 100 patient-years) generator complications in the TVP cohort. The leadless group also experienced 2 instances of new-onset heart failure (0.31 per 100 patient-years). In the transvenous group, there were a number of complications that were wholly absent in the leadless group, including lead-related complications (n = 36; 2.59 per 100 patient-years), electrode dislodgment (n = 4; 0.29 per 100 patient-years), infection (n = 66; 4.74 per 100 patient-years), and pocket complications (n = 16; 1.15 per 100 patient-years) (Figure 4). Most of the infectious complication encounters contained a code indicative of endocarditis, while only 10 (0.72 per 100 patient-years) contained only the 996.61 code. Of the 16 patients with pocket complications, 4 (0.29 per 100 patient-years) patients also had an infection, with only 2 (0.14 per 100 patient-years) of these infections occurring during the same hospital stay as the pocket complication.

Discussion

The principal finding of this analysis is that patients from the LEADLESS II IDE trial demonstrated fewer short- and mid-term complications when compared with a large propensity score–matched cohort of patients with single-chamber TVPs. The overall reduction in both short- and mid-term events was driven by a virtual elimination of lead, pocket, and infectious complications, suggesting that this disruptive technology has successfully targeted the most common sources of traditional pacemaker complications observed over the past 50 years. The TVP complications in this study are consistent with an extensive body of literature, showing that lead-related problems, thoracic trauma, vascular injury, pocket hematoma, and infection drive short-term complications and that lead-related problems dominate the mid-term complications.[2–7] The latter relate to electrical phenomena involving sensing, pacing, or insulation failures. These findings reinforce the use of a leadless pacemaker as an alternative to TVPs in patients requiring single-chamber ventricular pacing.

Both the short-term and mid-term TVP complication rates of 9.40% and 4.94% reported in our study exceeded those reported in The Mode Selection Trial (MOST) (4.8% at 30 days and 2.1% at 3 years)[18] and were slightly lower than those reported in the FOLLOWPACE study (12.4% at 2 months and 9.2% by 5 years).[3] Both MOST and FOLLOWPACE studies investigated dual-chamber devices, which are expected to have more complications than do single-chamber devices.[2,3,19,20] In addition, the FOLLOWPACE study did not exclude non–de novo systems while our study focused only on new implants. Similar to the FOLLOWPACE study and in contrast to the MOST trial, claims data capture complications occurring across the full spectrum of operators performing pacemaker surgery at community and urban hospitals and are not limited to the academic or tertiary medical centers with highly experienced operators. Furthermore, the MOST trial was performed between 1995 and 2001 and the FOLLOWPACE study between 2003 and 2007 while our patients were implanted between 2010 and 2014. A report from a large national survey demonstrates that the population receiving pacemakers has greatly expanded, and has become older and sicker,[21] which could lead to higher rates of complications.

The categories in which TVPs fared better had slightly lower rates of uncommon but potentially serious complications of pericardial effusion and vascular events. It should

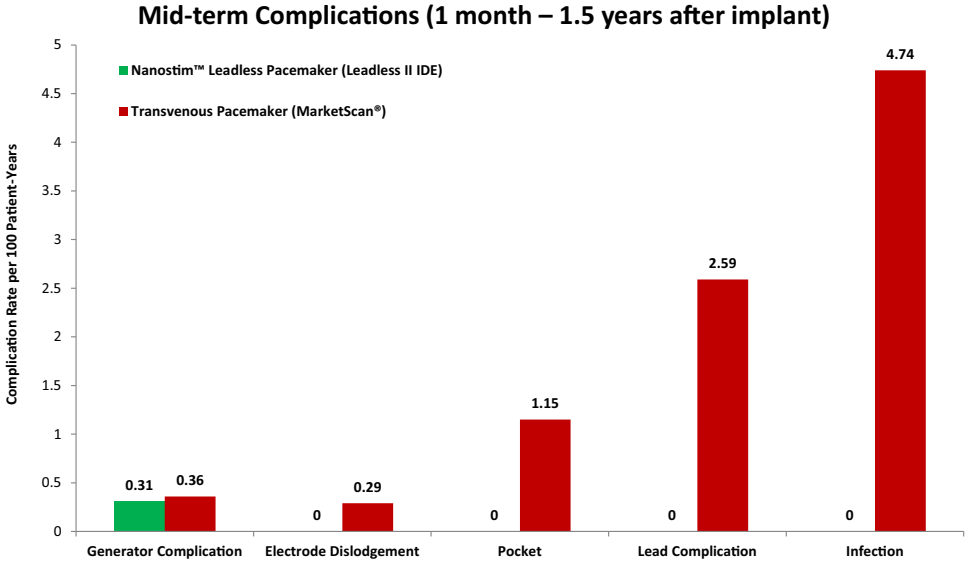


Figure 4 Mid-term complication rates presented per category for patients with leadless pacemaker and patients with transvenous pacemaker. The exact rate is shown at the top of each bar. One of the reported cardiac perforation complications also had an associated hemothorax as a result of a cardiopulmonary resuscitation performed during the procedure.

be noted that the introduction of femoral vascular complications with LCPs represents a true trade-off created by the paradigm shift away from pectoral surgical incisional access to percutaneous femoral vascular access; the introduction of an 18-F vascular delivery sheath provides challenges to achieving hemostasis after use of femoral instrumentation. However, the elimination of pocket-related and infection-related acute complications arguably more than compensates for the small increases in rates of vascular events in LCPs. The possibility of cardiac perforation and pericardial effusion exists with both technologies since decades of innovation in lead design and fixation mechanisms have not eliminated this problem even with transvenous leads.[22] The 0.35% pericardial effusion rate in this transvenous group is similar to other published data involving transvenous leads (0.3%–0.8%) and is lower than the 1.5% event rate in the leadless cohort.[3,18,22,23] It is concerning that 3 of the 7 patients with pericardial effusion in the LCP group required surgery. This suggests more traumatic tearing-type injuries that need to be mitigated by future iterations of LCP technology, as well as improvements in operator technique. Previous studies have associated acute pacemaker complications with operator experience and training.[2,4] Encouragingly, the original LEADLESS II study investigators reported a reduction in complications from 6.8% to 3.6% after 10 operator implants.[8] While the future performance of subsequent iterations of leadless delivery systems is unknowable, it is expected that design changes incorporating operator feedback as well as greater experience will improve acute implantation safety. It is possible that LCP delivery systems will always remain stiffer and more traumatic to cardiac tissue than will transvenous leads because of the support needed to introduce and steer the catheter-based device. Even in this scenario, it would be premature to equate small absolute differences in pericardial effusions to a net clinical benefit of avoiding complications associated with transvenous systems. Transvenous lead extractions carry significant risk in the event of vascular or cardiac tears.[24] Some of these complications may arise with LCPs if there is a need for extraction. No incidents of tricuspid valve injury occurred during the placement of LCPs in the trial, but there could be such incidents associated with LCP extraction. However, fully eliminating lead, pacemaker pocket, and infectious complications beyond the acute period will obviate the need for at least some of these procedures and extend some degree of still unknown benefit in avoiding procedure-related catastrophes. Finally, this field is simply too young to judge and compare the long-term implications that remain, as of yet, unknown; indeed, the end-of-service clinical experience of the leadless device will not be fully understood for another 10–15 years.

Development of a new technology can be accompanied by unexpected challenges. Field safety advisories were issued for the Nanostim leadless pacemaker due to battery malfunction and docking button detachments. The replacement battery for the Nanostim LCP has been approved by several regulatory agencies and the next generation LCP will include an updated docking button design.

Study limitations

Limitations of the present analysis include limitations of the MarketScan databases, which do not contain a random sample of patient claims data, but rather a cohort that is primarily drawn from large employers. Patients who are self-insured and those insured through small and medium employers are underrepresented, and those covered by Medicare Advantage and traditional Medicare plans are excluded. To avoid overestimating complication rates in the transvenous cohort, multiple diagnostic and procedure codes observed on the same or consecutive service dates were treated as a single occurrence; however, these could only have resulted in repeat occurrences to be undercounted in some scenarios. Similarly, single complications with encounters on nonconsecutive service dates could be overcounted. Furthermore, it was not possible to definitively associate every complication with the pacemaker implant. Some complications may have wrongly been attributed to pacemaker implants, and others may not have been identified if unanticipated claims codes were used. Finally, the severity of a complication could not be ascertained, as there is not a systematic way to identify cases requiring surgical management.

Another limitation of the analysis is lack of specific data on atrial fibrillation in the TVP cohort. Insurance claims do not distinguish between AF of different severities. Therefore, the distribution of various severities of AF may not have been the same between the matched TVP and LCP groups.

Limitations related to comparison of the 2 data sets include the differing definitions of complications and different types of complications that can occur with the different pacemaker systems. The LEADLESS II study included complications deemed serious by an independent committee. The TVP complications were not adjudicated and could have included both more and less severe events. One can only be sure that patients experiencing these complications had active encounters with the medical system, and the encounters resulted in the filing of insurance claims. Furthermore, since the LEADLESS II study was a clinical trial, it may have had more experienced implanters operating at academic centers and research hospitals as compared with TVP implanters from a full spectrum of US hospitals and with varying degrees of experience. Despite these limitations, the magnitude of the difference between complications in the LCP and TVP groups suggests that future studies will confirm the advantage of leadless technology.

Conclusion

This propensity score–matched analysis of leadless pacemakers from the LEADLESS II IDE study and TVPs from the MarketScan claims database suggests that leadless

pacemakers are associated with significant reduction in overall short- and mid-term complications, particularly among infectious, pocket-related, and lead-related events, but can be accompanied by more pericardial effusions, which are uncommon but may be serious enough to require surgery. Additional data about the long-term risk and complication profile of these devices are needed.

Appendix

Supplementary data

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.hrthm.2018.04.022.

References

1. Mond HG, Proclemer A. The 11th world survey of cardiac pacing and implantable cardioverter-defibrillators: calendar year 2009—a World Society of Arrhythmia's project. Pacing Clin Electrophysiol 2011;34:1013–1027.
2. Kirkfeldt RE, Johansen JB, Nohr EA, Jorgensen OD, Nielsen JC. Complications after cardiac implantable electronic device implantations: an analysis of a complete, nationwide cohort in Denmark. Eur Heart J 2014;35:1186–1194.
3. Udo EO, Zuithoff NP, van Hemel NM, de Cock CC, Hendriks T, Doevendans PA, Moons KG. Incidence and predictors of short- and long-term complications in pacemaker therapy: the FOLLOWPACE study. Heart Rhythm 2012;9:728–735.
4. Tobin K, Stewart J, Westveer D, Frumin H. Acute complications of permanent pacemaker implantation: their financial implication and relation to volume and operator experience. Am J Cardiol 2000;85:774–776, A9.
5. Al-Mohaissen MA, Chan KL. Prevalence and mechanism of tricuspid regurgitation following implantation of endocardial leads for pacemaker or cardioverter-defibrillator. J Am Soc Echocardiogr 2012;25:245–252.
6. Hauser RG, Hayes DL, Kallinen LM, Cannom DS, Epstein AE, Almquist AK, Song SL, Tyers GF, Vlay SC, Irwin M. Clinical experience with pacemaker pulse generators and transvenous leads: an 8-year prospective multicenter study. Heart Rhythm 2007;4:154–160.
7. Johansen JB, Jorgensen OD, Moller M, Arnsbo P, Mortensen PT, Nielsen JC. Infection after pacemaker implantation: infection rates and risk factors associated with infection in a population-based cohort study of 46299 consecutive patients. Eur Heart J 2011;32:991–998.
8. Reddy VY, Exner DV, Cantillon DJ, et al. Percutaneous implantation of an entirely intracardiac leadless pacemaker. N Engl J Med 2015;373:1125–1135.
9. Reynolds D, Duray GZ, Omar R, et al. A Leadless intracardiac transcatheter pacing system. N Engl J Med 2016;374:533–541.
10. Reddy VY, Knops RE, Sperzel J, et al. Permanent leadless cardiac pacing: results of the LEADLESS trial. Circulation 2014;129:1466–1471.
11. Knops RE, Tjong FV, Neuzil P, et al. Chronic performance of a leadless cardiac pacemaker: 1-year follow-up of the LEADLESS trial. J Am Coll Cardiol 2015;65:1497–1504.
12. Ritter P, Duray GZ, Steinwender C, et al. Early performance of a miniaturized leadless cardiac pacemaker: the Micra Transcatheter Pacing Study. Eur Heart J 2015;36:2510–2519.
13. Miller MA, Neuzil P, Dukkipati SR, Reddy VY. Leadless cardiac pacemakers: back to the future. J Am Coll Cardiol 2015;66:1179–1189.
14. Truven Health MarketScan® Research Databases. Truven Health Analytics; 2018. Available at: https://truvenhealth.com/markets/life-sciences/products/data-tools/marketscan-databases. Accessed May 23, 2018.
15. Therneau T. A Package for Survival Analysis in S. R package version 2.41-3; 2018. Available at: https://cran.r-project.org/web/packages/survival/index.html. Accessed May 23, 2018.
16. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 2011;42.
17. Heinze G, Ploner M, Dunkler D. coxphw: weighted estimation in Cox regression. R package version 4.0.1; 2018. Available at: https://www.jstatsoft.org/article/view/v084i02/v84i02.pdf. Accessed May 23, 2018.
18. Ellenbogen KA, Hellkamp AS, Wilkoff BL, Camunas JL, Love JC, Hadjis TA, Lee KL, Lamas GA. Complications arising after implantation of DDD pacemakers: the MOST experience. Am J Cardiol 2003;92:740–741.
19. Wiegand UK, Bode F, Bonnemeier H, Eberhard F, Schlei M, Peters W. Long-term complication rates in ventricular, single lead VDD, and dual chamber pacing. Pacing Clin Electrophysiol 2003;26:1961–1969.
20. Chauhan A, Grace AA, Newell SA, Stone DL, Shapiro LM, Schofield PM, Petch MC. Early complications after dual chamber versus single chamber pacemaker implantation. Pacing Clin Electrophysiol 1994;17:2012–2015.
21. Greenspon AJ, Patel JD, Lau E, Ochoa JA, Frisch DR, Ho RT, Pavri BB, Kurtz SM. Trends in permanent pacemaker implantation in the United States from 1993 to 2009: increasing complexity of patients and procedures. J Am Coll Cardiol 2012;60:1540–1545.
22. Moazzami K, Dolmatova E, Mazza V, Klapholz M, Waller A. Trends in cardiac tamponade among recipients of permanent pacemakers in the United States: 2008 to 2012. J Am Coll Cardiol 2016;67:776.
23. Cano O, Andres A, Alonso P, Osca J, Sancho-Tello MJ, Olague J, Martinez-Dolz L. Incidence and predictors of clinically relevant cardiac perforation associated with systematic implantation of active-fixation pacing and defibrillation leads: a single-centre experience with over 3800 implanted leads. Europace 2017;19:96–102.
24. Buiten MS, van der Heijden AC, Schalij MJ, van Erven L. How adequate are the current methods of lead extraction? A review of the efficiency and safety of transvenous lead extraction methods. Europace 2015;17:689–700.

BMJ Global Health

The impact of a cash transfer programme on tuberculosis treatment success rate: a quasi-experimental study in Brazil

Daniel J Carter,[1,2] Rhian Daniel,[2] Ana W Torrens,[3] Mauro N Sanchez,[4] Ethel Leonor N Maciel,[5] Patricia Bartholomay,[6] Draurio C Barreira,[7] Davide Rasella,[8] Mauricio L Barreto,[9,10] Laura C Rodrigues,[1,10] Delia Boccia[1]

To cite: J Carter D, Daniel R, Torrens AW, *et al*. The impact of a cash transfer programme on tuberculosis treatment success rate: a quasi-experimental study in Brazil. *BMJ Glob Health* 2019;**4**:e001029. doi:10.1136/bmjgh-2018-001029

Handling editor Nicola Foster

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjgh-2018-001029).

Received 2 July 2018
Revised 4 October 2018
Accepted 6 November 2018

For numbered affiliations see end of article.

Correspondence to
Dr Delia Boccia;
delia.boccia@lshtm.ac.uk

ABSTRACT

Background Evidence suggests that social protection policies such as Brazil's Bolsa Família Programme (BFP), a governmental conditional cash transfer, may play a role in tuberculosis (TB) elimination. However, study limitations hamper conclusions. This paper uses a quasi-experimental approach to more rigorously evaluate the effect of BFP on TB treatment success rate.

Methods Propensity scores were estimated from a complete-case logistic regression using covariates from a linked data set, including the Brazil's TB notification system (SINAN), linked to the national registry of those in poverty (CadUnico) and the BFP payroll.

Results The average effect of treatment on the treated was estimated as the difference in TB treatment success rate between matched groups (ie, the control and exposed patients, n=2167). Patients with TB receiving BFP showed a treatment success rate of 10.58 percentage points higher (95% CI 4.39 to 16.77) than patients with TB not receiving BFP. This association was robust to sensitivity analyses.

Conclusions This study further confirms a positive relationship between the provision of conditional cash transfers and TB treatment success rate. Further research is needed to understand how to enhance access to social protection so to optimise public health impact.

Key questions

What is already known?

- While encouraging, evidence about the impact of cash transfers on tuberculosis (TB) control is still scattered and conclusions are often hampered by important study limitations.

What are the new findings?

- This is the first study using a quasi-experimental design to evaluate the impact of Bolsa Familia on TB treatment success.
- Patients with TB enrolled in Bolsa Familia are more likely to complete their treatment successfully.
- Approximately half of patients with TB included in this study population were not enrolled in the cash transfer programme despite being eligible based on the income inclusion criterion.

What do the new findings imply?

- Conditional cash transfers like Bolsa Familia can contribute to TB elimination even if they were not designed for this purpose.
- Disparity in access is a missed opportunity to maximise TB impact of Bolsa Familia.

INTRODUCTION

Despite biomedical efforts, the global burden of tuberculosis (TB) remains considerable, with up to 1.5 million deaths from TB recorded in 2015.[1] TB treatment takes many months, and a proportion of patients are not cured, either because they abandon treatment, take treatment irregularly, are infected with drug-resistant TB, or die before completion of treatment.[1] The correlation between TB indicators and global poverty has been demonstrated both at ecological and individual levels, yet much of the morbidity and mortality in patients with TB still occur among the poorest segments of the population.[2] Social determinants impact vulnerability to TB at every stage of the disease pathway, from TB infection to clinical outcomes, including whether or not the patient was successfully treated.[3] Ending the global burden of TB requires bold policies and supportive systems able to recognise and tackle these social determinants.[4]

Recognising this social aspect of TB epidemiology, social protection is now a non-negotiable component to reach the TB elimination targets set by the WHO, including zero households affected by catastrophic costs, defined as TB-related expenditures when they exceed 20% of preillness annual household income.[5]

Brazil in particular has been an early adopter of the WHO's End TB Strategy,[6] as reflected by its long-term efforts to integrate development and health agendas. This is partially due to the long social protection tradition in Latin America, which in Brazil culminated with the creation of the Bolsa Família Programme (BFP) in 2003, one of the largest conditional cash transfer programmes in the world.[7]

In 2010, the BFP provided a variable monthly stipend to households meeting certain socioeconomic criteria: households earning less than R$70 a month (~US$22 at time of writing) and households with children, adolescents or pregnant women earning less than R$140 a month. BFP's targeting is not exact, and individuals reporting an income above R$140 can be found in the BFP payroll.[7] In order to receive BFP, families must be registered in the Cadastro Unico (single registry; CadÚnico), a registry of all low-income Brazilian families. In return for the transfers, recipients must comply with behavioural obligations (ie, school attendance; immunisation). BFP is not explicitly intended to target TB-affected households and only one-fourth of patients with TB in Brazil appear to be enrolled in the programme; given the intimate association between poverty and TB, underenrolment is likely.[8] Despite accumulating, the literature on the impact of conditional cash transfers on a variety of TB indicators is still limited, and there has been little methodologically rigorous evaluation of social protection interventions for TB prevention, care and control, including treatment outcomes.[9] There has also been some support in the literature for financial incentives having a small positive effect on TB outcomes,[10] but the underlying philosophy, mechanisms of action, as well as the ethical and sustainability implications for financial incentives may differ from cash transfers embedded into proper governmental social protection platforms.[11]

Despite its scarcity, the evidence is converging on a consistent positive impact of social protection on TB epidemiology and control, including some small-scale trials and studies in Peru,[12] Moldova[13] and South Africa.[14–16]As for Brazil, the literature is even more rich even if evidence does not necessarily follow from proper controlled trials.[15–18] Torrens *et al*[8] have already attempted to estimate the impact of BFP on TB treatment success rates and found out that patients with TB enrolled in BFP were approximately 7% more likely to be successfully treated after treatment than a control group.[8] While the findings of this study are consistent with what observed in the literature, conclusions are hampered by the potential biased nature of the control group.[8]

For an unbiased estimate of the proportion of patients cured attributable to BFP, we must construct a control group as similar as possible to the group of BFP recipients. This group of BFP recipients on average have some TB treatment success rate. We wish to estimate the difference in that treatment success rate if, counter to fact, that group of patients had not received BFP, but had the same sociodemographic characteristics and were thus still enrolled in CadÚnico.

To this aim, we approach the same routine data source as in Torrens *et al*[8] using a quasi-experimental approach to construct a more appropriate control group and to then determine a more rigorous estimate of the effect of BFP on TB treatment success rate among those who receive it. Specifically, we aimed to: (1) use propensity score matching to create a control group balanced for propensity to receive BFP, (2) provide an estimate of the average treatment effect of BFP on TB treatment success rate among recipients and (3) to reflect on the utility of the resulting estimate for changing TB policy.

METHODS

Conceptual framework: directed acyclic graph

A directed acyclic graph (DAG) was proposed for conceiving of the causal relationships between the outcome, the exposure and all the variables hypothesised to be on the causal pathway (figure 1). Each node in the DAG consists of a high-level construct measured by proxy variables taken from the set of covariates available (table 1). The nodes in this DAG were constructed based on a variety of theoretical literature, and the grouping of covariates under one node denotes that they are considered to be measures of that underlying construct for the purposes of this paper.[3 19 20] Online supplementary appendix 1 outlines explicitly which covariates fall under each node.

The DAG outlines potential mechanisms by which BFP ('the exposure') is proposed to affect treatment success rate ('the outcome'). These include via access to directly observed treatment and via increased capacity for mitigation of catastrophic costs (expenditure). We provide an estimate for the direct effect of social protection outside of these pathways, which may include expanded access to healthcare through means other than Directly Observed Therapy (DOT), increased psychological well-being or greater integration into governmental systems in general. The DAG also outlines pathways between treatment success rate and income (and therefore access to BFP), through complex relationships between demographics, geography and socioeconomic factors. The 'treatment success' outcome includes those who completed treatment with or without bacteriological confirmation.

Data handling

The data for this study arose from a linkage between the 2010 TB data set from SINAN (Brazil's national Notifiable Disease Surveillance System) and the 2011 CadÚnico data set. The CadÚnico data set was itself linked to the Bolsa Familia payroll held by the Caixa Federal (Federal Bank). The linkage added the demographic and social information from CadÚnico and the BFP payroll to every patient with TB in the SINAN data set.

Of the complete SINAN-CadÚnico-BFP data set (n=180 046), only individuals who were new TB cases registered in

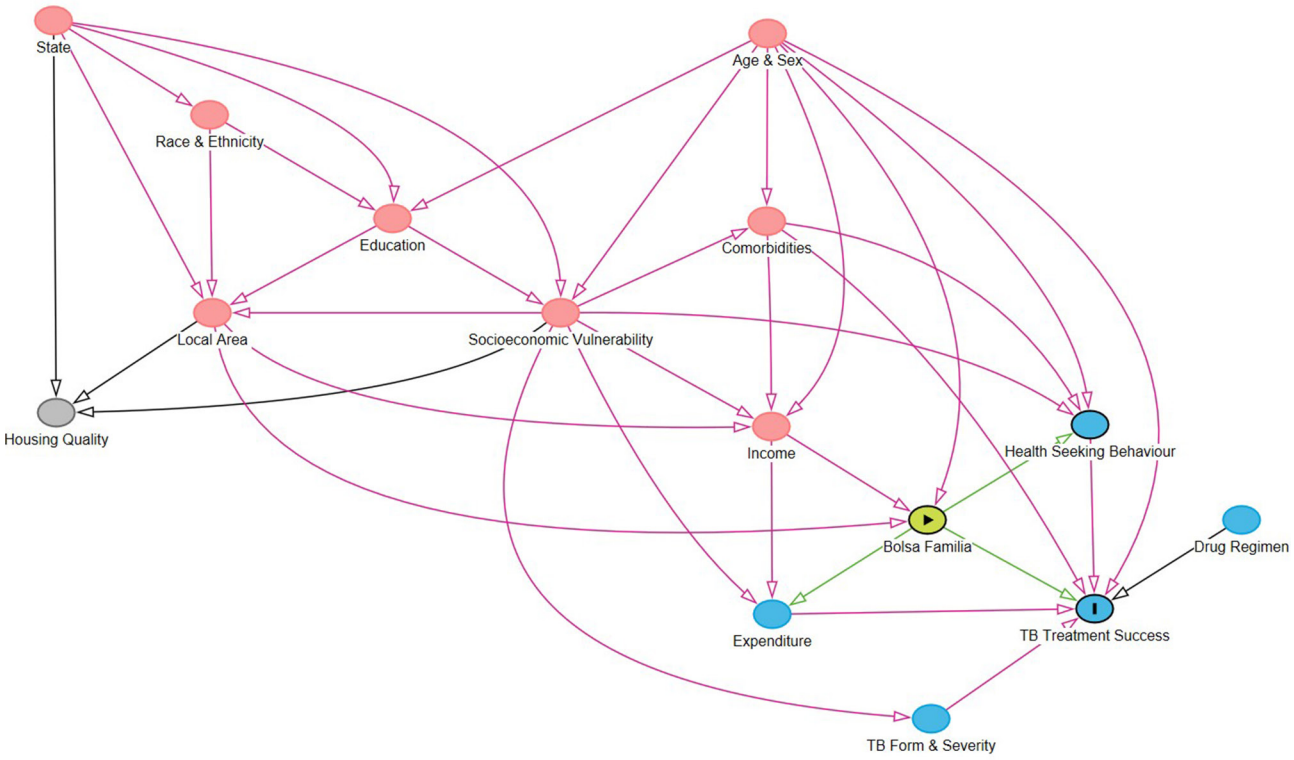


Figure 1 Directed acyclic graph (DAG) outlining the pathways linking Bolsa Familia with tuberculosis (TB) outcomes. A DAG was built to conceptualise the potentially causal relationships between constructs relevant for measuring the impact of Bolsa Familia on TB treatment success rate. Red nodes are ancestors of both the outcome and the exposure (ie, confounders) while grey nodes are unassociated with the outcome and exposure. Blue nodes are ancestors of the outcome. The DAG links nodes that represent constructs that are measured by covariates table 2).

CadUnico in 2010 with a non-missing treatment outcome variable were retained for this study (n=16 760). Exposed individuals (defined here as those receiving BFP) were further restricted to those whose receipt of BFP preceded case closure. Case closure is defined as the date on which an outcome (eg, treated, unsuccessful completion of treatment, death) is recorded. The final data set used for analysis included 13 029 individuals, 6940 of whom received BFP. The data set contained a set of 60 covariates that could be used for propensity score matching (ie, categorical or numerical data).

Many of these 60 covariates had a considerable amount of missing data. Data were assumed to be missing completely at random. Variables that were recorded as missing in over 50% of individuals were omitted from the analysis. These variables included house type (permanent/improvised), roof, floor, and wall material, number of people and families in the home, number of bedrooms and bathrooms, variables relating to employment status, expenditure on rent and transport, and receipt of pension, unemployment benefit and alimony. It is conceivable that rent and transport expenditure could be important confounders of treatment success rate given the potential of cash transfers for mitigating catastrophic costs, but neither are conditionally associated with both outcome and exposure in the observed data and expenditure is represented by other retained variables.[21]

The omission of variables with this level of missing data resulted in 45 covariates to be considered for use in propensity score estimation. A sensitivity analysis was run omitting all variables with over 25% missing data, which further omitted water expenditure and years of formal education. At both missing data thresholds, at least one proxy covariate remained under each node of the DAG such that no high-level construct was unrepresented by the available covariates.

Propensity score matching

Without applying propensity score approaches or other approaches to control for confounding, it is likely that the values of the available covariates between the exposed and the unexposed (and those who experience or do not experience the outcome) vary, which potentially biases comparisons between groups. We wish to achieve a 'balance' in these values, which may approximate the balance produced by conventional randomisation procedures. We wish to first determine the likelihood of receiving BFP given the covariate values, which is represented by the propensity score. If the propensity score is then balanced between groups by matching, it is as though the covariates that were used to estimate the propensity score were themselves balanced.[22]

Propensity scores were estimated by logistic regression. One of two criteria must be met for a variable to be included in this logistic regression: (A) conditional

Table 1 Variables to operationalise constructs included in the statistical models

Node (construct)	Covariates included in the model	Covariates excluded from the model (missing data threshold)	Covariates excluded from the model (no available measure)
State	State		
Race	Race, indigenous, quilombola		
Local area	Urbanicity, running water, sewage, electricity, water store, garbage collection	House type	Transit access
Education	Years of education, literacy		
Socioeconomic vulnerability	Child work, institutionalisation, work-acquired TB	Employment, pension receipt, unemployment benefit, alimony receipt	Food security, adequate nutrition, perception of poverty
Age and sex	Age, sex		Gender identity
Comorbidities	AIDS, alcohol use disorder, diabetes, HIV, mental disorder, other chronic illness		General mental health, stress
Income	Income		
Expenditure	(on) Food, energy, gas, water	(on) Rent, transport	Medical costs
Health-seeking behaviour	Directly observed treatment		Engagement with primary care
TB form and severity	Chest X-ray, initial sputum smear, pulmonary/extrapulmonary, throat culture, tuberculin skin test		MDR-TB (is included in outcome as non-successful treatment)
Drug regimen	Rifampicin, isoniazid, ethambutol, streptomycin, pyrazinamide, ethionamide, other drugs		

Not all covariates included under one of the constructs in the directed acyclic graph (DAG) were included in the propensity score model. Table 1 summarises which covariates were included and which were excluded. Some covariates that might reasonably be part of the pathways encoded in this DAG were excluded as there was no adequate measure of them in these linked administrative data. Other covariates were excluded by the missing data threshold, which itself was chosen to balance measurability of each of the constructs with the loss of sample size from undertaking a complete case analysis.
The housing quality node was not included in the model as it was not associated with outcome (TB mortality) or exposure. The housing node included measurable covariates of roof, floor, and wall material, number of people in the home, and the number of bedrooms and bathrooms, as well as the unmeasurable covariate of indoor air pollution.
MDR-TB, multidrug-resistant tuberculosis; TB, tuberculosis.

association with the outcome given exposure, to improve precision or (B) both association with exposure and conditional association with outcome given exposure, to account for confounding.[23] These criteria apply to both mediators and confounders and can be determined from the DAG (figure 1). All DAG nodes meet these criteria but housing and thus the covariates used to model the propensity score were all non-housing covariates meeting the missing data threshold. Quadratic forms of the continuous covariates were used in the logistic regression but sensitivity analyses were performed without including them. Two-way interactions between gender and all variables and age and all variables were also used, given it is likely that these covariates would differ in effect across strata.

Each patient who did not receive BFP (ie, not exposed) was matched to a patient who did receive it (ie, exposed) closest in propensity score, within a particular 'caliper' of 0.1 SD from the mean propensity score. Matching was done with replacement and multiple matches to minimise both bias and variance, following Caliendo and Kopeinig.[24] Multiple matches were weighted to form one matched control for each patient. Standardised mean differences and overlap plots were examined to assess whether balance was improved by matching.

Throughout the literature, complete cases are used for propensity score matching, and this is the approach used in this paper.[24] This reduced the data set to 2167 individuals at the 50% missing data threshold and 3048 individuals at the 25% threshold.

Estimating the impact of Bolsa Familia

Taking the difference of the proportion of treatment success between matched groups resulted in an estimate of the average effect of treatment on the treated (ATT), or the (causal) risk difference in the exposed. The procedure used in Abadie and Imbens[25] was used to estimate the SE of the ATT and thus the CIs. The CIs thus account for the uncertainty due to the matching procedure, but do not account for the uncertainty due to the fact that the estimated propensity score is itself a function of the data; this latter feature leads to conservative inferences.[25] The

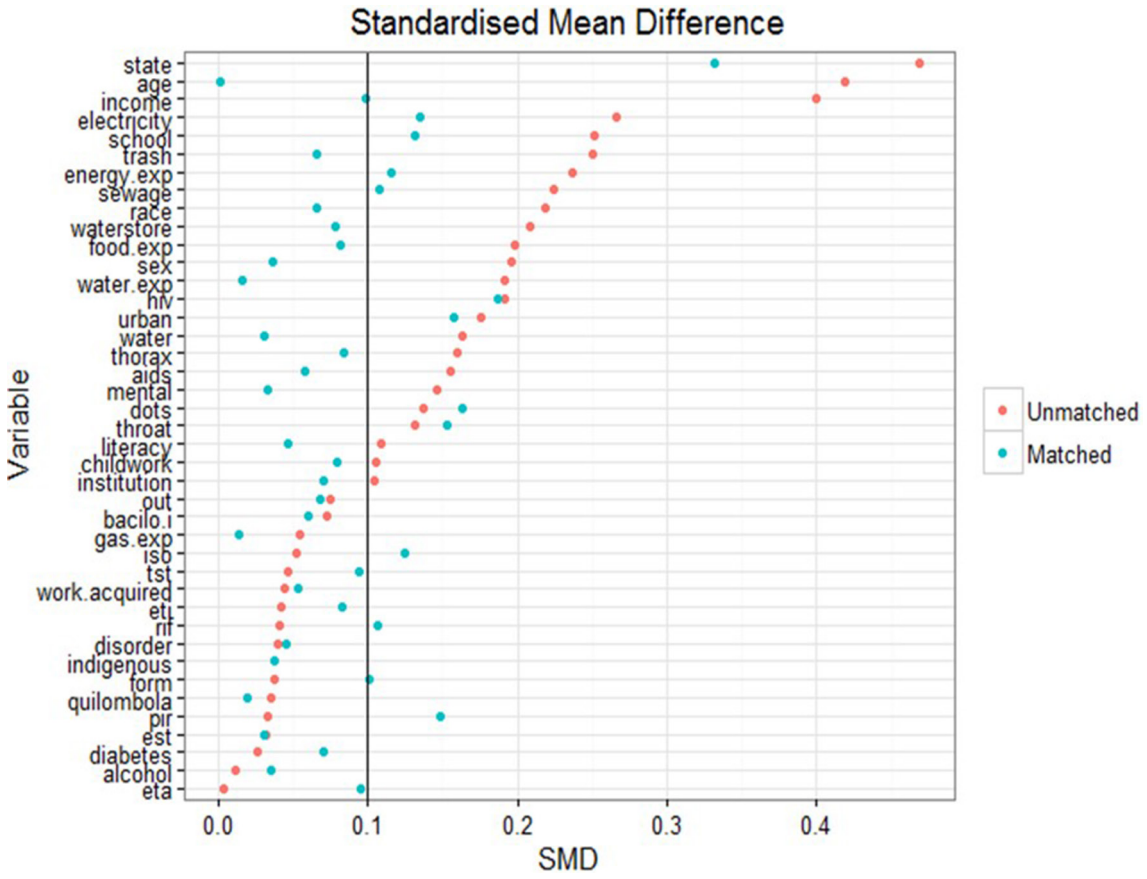


Figure 2 Standardised mean difference (SMD). The change in SMD in the matched and unmatched groups for each variable. A smaller difference indicates improved balance between groups; being below the threshold of 0.1 is conservatively considered to be effectively balanced. Balance has been largely improved by matching though some imbalance remains between groups. bacilo.i, initial sputum smear; disorder, any other chronic illness; est, streptomycin; eta, ethambutol; eti, ethionamide; exp, expenditure; iso, isoniazid; mental, mental disorder; pir, pyrazinamide; rif, rifampicin; thorax, chest X-ray; throat, throat culture; tst, tuberculin skin test.

ATT was also estimated by a multiple imputation-based sensitivity analysis, and point estimates from this are provided for comparative purposes in online supplementary appendix 2.

Statistical software

All analyses were conducted in R V.3.4.1 and the MatchIt package was used for the propensity score matching procedure.

RESULTS

Propensity score matching: covariate balance

A complete balance table is presented in table 1 in online supplementary appendix 1 for the match produced by model A for all covariates included in the propensity score matching exercise. There is good similarity of the covariates after matching, suggesting a reasonable balance was obtained between groups. Prior to matching, there were some imbalances found between BFP recipients and non-recipients on important covariates. Figure 2 presents the changes in standardised mean difference between those receiving BFP (ie, exposed) and those not receiving BFP (ie, not exposed) before and after matching. Figure 3 presents overlap plots to demonstrate the similarity of the propensity score values between groups.

Propensity score matching in general resulted in improved balance of the values of covariates between cases and controls. A standardised mean difference of below 0.1 implies that groups do not differ greatly between values of the covariate.[23] Though the matching process only brought 50% of the imbalanced variables below this threshold, a large improvement was seen on the balance of important upstream covariates like age (0.42 to 0.01), income (0.40 to 0.09) and schooling (0.24 to 0.12). The change in distributions of these variables after matching can be seen in figure 3. On average, those receiving BFP in the unmatched cohort were younger (34.5 vs 41.3 years), poorer (R$65.2 vs R$197.4 per month) and less educated (89.2% vs 83.5% not completed secondary school).

From figure 3, up 20.9% of patients with TB fall under the R$70 income threshold for unconditional receipt of BFP and therefore are theoretically eligible for the programme, but yet excluded from it. A further 29.4% fall under the R$140 income threshold and could therefore potentially be eligible for BFP.

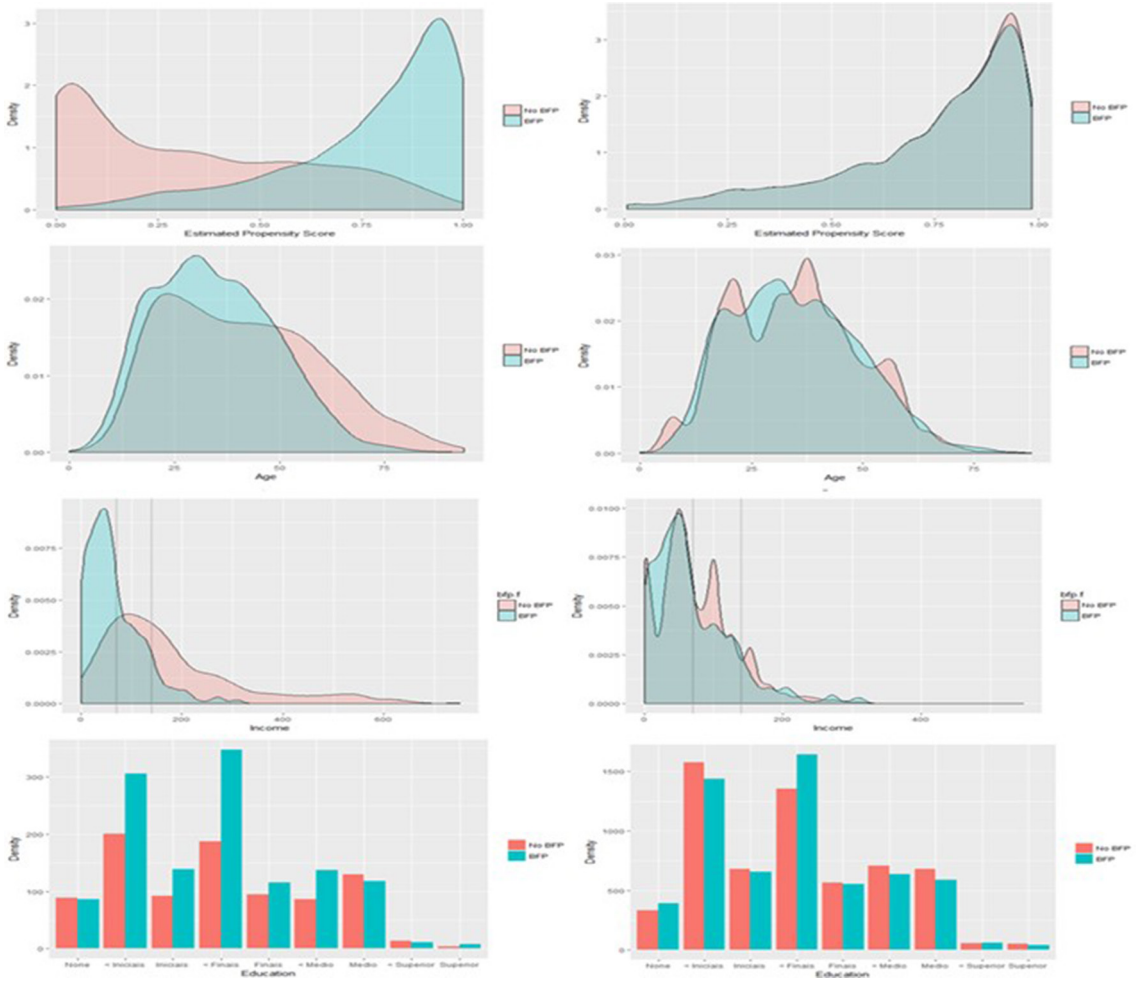


Figure 3 Overlap in estimated propensity scores between those receiving and those not receiving Bolsa Família Programme (BFP) before matching (top left) and after matching (top right). Overlap has been substantially improved by matching to treated (exposed) patients, suggestive of the groups being balanced on the propensity score. The region of overlap extends between 0 and 1. Also presented are similar plots of variable distribution before and after matching for income, age and schooling (from top to bottom). Dotted lines on the income distributions mark the thresholds for BFP eligibility.

Estimating the impact of Bolsa Familia

In total, four estimates of the ATT were produced (table 2). Model A is the primary model of interest as it is the most complex model specification. Models B–D represent sensitivity analyses on model A to investigate how sensitive the results are to simplifying changes to these modelling and missing data decisions.

The ATT from model A was estimated to be 10.58 (95% CI 4.39 to 16.77) (table 2). Thus, among patients with TB who receive BFP, we expect a treatment success rate of 10.58 percentage points higher than if those patients had not received the benefit. The proportion successfully treated in those who did not receive BFP was 76.6% compared with 87.2% in the BFP recipients. This average treatment effect is protective even when a simpler model is used and when the missing data threshold at which covariates are omitted is reduced to 25%, with ATT estimates between 6.31 and 7.21 (table 2). It is also in broad agreement with an ATT point estimate of 7.22 obtained from a multiple imputation approach (online supplementary appendix 2). Expressed as number needed to treat, the estimated ATT implies that on average, among patients with TB who received Bolsa Familia before acquiring TB, one unsuccessful treatment outcome was averted because of Bolsa Familia for every nine patients.

DISCUSSION

Summary: interpretation of results

This is the first study that uses a quasi-experimental approach to estimate the impact of a conditional cash transfer programme on TB treatment success rates.[9] Across all models, results have shown a substantial absolute increase in TB treatment success rate (between 7% and 11%) among those who receive BFP. This seems to suggest a consistent positive association between receiving BFP on a key indicator of TB control: treatment success rate. This is in line with the studies of Torrens *et al*[8] and Durovni *et al*[15] and a few other previous studies evaluating the relationship between social protection and TB outcomes undertaken using less rigorous methodologies and adjusting for only a subset of potential

Table 2 Results of propensity score matching estimates of the ATT for four models

Models* n controls=898 n exposed=1269	ATT	95% CI	Controls matched (unweighted), n	Exposed dropped, n	Pairs matched (weighted), n	Unique controls, n
Model A†	10.58	(4.39 to 16.77)	6021	109	1160	545
Model B‡	7.21	(1.33 to 13.09)	6468	21	1248	656 (D2)
Models* n controls=1319 n exposed=1729	**ATT**	**95% CI**	**Controls matched (unweighted), n**	**Exposed dropped, n**	**Pairs matched (weighted), n**	**Unique controls, n**
Model C*	6.31	(1.46 to 11.16)	8895	70	1659	955
Model D*‡	7.06	(2.57 to 11.56)	9272	17	1712	1001

The matching used was many-to-one with replacement. Some exposed patients were not similar enough to any control patients according to the calliper threshold and these individuals were dropped from the analysis (exposed dropped). Some controls were not similar enough to any exposed patients and were thus not used as potential matches and dropped from the analysis. The remaining controls (unique controls) were then 'copied' a number of times to be used as potential matches (controls matched unweighted). Each control was not matched individually, but rather weighted to form one matched comparator for each treatment patient. These matched comparator patients were matched to the treatment patients to form matched pairs (pairs of controls and treated cases matched). The number of pairs may thus be higher than the total initial sample size as some controls were used more than once and some were not used at all.
*Models C and D omit variables with >25% missing data.
†Model A includes linear and quadratic forms of continuous covariates and omits variables with >50% missing data to estimate the propensity score. Variables included in the final propensity score are those listed in bold in the caption to figure 1.
‡Models B and D omit quadratic forms of continuous covariates.
ATT, average effect of treatment on the treated.

confounders, which also demonstrate a protective effect of similar scale.[13 26] Given the already relatively high treatment success rate in Brazil, it can be expected that the size of impact may be even higher in settings within and outside Brazil, with lower treatment success rates and a less effective TB control programme. Similar propensity score approaches have already been used to evaluate the effect of cash transfers in HIV/AIDS, but not on TB.[27]

Another important and somewhat unexpected finding of our analysis is that the profile of patients with TB enrolled in BFP was not overtly dissimilar from patients with TB who have not received BFP even before matching. Figure 2 suggests that the most imbalanced covariates for receipt of BFP (based on the standardised mean difference) were state of residence, income, age and schooling. There may also be differences between recipients and non-recipients based on measures of the infrastructure of the local area (sewage, electricity, trash disposal). Patients with TB not benefiting from BFP transfers appear to be broadly similar to patients with TB who are BFP recipients under a number of other sociodemographic characteristics, particularly on comorbidities such as diabetes and alcohol abuse, as well as on DOT prevalence table 2 in online supplementary appendix 1. This suggests there may be some shared vulnerability among patients with TB (ie, concomitant socioeconomic stressors, diverse ability to navigate complex social services), who are not captured by the current BFP targeting and enrolment process, leading to some degree of disparity in access to social protection and specifically BFP in Brazil. Even when looking strictly to the BFP eligibility criterion (ie, income), our results show that up to 51.3% of patients may be theoretically eligible for BFP, but yet left out. This seems to further suggest that the income threshold for BFP is insufficiently specific to ensure access to vulnerable patients with TB.

STRENGTH AND LIMITATIONS

The utilisation of quasi-experimental approach is a major strength of this paper. Quasi-experimental approaches like propensity score matching require fewer assumptions about the data than traditional parametric counterparts. The specification of the estimand and population parameters of interest are an additional strength to using propensity score matching, and the risk of bias from residual confounding is minimised compared with prior work by careful use of a DAG.[28] While the use of propensity scores for matching has recently drawn some criticism,[29] the diagnostic plots demonstrated in figures 2 and 3 show that balance was improved by matching, and a number of model specifications for the propensity score were tested and found to demonstrate a similar positive impact.

Indeed, a clear strength of this work is the comparability of the control group. As demonstrated in figure 3, those in the exposed group and those in the control group have a very similar distribution of propensity to receive BFP. This overlap suggests that we are only comparing patients with similar covariate profiles: while some of the control patients may not be eligible on paper for BFP, in the complex context of real-world receipt of BFP, the not-exposed group (our 'control' group) resemble almost exactly those patients with TB who receive BFP on

all measured variables and are representative of a broad range of patients with TB from across Brazil. This is a methodological improvement over the control groups seen in prior work which greatly strengthens the quality of evidence available to policymakers.

The control group in the study of Durovni *et al*[15] was taken from a pool of all patients with TB rather than those who are registered in CadÚnico, and therefore some patients ineligible for BFP were included in the control group. The control group in the study of Torrens *et al*[8] was taken from patients with TB who were eligible in theory for BFP, but who had not received any money from the programme until after treatment. This control group had characteristics different from those patients with TB not eligible for the programme on demographic and socioeconomic variables examined by the authors. Both of these control groups may have potentially biased the resulting estimate of proportion of patients cured attributable to BFP.

This quasi-experimental approach also implies the possibility of drawing causal conclusions. The estimand used in this study, the average treatment effect on the treated, could be given a causal interpretation if particular 'identifying' assumptions hold, including: (1) positivity, which implies that no individual has a probability of 1 of receiving BFP conditional on their confounders; (2) consistency, which implies that different variations of receiving BFP do not have different effects on TB outcomes; and (3) conditional exchangeability, which implies that there is no residual confounding. We note that while BFP might appear to create a structural violation of the positivity assumption with its income threshold, examining the threshold itself it was noted that the cut-off was often inaccurately applied and thus very few random positivity violations were encountered in the matched set. With regard to the consistency assumption, we specifically assumed that receipt of any amount of transfer for any amount of time was sufficient in this context, but further work should investigate dose–response relationships between cash transfers and TB. Drawing causal conclusions is however hampered by the non-interference assumption, which in this context assumes that the exposure received by one individual does not affect the outcome of the other. The results of this study suggest that the size of effect found may be too large to ignore this assumption and work should be undertaken to investigate the effect of social protection on TB transmission. Another potential violation of this assumption is that BFP increases the probability of treatment success in recipients and in other cases through community effects of the cash transfer.

In conclusion, while most identifying assumptions are potentially plausible, we cannot draw conclusions about causality given the interference limitations outlined above. The circumstances under which causal inferences can be drawn with interference is an area of ongoing research.[30]

Another limitation to this work is the data quality. The missing data results in a relatively small sample size used for matching and we cannot rule out the possibility of residual confounding from covariates that are mostly missing or remain unbalanced. Remaining imbalance on the state variable suggests data may be missing conditionally at random on the state variable. As information on it is housed within a separate register, we were unable to assess the impact of the Family Health Strategy, (FHS) though previous work suggests the effect of BFP is independent of Family Health Strategy (FHS) coverage.[15] While an approach combining multiple imputation and propensity score matching would have mitigated this problem, there remain many gaps in the literature on the practical implementation of these techniques together (see online supplementary appendix 2). Furthermore, the data linkage is cross-sectional and thus time-varying confounding cannot be accounted for with these data; better data availability longitudinally would allow for measurement on more direct measures of TB control, such as incidence.

The choice of a dichotomous outcome variable may be another limitation: non-success outcomes include continued disease after regimen completion, treatment abandonment, death from TB, death from other causes and development of multidrug-resistant TB, which may have heterogeneous risk factors. Loss to follow-up and transferred cases are also not considered by this analysis—the analysis is agnostic about whether these patients were cured or not cured. The results may be different if each non-success outcome were addressed in turn, but this would require a larger sample size and may be best addressed in a descriptive study.

Policy implications

Despite the above limitations, these findings preliminarily suggest that: (1) there is a considerable proportion of patients with TB eligible for BFP that for unknown reasons seem to be left out from the programme; (2) almost half of the patients with TB will not be eligible for BFP according to income thresholds, and thus there is room for a more comprehensive or multidimensional targeting approach not only using income as eligibility criteria. Given the 7%–11% absolute increase in treatment success rate seen among those receiving BFP from our work, from a health rights perspective, it must be considered how best to deliver a protective programme to vulnerable patients in Brazil.

BFP was not designed to address specific diseases, not least TB: TB status is not a targeting criterion and none of the conditionalities currently imposed by the programme have any direct implication for TB care and/or TB control. Despite the suggested positive impact, ethical and equity issues make unlikely that TB will become one of the eligibility criteria of BFP. Nonetheless, access could be expanded, and thus impact maximised, by making BFP more TB sensitive through a more inclusive, although non-stigmatising, targeting strategy. Higher

impact could in fact be achieved by simply ensuring that patients who are already eligible by definition for the programme receive the benefits, or at least receive them while on treatment. To this purpose, further research is urgently needed to understand determinants of access to BFP from patients with TB and to explore those supply and demand side barriers that delay the transfer of benefits once patients with TB are legitimately enrolled.

Understanding how to effectively and cost-effectively remove these individual and system-level barriers and what may be the ultimate impact on the Brazilian TB epidemic is a priority research area, whose lessons may be transferrable to other settings.

Nonetheless, it can be anticipated that the removal of these barriers may require the implementation of more efficient BFP delivery models, including the 'single window' approach which entails an integrated delivery of TB care services and social protection.[31] According to this model, the access to the most appropriate social protection schemes is determined and facilitated at the primary healthcare level where ad hoc staff (eg, social workers) are trained to assess the social protection needs of patients with TB and provide information, legal and administrative advices, and referrals to various services so to allow patients to access benefits from one 'single window' without having to navigate across complex and multiple service points.[31]

Another emerging model for the delivery of social protection is the 'cash plus' model in which the provision of cash transfers is combined with another form of social support when the provision of in-kind benefits is not deemed sufficient to reduce households' vulnerabilities (including health-related vulnerabilities).[32]

In the case of TB in Brazil, this 'plus' component can be represented by a top-up of the cash benefit to account for the TB-related catastrophic costs incurred by the households; or the provision of a food basket to improve nutrition of cash beneficiaries and therefore their treatment outcome; or the improvement of housing and ventilation conditions to interrupt intrahousehold transmission of TB. To identify the most relevant 'intensifier' of any cash transfer intervention it will be essential also to understand thoroughly the most likely pathway through which this impact takes place. This requires the development of a setting-specific, epidemiologically driven conceptual framework and a more comprehensive collection of data for the variables in the causal pathway.

To be useful the above research agenda should rely on both quantitative and qualitative methods to embrace the complexity of pathways likely to underlie impact and the multifaced nature of determinants of access to cash transfers in the context of TB-affected communities.

CONCLUSIONS

Overall, the strength of evidence and size of effect of the ATT estimated in this work seems to suggest that expanding social protection to a wider population of patients with TB may represent a valid mechanism for improving TB outcomes beyond the traditional biomedical approach. This is consistent with the need of a multisectoral accountability framework expressed during the last WHO-Global Ministerial Conference held in Moscow in November 2017 which demands a more pervasive integration of TB programmatic action within development models and infrastructures.[33] It is essential that, like in this work, recent developments in quasi-experimental methodology continue to be integrated with the evidence base for bold policies in development. With stronger evidence available, the rapid implementation of bold policies may be justified in TB contexts and the global public health community will be a large step closer to achieving the aims of the WHO's End TB Strategy.

Author affiliations
[1]Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK
[2]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK
[3]Tropical Medicine Department, University of Brasília, Brasília, Brazil
[4]Federal University of Brasília, Brasília, Brazil
[5]Federal University of Espírito Santo, Vitoria, Brazil
[6]National Tuberculosis Programme/Ministry of Health, Brasília, Brazil
[7]National Tuberculosis Programme/Ministry of Health of Brazil, Brasília, Brazil
[8]Centro de Pesquisas Gonçalo Muniz, Fundação Oswaldo Cruz, Salvador, Brazil
[9]Institute of Collective Health, Federal University of Bahia, Salvador, Brazil
[10]Centro de Integração de Dados de Conhecimentos para Saúde (CIDACS), Fundação Oswaldo Cruz, Salvador, Brazil

Present affiliations The present affiliation of Rhian Daniel is: Division of Population Medicine, Cardiff University, Wales, United Kingdom.

Contributors DJC was in charge of the data analysis and drafted the first version of the manuscript. RD provided statistical supervision of the data analysis. DB conceived the study, planned the data analysis and contributed to the results interpretation and paper writing and submission to the journal. ELNM, MNS, AWT, DR, LCR, MLB, DCB and PB made the data available, supported the data analysis and results interpretation.

Funding This work was sponsored by a grant from the Wellcome Trust to the PI (No 104473/Z/14/Z).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Research Ethics Committee of the Institute of Center of Health Sciences of the Federal University of Espirito Santo (protocol number 242831).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The data used in this paper, including the unpublished data, belong to the Brazilian Ministry of Health and the Ministry of Social Development. The data set builds upon the data linkage between the Brazilian National TB Registry and the CadUnico database for year 2010. Permission for data sharing should be addressed directly to the Ministry of Health and the Ministry of Social Development in Brazil.A preliminary open access version of this paper has been prepublished on bioRxiv (https://www.biorxiv.org/about-biorxiv) and is available at https://www.biorxiv.org/content/early/2018/04/30/311589.

REFERENCES

1. WHO. *Global Tuberculosis Report*. Geneva, Switzerland: World Health Organization, 2017.
2. Raviglione M, Zumla A, Marais B, *et al*. A sustainable agenda for tuberculosis control and research. *The Lancet* 2012;379:1077–8.
3. Lönnroth K, Jaramillo E, Williams BG, *et al*. Drivers of tuberculosis epidemics: The role of risk factors and social determinants. *Social Science & Medicine* 2009;68:2240–6.
4. Lönnroth K, Raviglione M. The WHO's new end TB strategy in the post-2015 era of the sustainable development goals. *Trans R Soc Trop Med Hyg* 2016;110:148–50.
5. Rudgard WE, Evans CA, Sweeney S, *et al*. Comparison of two cash transfer strategies to prevent catastrophic costs for poor tuberculosis-affected households in low- and middle-income countries: an economic modelling study. *PLoS Med* 2017;14:e1002418.
6. Uplekar M, Weil D, Lonnroth K, *et al*. WHO's new end TB strategy. *The Lancet* 2015;385:1799–801.
7. Soares TS, Familia B, design its. *its impact and possibilities for the future. Working paper Number 89*. Brazilia, Brazil: International Policy Centre for Inclusive Growth, 2012.
8. Torrens AW, Rasella D, Boccia D, *et al*. Effectiveness of a conditional cash transfer programme on TB cure rate: a retrospective cohort study in Brazil. *Trans R Soc Trop Med Hyg* 2016;110:199–206.
9. Richterman A, Steer-Massaro J, Jarolimova J, *et al*. Cash interventions to improve clinical outcomes for pulmonary tuberculosis: systematic review and meta-analysis. *Bull World Health Organ* 2018;96:471–83.
10. van Hoorn R, Jaramillo E, Collins D, *et al*. The effects of psycho-emotional and socio-economic support for tuberculosis patients on treatment adherence and treatment outcomes – a systematic review and meta-analysis. *Plos One* 2016;11:e0154095.
11. Freitas de Andrade K, Silva Nery J, Andrade de Souza R. Effects of social protection on tuberculosis treatment outcomes in low or middle-income and in high-burden countries: systematic review and meta-analysis. *Cad Saude Publica* 2018;34.
12. Wingfield T, Tovar MA, Huff D, *et al*. A randomized controlled study of socioeconomic support to enhance tuberculosis prevention and treatment, Peru. *Bull World Health Organ* 2017;95:270–80.
13. Ciobanu A, Domente L, Soltan V, *et al*. Do incentives improve tuberculosis treatment outcomes in the Republic of Moldova? *Public Health Action* 2014;4:59–63.
14. Lutge E, Lewin S, Volmink J, *et al*. Economic support to improve tuberculosis treatment outcomes in South Africa: a pragmatic cluster-randomized controlled trial. *Trials* 2013;14:154.
15. Durovni B, Saraceni V, Puppin MS, *et al*. The impact of the Brazilian family health strategy and the conditional cash transfer on tuberculosis treatment outcomes in Rio de Janeiro: an individual-level analysis of secondary data. *J Public Health (Oxf)* 2018;40:e359–66.
16. Rudgard WE, das Chagas NS, Gayoso R, *et al*. Uptake of governmental social protection and financial hardship during drug-resistant tuberculosis treatment in Rio de Janeiro, Brazil. *Eur Respir J* 2018;51:1800274.
17. Nery JS, Rodrigues LC, Rasella D, *et al*. Effect of Brazil's conditional cash transfer programme on tuberculosis incidence. *Int J Tuberc Lung Dis* 2017;21:790–6.
18. Boccia D, Rudgard W, Shrestha S, *et al*. Modelling the impact of social protection on tuberculosis: the S-PROTECT project. *BMC Public Health* 2018;18:786.
19. Hargreaves JR, Boccia D, Evans CA, *et al*. The social determinants of tuberculosis: from evidence to action. *Am J Public Health* 2011;101:654–62.
20. Maciel EL, Reis-Santos B. Determinants of tuberculosis in Brazil: from conceptual framework to practical application. *Rev Panam Salud Publica* 2015;38:28–34.
21. Wingfield T, Boccia D, Tovar M, *et al*. Defining catastrophic costs and comparing their importance for adverse tuberculosis outcome with multi-drug resistance: a prospective cohort study, Peru. *PLoS Medicine* 2014;11:e1001675.
22. Rosenmbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
23. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;46:399–424.
24. Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. *J Econ Surv* 2008;22:31–72.
25. Abadie A, Imbens GW. Bias-corrected matching estimators for average treatment Effects. *J Bus Econ Stat* 2011;29:1–11.
26. Sripad A, Castedo J, Danford N, *et al*. Effects of Ecuador's national monetary incentive program on adherence to treatment for drug-resistant tuberculosis. *Int J Tuberc Lung Dis* 2014;18:44–8.
27. Cluver L, Boyes M, Orkin M, *et al*. Child-focused state cash transfers and adolescent risk of HIV infection in South Africa: a propensity-score-matched case-control study. *Lancet Glob Health* 2013;1:e362–e370.
28. Williamson E, Morley R, Lucas A, *et al*. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res* 2012;21:273–93.
29. King G, Nielsen R, 2016. Why propensity scores should not be used for matching. Available from: http://j.mp/2ovYGsW [Accessed 14 Aug 2018].
30. Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–92.
31. Ebken C. *Single Window Services in Social Protection: rationale and design features in developing country contexts. Discussion papers on social protection*. Bonn, Germany: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, 2014.
32. Roelen K, Devereux S, Abdulai A. *How to make 'cash plus' work: linking cash transfers to services and sectors, innocenti working papers no. 10*. UNICEF Innocenti, 2017.
33. Raviglione M, Uplekar M, Weil D, *et al*. Tuberculosis makes it onto the international political agenda for health…finally. *Lancet Glob Health* 2018;6:e20–e21.

Appendix 1 - Baseline Characteristics.

The following table presents the baseline characteristics of the unmatched population and the same characteristics after matching. Variables correspond to the covariates of the DAG in Figure 1 in the manuscript that were included in the propensity score model.

Table 1. Balance on covariates both before and after matching, stratified by receipt of Bolsa Familia.

	Unmatched		*Matched (unweighted)*	
Group	**No BFP**	**BFP**	**No BFP**	**BFP**
N	898	1269	6021	6021
Sex (%)				
F	418 (46.5)	711 (56.0)	3432 (57.0)	3323 (55.2)
I	0 (0.0)	1 (0.1)	0 (0.0)	0 (0.0)
M	480 (53.5)	557 (43.9)	2589 (43.0)	2698 (44.8)
Age (mean (sd))	41.28 (18.19)	34.45 (14.21)	34.86 (14.26)	34.85 (14.20)
Race (%)				
Branca	311 (34.6)	321 (25.3)	1430 (23.8)	1550 (25.7)
Preta	122 (13.6)	199 (15.7)	1081 (18.0)	965 (16.0)
Amarela	10 (1.1)	7 (0.6)	26 (0.4)	36 (0.6)
Parda	451 (50.2)	736 (58.0)	3484 (57.9)	3470 (57.6)
Indigena	4 (0.4)	6 (0.5)	0 (0.0)	0 (0.0)

Indigenous = Not Indigenous (%)	896 (99.8)	1268 (99.9)	6007 (99.8)	6016 (99.9)
Quilombola = Not Quilombola (%)	895 (99.7)	1267 (99.8)	6012 (99.9)	6016 (99.9)
Years of Education (n years)				
None	89 (9.9)	86 (6.8)	333 (5.5)	394 (6.5)
Fundamental I incomplete (< 5)	201 (22.4)	306 (24.1)	1578 (26.2)	1439 (23.9)
Fundamental I complete (5)	92 (10.2)	139 (11.0)	680 (11.3)	657 (10.9)
Fundamental II incomplete (< 9)	187 (20.8)	348 (27.4)	1357 (22.5)	1645 (27.3)
Fundamental II complete (9)	95 (10.6)	116 (9.1)	568 (9.4)	555 (9.2)
Medio incomplete (< 12)	86 (9.6)	137 (10.8)	710 (11.8)	638 (10.6)
Medio complete (12)	130 (14.5)	118 (9.3)	681 (11.3)	590 (9.8)
Superior incomplete (< 16)	14 (1.6)	11 (0.9)	59 (1.0)	60 (1.0)
Superior complete (16)	4 (0.4)	8 (0.6)	55 (0.9)	43 (0.7)
Literacy = Illiterate (%)	142 (15.8)	153 (12.1)	632 (10.5)	720 (12.0)
Urban (%)				
Urban	838 (93.3)	1121 (88.3)	5064 (84.1)	5383 (89.4)
Rural	54 (6.0)	129 (10.2)	801 (13.3)	544 (9.0)
Periurban	6 (0.7)	19 (1.5)	156 (2.6)	94 (1.6)
Water = No Running Water (%)	76 (8.5)	172 (13.6)	800 (13.3)	738 (12.3)
Sewage (%)				
Sewage System	461 (51.3)	575 (45.3)	2705 (44.9)	2849 (47.3)
Septic Tank	176 (19.6)	238 (18.8)	1077 (17.9)	1103 (18.3)

Tank	192 (21.4)	372 (29.3)	1953 (32.4)	1717 (28.5)
Open Air	49 (5.5)	59 (4.6)	246 (4.1)	274 (4.6)
Into Water	1 (0.1)	10 (0.8)	2 (0.0)	10 (0.2)
Other	19 (2.1)	15 (1.2)	38 (0.6)	68 (1.1)
Electricity (%)				
Own Metered	792 (88.2)	1002 (79.0)	4801 (79.7)	4839 (80.4)
Central Metered	33 (3.7)	64 (5.0)	375 (6.2)	304 (5.0)
Unmetered	53 (5.9)	134 (10.6)	448 (7.4)	591 (9.8)
Gas or Oil	2 (0.2)	9 (0.7)	0 (0.0)	0 (0.0)
Candle	4 (0.4)	8 (0.6)	24 (0.4)	37 (0.6)
Other	14 (1.6)	52 (4.1)	373 (6.2)	250 (4.2)
Water Source (%)				
Pipe Network	801 (89.2)	1040 (82.0)	5032 (83.6)	5067 (84.2)
Well	64 (7.1)	154 (12.1)	657 (10.9)	648 (10.8)
Cistern	7 (0.8)	14 (1.1)	101 (1.7)	50 (0.8)
Other	26 (2.9)	61 (4.8)	231 (3.8)	256 (4.3)
Trash (%)				
Direct Collect	759 (84.5)	1003 (79.0)	4923 (81.8)	4856 (80.7)
Indirect Collect	14 (1.6)	54 (4.3)	204 (3.4)	223 (3.7)
Household	48 (5.3)	127 (10.0)	547 (9.1)	511 (8.5)
Street	18 (2.0)	22 (1.7)	76 (1.3)	110 (1.8)

Other	59 (6.6)	63 (5.0)	271 (4.5)	321 (5.3)
Thorax X-Ray (%)				
Suspect	763 (85.0)	1009 (79.5)	4954 (82.3)	4821 (80.1)
Normal	43 (4.8)	70 (5.5)	286 (4.8)	321 (5.3)
Other Pathology	11 (1.2)	14 (1.1)	26 (0.4)	62 (1.0)
Not Undertaken	81 (9.0)	176 (13.9)	755 (12.5)	817 (13.6)
Initial Bacilloscopy (%)				
Positive	206 (22.9)	329 (25.9)	1452 (24.1)	1543 (25.6)
Negative	291 (32.4)	385 (30.3)	1725 (28.6)	1813 (30.1)
Not Performed	401 (44.7)	555 (43.7)	2844 (47.2)	2665 (44.3)
Form (%)				
Pulmonary	763 (85.0)	1079 (85.0)	5337 (88.6)	5143 (85.4)
Extrapulmonary	108 (12.0)	159 (12.5)	596 (9.9)	738 (12.3)
Both P & E	27 (3.0)	31 (2.4)	88 (1.5)	140 (2.3)
Throat Culture (%)				
Positive	79 (8.8)	111 (8.7)	587 (9.7)	508 (8.4)
Negative	88 (9.8)	83 (6.5)	571 (9.5)	399 (6.6)
In Progress	40 (4.5)	74 (5.8)	450 (7.5)	330 (5.5)
Not Performed	691 (76.9)	1001 (78.9)	4413 (73.3)	4784 (79.5)
Tuberculin Skin Test (%)				
No Reaction	61 (6.8)	79 (6.2)	414 (6.9)	366 (6.1)

Some Reaction	24 (2.7)	29 (2.3)	73 (1.2)	136 (2.3)
Strong Reaction	160 (17.8)	244 (19.2)	1198 (19.9)	1104 (18.3)
Not Performed	653 (72.7)	917 (72.3)	4336 (72.0)	4415 (73.3)
Directly Observed Treatment (%)				
DOT	434 (48.3)	692 (54.5)	3595 (59.7)	3198 (53.1)
No DOT	458 (51.0)	563 (44.4)	2417 (40.1)	2768 (46.0)
Unknown	6 (0.7)	14 (1.1)	9 (0.1)	55 (0.9)
Rifampicin = Not Taking (%)	20 (2.2)	21 (1.7)	202 (3.4)	102 (1.7)
Isoniazid = Not Taking (%)	18 (2.0)	17 (1.3)	193 (3.2)	81 (1.3)
Ethambutol = Not Taking (%)	245 (27.3)	348 (27.4)	1439 (23.9)	1689 (28.1)
Streptomycin = Not Taking (%)	889 (99.0)	1260 (99.3)	5990 (99.5)	5975 (99.2)
Pyrazinamide = Not Taking (%)	26 (2.9)	30 (2.4)	320 (5.3)	148 (2.5)
Ethionamide = Not Taking (%)	882 (98.2)	1253 (98.7)	5991 (99.5)	5944 (98.7)
Other Drugs = Not Taking (%)	861 (95.9)	1234 (97.2)	5765 (95.7)	5842 (97.0)
AIDS = No AIDS (%)	819 (91.2)	1207 (95.1)	5778 (96.0)	5705 (94.8)
Alcoholism = No Alcoholism (%)	809 (90.1)	1139 (89.8)	5451 (90.5)	5387 (89.5)
Diabetes = No Diabetes (%)	831 (92.5)	1183 (93.2)	5502 (91.4)	5614 (93.2)
HIV (%)				
Positive	89 (9.9)	69 (5.4)	261 (4.3)	354 (5.9)
Negative	526 (58.6)	744 (58.6)	3966 (65.9)	3507 (58.2)
In Progress	40 (4.5)	84 (6.6)	205 (3.4)	376 (6.2)

Not Undertaken	243 (27.1)	372 (29.3)	1589 (26.4)	1784 (29.6)
Mental Disorder = No Mental Disorder (%)	875 (97.4)	1260 (99.3)	5991 (99.5)	5975 (99.2)
Other Disorder = No Disorder (%)	777 (86.5)	1115 (87.9)	5379 (89.3)	5292 (87.9)
Food Expenditure (mean (sd))	209.50 (131.61)	184.69 (118.13)	192.67 (114.22)	183.22 (117.68)
Energy Expenditure (mean (sd))	40.17 (34.37)	32.45 (30.64)	36.22 (30.73)	32.70 (30.38)
Gas Expenditure (mean (sd))	33.69 (20.21)	32.75 (13.26)	33.04 (11.53)	32.86 (13.15)
Water Expenditure (mean (sd))	22.33 (20.15)	18.28 (21.99)	18.83 (18.93)	18.52 (20.43)
Child Work = No Child Worker (%)	881 (98.1)	1223 (96.4)	5893 (97.9)	5814 (96.6)
Institutionalised (%)				
None	820 (91.3)	1181 (93.1)	5688 (94.5)	5619 (93.3)
Military	30 (3.3)	33 (2.6)	173 (2.9)	172 (2.9)
Asylum	0 (0.0)	1 (0.1)	0 (0.0)	0 (0.0)
Orphanage	0 (0.0)	2 (0.2)	0 (0.0)	0 (0.0)
Psychiatric	1 (0.1)	1 (0.1)	0 (0.0)	0 (0.0)
Other	22 (2.4)	27 (2.1)	67 (1.1)	113 (1.9)
Unknown	25 (2.8)	24 (1.9)	93 (1.5)	117 (1.9)
Work Acquired TB (%)				
Got at work	20 (2.2)	24 (1.9)	109 (1.8)	114 (1.9)
Not at work	790 (88.0)	1134 (89.4)	5281 (87.7)	5370 (89.2)
Unknown	88 (9.8)	111 (8.7)	631 (10.5)	537 (8.9)

State (%)				
Rondonia	3 (0.3)	7 (0.6)	0 (0.0)	0 (0.0)
Acre	6 (0.7)	14 (1.1)	47 (0.8)	46 (0.8)
Amazonias	47 (5.2)	112 (8.8)	514 (8.5)	528 (8.8)
Roraima	2 (0.2)	5 (0.4)	1 (0.0)	15 (0.2)
Para	22 (2.4)	39 (3.1)	304 (5.0)	187 (3.1)
Amapa	0 (0.0)	2 (0.2)	0 (0.0)	0 (0.0)
Tocantins	7 (0.8)	3 (0.2)	15 (0.2)	16 (0.3)
Maranhao	48 (5.3)	70 (5.5)	288 (4.8)	353 (5.9)
Piaui	19 (2.1)	32 (2.5)	76 (1.3)	122 (2.0)
Ceara	118 (13.1)	214 (16.9)	1267 (21.0)	1040 (17.3)
Rio Grande do Norte	7 (0.8)	20 (1.6)	17 (0.3)	72 (1.2)
Paraiba	7 (0.8)	23 (1.8)	62 (1.0)	118 (2.0)
Pernambuco	27 (3.0)	63 (5.0)	128 (2.1)	287 (4.8)
Alagoas	18 (2.0)	35 (2.8)	294 (4.9)	166 (2.8)
Sergipe	15 (1.7)	36 (2.8)	150 (2.5)	166 (2.8)
Bahia	68 (7.6)	106 (8.4)	387 (6.4)	504 (8.4)
Minas Gerais	60 (6.7)	81 (6.4)	436 (7.2)	404 (6.7)
Espirito Santo	39 (4.3)	34 (2.7)	126 (2.1)	158 (2.6)
Rio de Janeiro	55 (6.1)	110 (8.7)	375 (6.2)	536 (8.9)
Parana	102 (11.4)	82 (6.5)	485 (8.1)	420 (7.0)

Santa Caterina	53 (5.9)	18 (1.4)	62 (1.0)	91 (1.5)
Rio Grande do Sul	117 (13.0)	115 (9.1)	758 (12.6)	588 (9.8)
Mato Grosso do Sul	18 (2.0)	16 (1.3)	63 (1.0)	68 (1.1)
Mato Grosso	17 (1.9)	20 (1.6)	116 (1.9)	80 (1.3)
Goias	14 (1.6)	9 (0.7)	24 (0.4)	41 (0.7)
Distrito Federal	9 (1.0)	3 (0.2)	26 (0.4)	15 (0.2)
Income (mean (sd))	197.39 (465.17)	65.22 (56.04)	74.05 (57.81)	68.36 (58.05)

Covariates are grouped by DAG node including *Sex & Age* (Sex, Age), *Race* (Race, Indigenous, Quilombola), *Education* (Education Level, Literacy), *Local Area* (Urban, Running Water, Sewage, Electricity, Water Store, Trash), *Type of TB* (Thorax X-Ray, Initial Bacilloscopy, Form, Throat Culture, Tuberculin Skin Test), *Directly Observed Treatment* (DOT), *Drugs* (Rifampicin, Isoniazid, Ethambutol, Streptomycin, Pyrazinamide, Ethionamide, Other Drugs), *Comorbidities* (AIDS, Alcoholism, Diabetes, HIV, Mental Disorder, Other Disorder), *Expenditure* (on Food, Energy, Gas, and Water), *Social Vulnerability* (Child Worker, Institutionalised, Work Acquired TB), *State*, and *Income.* Where duplicate variables existed, SINAN was used preferentially.

Appendix 2 – Missing Data

Though it was not the primary analytical method for this work, a confirmatory sensitivity analysis using multiple imputation was undertaken using the MICE (multiple imputation by chained equations) approach, as implemented in the MICE package in R.[1-2] The MICE package defaults of predictive mean matching and polytomous regression were used as the imputation methods for numeric and categorical variables respectively, creating 5 multiply imputed datasets.

The literature is still unclear as to whether the pooling of propensity scores themselves or the pooling of treatment estimates is the better approach after multiple imputation. Here, we followed Leyrat et al. (2016) and applied Rubin's rules to pool the ATT estimates from each imputed dataset.[3-4] The ATT estimates were based on a comparison between groups that were matched on the propensity score estimated by the same model specification used for Model A. The resulting estimated ATT was 7.22, in broad agreement with other results.

An approach combining multiple imputation and propensity score methods was not used for the primary analysis due to numerous unresolved questions that admit the possibility for an unknown amount of bias with regards to the estimation of variance after pooling, the timing of pooling datasets, the best number of datasets to impute, the best method for handling imputations, at a minimum. Practical guidelines for methods that more efficiently and robustly account for the incompleteness of data within estimation methods based on the propensity score are needed, but research in this area is ongoing.

References

1. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45. Available: http://www.jstatsoft.org/v45/i03
2. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple Imputation by Chained Equations: What is it and how does it work? Int J Methods Psychiatr Res. 2011;20: 40–49. doi:10.1002/mpr.329
3. Leyrat C, Seaman SR, White IR, Douglas I, Smeeth L, Kim J, et al. Propensity score analysis with partially observed confounders: how should multiple imputation be used? ArXiv160805606 Stat. 2016; Available: http://arxiv.org/abs/1608.05606
4. Rubin DB. Inference and missing data. Biometrika. 1976;63: 581–592. doi:10.1093/biomet/63.3.581

JAMA | Original Investigation

Association Between Early Participation in Physical Activity Following Acute Concussion and Persistent Postconcussive Symptoms in Children and Adolescents

Anne M. Grool, MD, PhD; Mary Aglipay, MSc; Franco Momoli, PhD; William P. Meehan III, MD; Stephen B. Freedman, MDCM, MSc; Keith Owen Yeates, PhD; Jocelyn Gravel, MD; Isabelle Gagnon, PhD; Kathy Boutis, MD; Willem Meeuwisse, MD, PhD; Nick Barrowman, PhD; Andrée-Anne Ledoux, PhD; Martin H. Osmond, MDCM; Roger Zemek, MD; for the Pediatric Emergency Research Canada (PERC) Concussion Team

IMPORTANCE Although concussion treatment guidelines advocate rest in the immediate postinjury period until symptoms resolve, no clear evidence has determined that avoiding physical activity expedites recovery.

OBJECTIVE To investigate the association between participation in physical activity within 7 days postinjury and incidence of persistent postconcussive symptoms (PPCS).

DESIGN, SETTING, AND PARTICIPANTS Prospective, multicenter cohort study (August 2013-June 2015) of 3063 children and adolescents aged 5.00-17.99 years with acute concussion from 9 Pediatric Emergency Research Canada network emergency departments (EDs).

EXPOSURES Early physical activity participation within 7 days postinjury.

MAIN OUTCOMES AND MEASURES Physical activity participation and postconcussive symptom severity were rated using standardized questionnaires in the ED and at days 7 and 28 postinjury. PPCS (≥3 new or worsening symptoms on the Post-Concussion Symptom Inventory) was assessed at 28 days postenrollment. Early physical activity and PPCS relationships were examined by unadjusted analysis, 1:1 propensity score matching, and inverse probability of treatment weighting (IPTW). Sensitivity analyses examined patients (≥3 symptoms) at day 7.

RESULTS Among 2413 participants who completed the primary outcome and exposure, (mean [SD] age, 11.77 [3.35] years; 1205 [39.3%] females), PPCS at 28 days occurred in 733 (30.4%); 1677 (69.5%) participated in early physical activity including light aerobic exercise (n = 795 [32.9%]), sport-specific exercise (n = 214 [8.9%]), noncontact drills (n = 143 [5.9%]), full-contact practice (n = 106 [4.4%]), or full competition (n = 419 [17.4%]), whereas 736 (30.5%) had no physical activity. On unadjusted analysis, early physical activity participants had lower risk of PPCS than those with no physical activity (24.6% vs 43.5%; Absolute risk difference [ARD], 18.9% [95% CI,14.7%-23.0%]). Early physical activity was associated with lower PPCS risk on propensity score matching (n = 1108 [28.7% for early physical activity vs 40.1% for no physical activity]; ARD, 11.4% [95% CI, 5.8%-16.9%]) and on inverse probability of treatment weighting analysis (n = 2099; relative risk [RR], 0.74 [95% CI, 0.65-0.84]; ARD, 9.7% [95% CI, 5.7%-13.7%]). Among only patients symptomatic at day 7 (n = 803) compared with those who reported no physical activity (n = 584; PPCS, 52.9%), PPCS rates were lower for participants of light aerobic activity (n = 494 [46.4%]; ARD, 6.5% [95% CI, 5.7%-12.5%]), moderate activity (n = 176 [38.6%]; ARD, 14.3% [95% CI, 5.9%-22.2%]), and full-contact activity (n = 133 [36.1%]; ARD, 16.8% [95% CI, 7.5%-25.5%]). No significant group difference was observed on propensity-matched analysis of this subgroup (n = 776 [47.2% vs 51.5%]; ARD, 4.4% [95% CI, −2.6% to 11.3%]).

CONCLUSIONS AND RELEVANCE Among participants aged 5 to 18 years with acute concussion, physical activity within 7 days of acute injury compared with no physical activity was associated with reduced risk of PPCS at 28 days. A well-designed randomized clinical trial is needed to determine the benefits of early physical activity following concussion.

JAMA. 2016;316(23):2504-2514. doi:10.1001/jama.2016.17396

Editorial page 2491

Author Video Interview and JAMA Report Video

CME Quiz at jamanetworkcme.com

Author Affiliations: Author affiliations are listed at the end of this article.

Group Information: The Pediatric Emergency Research Canada (PERC) Concussion Team members are listed at the end of this article.

Corresponding Author: Roger Zemek, MD, Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Rd, Ottawa, ON K1H 8L1, Canada (rzemek@cheo.on.ca).

Rest has long been considered the cornerstone of concussion management,[1] and pediatric guidelines universally recommend an initial period of cognitive and physical rest following a concussion.[1,2] Cognitive rest recommendations include modification of school attendance and mental activities.[1] Physical rest recommendations advocate avoidance of physical activity until postconcussive symptoms have resolved, endorsing gradual resumption of activities only if symptoms are not exacerbated.[1,2]

Due to limited high-quality evidence, existing physical rest guidelines are based on consensus and precautionary principles.[2,3] There is limited evidence that following these guidelines results in a positive effect on prognosis.[4,5] Although strenuous exercise in patients recovering from concussion may be deleterious and increase re-injury risk,[6] recent literature suggests that protracted rest may hamper concussion recovery,[7] leading to secondary symptoms of fatigue, depression, anxiety, and physiological deconditioning.[3,8] Increasing evidence suggests the introduction of controlled, light aerobic physical activity following pediatric concussion may be safe[9] while promoting recovery[9] by enhancing physical, psychological, and academic outcomes.[6,9,10] These preliminary findings indicate that gradual resumption of pre-injury activities could begin as soon as tolerated provided there is no increased risk of re-injury.[11]

The objective of this study was to examine the association between participation in physical activity within 7 days postinjury and the occurrence of persistent postconcussive symptoms (PPCS) following concussion in children and adolescents. It was hypothesized that early participation in physical activity would be associated with lower PPCS rates compared with no physical activity.

Key Points

Question Is participation in physical activity within 7 days following acute concussion associated with lower rates of persistent postconcussive symptoms in children and adolescents compared with conservative rest?

Findings In this prospective, multicenter cohort study of 3063 children and adolescents aged 5.00 to 17.99 years after propensity matching, the proportion with postconcussive symptoms at 28 days was 28.7% with participation in early physical activity vs 40.1% with conservative rest, a significant difference.

Meaning Participation in physical activity within 1 week after injury may benefit symptom recovery following acute concussion in children and adolescents.

Methods

This research comprises a planned secondary analysis of the Predicting Persistent Postconcussive Problems in Pediatrics (5P) study,[12,13] a prospective, multicenter cohort study that recruited participants from August 2013 until June 2015 at 9 Pediatric Emergency Research Canada (PERC) network tertiary pediatric emergency departments (EDs).

Participants

This study enrolled 3063 participants aged 5.00 to 17.99 years with ED presentation for acute head injury[12,13] occurring within the preceding 48 hours, who met concussion diagnosis criteria according to the 2012 Zurich consensus statement.[1] Exclusion criteria were a Glasgow Coma Scale score of 13 or less; any abnormality on brain computed tomography or magnetic resonance imaging; neurosurgical intervention, intubation, or intensive care unit admission; multisystem injury requiring hospitalization; severe preexisting neurological developmental delay resulting in communication difficulties; intoxication; absence of trauma as the primary event; previously enrolled in this same study; insurmountable language barrier; or inability to follow-up by phone or electronic-mail. The 5P study was approved by ethics committees of each participating institution, and a written informed consent and assent was obtained from all participants or parents as appropriate.

Study Protocol

Trained research assistants completed standardized assessments of all patients in the ED.[12] Data were collected and managed using Research Electronic Data Capture (REDCap) tools hosted at the Children's Hospital of Eastern Ontario Research Institute.[14] Patients and parents provided information on demographics and past history (ie, prior concussion, headache, developmental or psychiatric conditions), as well as injury characteristics using the Acute Concussion Evaluation inventory,[15] a validated scale used to identify concussion in children and adolescents aged 3 to 18 years. Patients and parents quantified pre-injury and current symptoms (ie, physical, emotional, cognitive, and sleep) using the Post-Concussion Symptom Inventory (PCSI).[16] Cognitive status, physical examination, and balance assessments were completed using the Child-Sport Concussion Assessment Tool-3rd Edition (Child-SCAT3) evaluation.[1,17]

Follow-up Procedures

Enrolled patients were offered web-based survey or telephone follow-up at 7 and 28 days postenrollment.[12] Patients received email reminders 24 hours after each survey deadline; research assistants telephoned nonresponders as many as 5 times and offered verbal interviews. Surveys were parent reported for children aged 5.00 to 7.99 years and patient reported for all other participants. Current level of physical activity was self-categorized as no activity (eg, physical rest), light aerobic exercise (eg, walking, swimming, or stationary cycling), sport-specific exercise (eg, running drills in soccer or skating drills in ice hockey), noncontact training drills (eg, complex passing drills), full-contact practice (eg, normal training activities), and return to competition (eg, normal game play).[12] Early physical activity participation was defined as any level of physical activity other than no activity at 7 days postenrollment. Early physical activity subcategories were defined as no activity, light aerobic exercise, moderate exercise (sport-specific exercise or noncontact training drills), or full exercise (full-contact practice or return to competition). Questions regarding physical activity were based on Zurich

Consensus Statement on Concussion in Sport return-to-play[1] steps; these questions have not been validated.

Primary Outcome Measure

Primary outcome was the presence of PPCS, defined as at least 3 new or worsening individual symptoms compared with the preconcussion status measured at day 28 according to the validated PCSI.[12,13,18] An individual symptom was defined as a positive difference between the current minus the perceived pre-injury symptom rating as completed 28 days postenrollment.[12,13]

Statistical Analysis

Frequencies and descriptive statistics were used to summarize patient baseline characteristics for the overall sample and by early physical activity. Missing data were managed via listwise deletion (ie, participants were excluded from the analysis if any single values were missing).

The proportion of PPCS in each group was computed, along with a Wilson score 95% CI, a method for obtaining a CI for a proportion.[19] The unadjusted association between early physical activity and PPCS was estimated using the sample relative risk (RR) and the sample Absolute risk difference (ARD).

Propensity scores[20] were developed to account for potential confounding by observed baseline characteristics.[21,22] A propensity score was derived to reflect the probability of a participant having engaged in early physical activity given an observed set of baseline characteristics. Propensity score methods replace an entire set of baseline characteristics with a single composite score, and this can be accomplished with numbers of potential confounders in excess of what is possible with conventional regression methods.[23-25] Clinically relevant variables (defined a priori) and those that may be associated with early physical activity were included in the models. Continuous variables were categorized based on the Youden index[26] or through visualization using locally weighted polynomial regression (LOESS) curves. The following variables were included as predictors of early activity using multivariate logistic regression to calculate the propensity score: age group, sex, duration of prior concussion (no prior concussion or concussion with symptoms lasting <1 week vs prior concussion with symptoms lasting ≥1 week), personal history of migraines, family history of migraines, learning disability, attention-deficit/hyperactivity disorder, developmental disorder, anxiety, depression, sleep disorder, other psychiatric disorder, loss of consciousness duration (did not lose consciousness or loss of consciousness <3 minutes vs loss of consciousness ≥3 minutes), time between head injury and triage, seizure, early symptoms on the Acute Concussion Evaluation (appears dazed and confused, confused about events, answers questions slowly, repeats questions, forgetful), balance tandem stance (0-3 errors vs ≥4 errors or physically unable), sports injury, all 20 parent reported indicators of the Postconcussion Symptom Inventory,[16] and site.

To examine the outcome associated with early activity,[21] participants who did and did not engage in early physical activity were matched 1:1 in random order on the logit of the propensity scores using a greedy algorithm and nearest-neighbor approach (maximum caliper distance, 0.1) using the MatchIt package in R (R Project for Statistical Computing).[27] Equivalence between matched participants (activity vs nonactivity groups) was assessed by testing for differences in covariates using χ^2 analyses and Mann-Whitney *U* tests where appropriate. Standardized mean differences were calculated using the R package Tableone. After obtaining a matched data set, the association between early participation in physical activity and PPCS was estimated using the sample RR and the sample ARD.

Inverse probability of treatment weighting (IPTW) was used to investigate the association of early participation in physical activity among the entire population of youth recovering from acute concussion when this population is hypothetically moved from no early activity to participation in early activity. Participants were weighted by the inverse of the probability of engaging in physical activity at day 7. The association between early participation in physical activity and PPCS was estimated using the RR obtained from log-binomial regression and the ARD obtained from identity link binomial regression. In both cases, the IPTW weights were used with a quasibinomial model to obtain robust variance estimates. To avoid convergence issues, the R package glm2 was used. Group differences were assessed by calculating IPTW proportions, weighted medians, and standardized mean differences.

Because the self-report questionnaire at day 7 does not differentiate between the timing of activity and symptoms within the first week postinjury, 2 sensitivity analyses were performed. First, the original analyses were repeated by replacing the total ED symptom load with the total score at day 7. Second, a subanalysis of only patients remaining symptomatic with at least 3 symptoms at day 7 was performed, thus excluding recovered patients and those with minimal symptomatology. A sensitivity analysis was also conducted to investigate a possible interaction between age and physical activity in the model for PPCS. A quasibinomial model with a log link included an effect for early exercise, age (as a continuous variable), and the product of these 2 variables. Two sided *P* values of less than 0.05 were considered statistically significant. All analyses were performed using IBM SPSS Statistics version 23 (IBM Corp) and R version 3.0.2.

Results

In total, 2584 of 3063 (84.4%) patients completed the primary outcome assessment (**Figure 1**). Of these, 171 were excluded because of missing data on participation in physical activity at day 7, resulting in a cohort of 2413 patients. Baseline characteristics for the total cohort, and for participant groups with and without early physical activity are summarized in **Table 1**. Overall, 733/2413 (30.4%) patients met criteria for PPCS.

Early Participation in Physical Activity

At 7 days postenrollment, 1677 (69.5%) patients reported participating in physical activity including light aerobic exercise

(795 [32.9%]), sport-specific exercise (214 [8.9%]), noncontact training drills (143 [5.9%]), full contact practice (106 [4.4%]), or return to competition (419 [17.4%]), and 736 patients (30.5%) reported no physical activity. Of the 1677 patients who engaged in early physical activity, 523 (31.3%) were symptom free and 803 (48.0%) had at least 3 persistent or worsening postconcussive symptoms at day 7. Of those reporting engaging in no physical activity at day 7, 584 (79.5%) had at least 3 persistent or worsening postconcussive symptoms at day 7.

Bivariable Analysis

In bivariable analysis (unweighted sample), early participation in any type of physical activity compared with no physical activity was associated with lower risk of PPCS (413 [24.6%] patients vs 320 [43.5%] patients; RR, 0.75 [95% CI, 0.70-0.80]; ARD, 18.9% [95% CI, 14.7%-23.0%]). When early physical activity subcategories were distinguished, participation in light aerobic exercise (250 patients [31.4%]; RR, 0.82 [95% CI, 0.76-0.89]; ARD, 12.0% [95% CI, 7.2%-16.8%]), moderate exercise (87 patients [24.4%]; RR, 0.75 [95% CI, 0.69-0.81]; ARD, 19.1% [95% CI, 13.2%-24.6%]), and full exercise (76 patients [14.5%]; RR, 0.66 [95% CI, 0.61-0.71]; ARD, 29.0% [95% CI, 24.2%-33.5%]) were all associated with significantly lower risk of PPCS as compared with the no activity group (320 patients [43.5%]; **Table 2**).

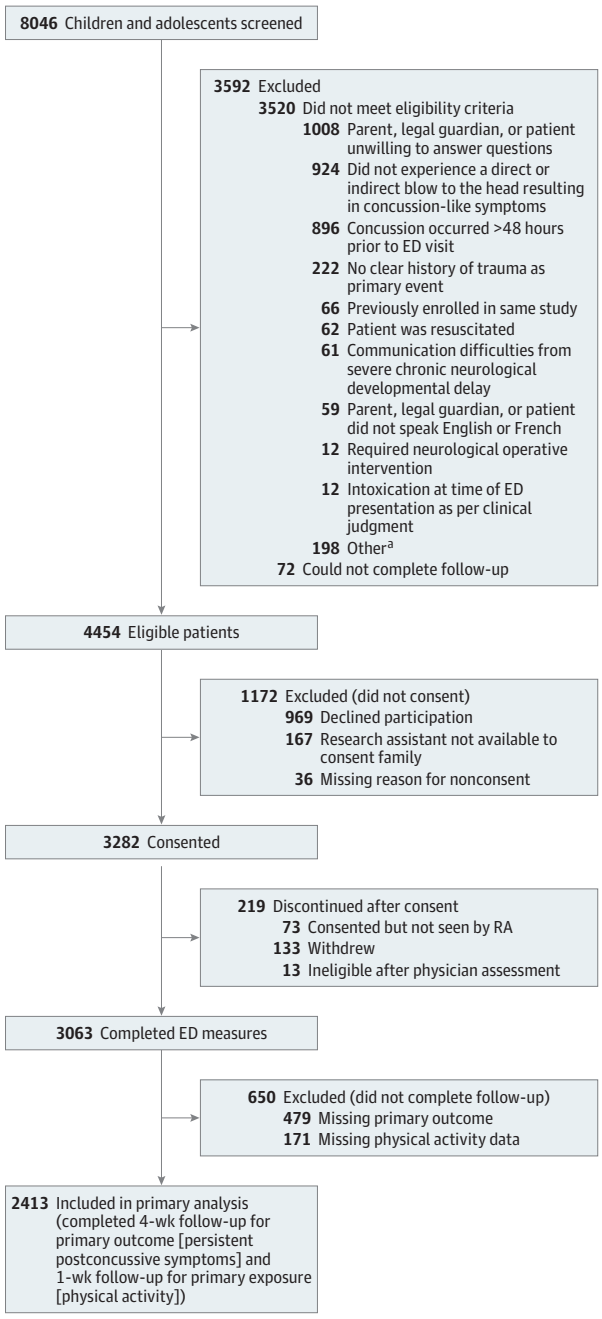
Propensity Score-Matched Analysis

Prior to matching, median propensity to engage in physical activity in the activity group was 0.74 (interquartile range [IQR], 0.65-0.80) vs 0.66 (IQR, 0.55-0.75) in the nonactivity group. Matching resulted in 554 children and adolescents participating in early physical activity matched to 554 children not participating in activity. Because more participants reported engaging in physical activity than not, 900 who participanted in physical activity were unmatched in contrast to 91 nonactivity participants. The distribution of propensity scores in the early activity and nonactivity groups are shown in before matching (**Figure 2**A) and after matching (Figure 2B). Following propensity score matching, mean (SD) propensity for early physical activity was similar for those reporting activity (0.641 [0.176]) vs no physical activity (0.627 [0.171]) and also resulted in between-group balance on baseline characteristics (Table 1). In propensity score–matched bivariable analysis (n = 1108), early participation in physical activity remained significantly associated with lower PPCS risk (n = 159 [28.7%]) vs n = 222 [40.1%]; RR, 0.84 [95% CI, 0.77-0.92]; ARD, 11.4% [95% CI, 5.8%-16.9%]).

Inverse Probability Treatment Weighting Analysis

IPTW, formed by those with complete data on exercise and all 43 covariates included in the propensity analysis, also resulted in between-group balance on baseline characteristics (Table 1; n = 2099). Figure 2C shows the weighted distribution of propensity scores in the early activity and nonactivity groups. Mean (SD) weight was 2.00 (1.52) with a range of 1.03 to 12.9. The highest values for standardized mean differences in the weighted data were 0.113 for wearing a mouth guard and 0.101 for wearing a helmet; all other baseline variables had standardized mean difference values of less than 0.1. PPCS remained significantly less likely in the early physical activity group in IPTW log-binomial regression analysis compared with the no physical activity group (RR, 0.74 [95% CI, 0.65-0.84]; ARD, 9.7% [95% CI, 5.7%-13.7%]).

Figure 1. Flow Diagram of Participants With and Without Early Physical Activity Following Acute Concussion



ED indicates emergency department; RA, research assistant.

[a] Includes those for whom reason was not specified or was missing data due to the fact that 1 of the 9 research ethics boards did not permit the collection of reasons for meeting exclusion criteria due to provincial regulations.

Table 1. Baseline Characteristics Total Sample, Unweighted Sample, Propensity Score-Matched Sample, and Inverse Probability of Treatment-Weighted Sample

Characteristic	Physical Activity at 7 Days, No. (%)									
		Unweighted Sample (n = 2413)			Propensity 1:1 Matching (n = 1108)			IPTW (n = 2099)[a,b]		
	Total (N = 3063)	No (n = 736)	Yes (n = 1677)	Standardized Mean Difference	No (n = 554)	Yes (n = 554)	Standardized Mean Difference	No (n = 645)	Yes (n = 1454)	Standardized Mean Difference
Age, y[c]										
5-7	534 (17.4)	73 (9.9)	343 (20.5)		69 (12.5)	79 (14.3)		352.5 (16.9)	379.4 (18.0)	
8-12	1282 (41.9)	268 (36.4)	762 (45.4)	0.434	226 (40.8)	232 (41.9)	0.068	935.9 (44.8)	906.1 (43.1)	0.040
13-18	1247 (40.7)	395 (53.7)	572 (34.1)		259 (46.8)	243 (43.9)		798.5 (38.3)	818.5 (38.9)	
Female sex[c]	1205 (39.3)	350 (47.6)	605 (36.1)	0.234	237 (42.8)	243 (43.9)	0.022	822.4 (39.4)	822.5 (39.1)	0.007
Time between head injury and triage, median (SD), h[c]	8.7 (11.8)	10.3 (12.8)	8.0 (11.0)	0.197	9.5 (12.1)	9.6 (12.1)	0.004	9.0 (11.7)	8.8 (11.6)	0.011
Previous concussion lasting ≥1 wk[c]	390 (12.8)	139 (19.0)	164 (9.8)	0.262	86 (15.5)	73 (13.2)	0.067	243.7 (11.7)	248.4 (11.8)	0.004
Personal migraine history	392 (12.9)	103 (14.1)	198 (11.9)	0.066	69 (12.5)	82 (14.8)	0.068	246.2 (11.8)	245.9 (11.7)	0.003
Family history of migraine[c]	1436 (48.2)	345 (48.3)	808 (49.2)	0.019	263 (47.5)	264 (47.7)	0.004	1000.2 (47.9)	1021.3 (48.5)	0.012
Learning disabilities[c]	243 (8.0)	53 (7.2)	124 (7.4)	0.007	42 (7.6)	42 (7.6)	<0.001	144.3 (6.9)	158.2 (7.5)	0.023
ADD or ADHD[c]	268 (8.7)	56 (7.7)	137 (8.2)	0.02	42 (7.6)	44 (7.9)	0.013	139.6 (6.7)	161.7 (7.7)	0.039
Other developmental disorder[c]	122 (4.0)	34 (4.6)	47 (2.8)	0.096	21 (3.8)	20 (3.6)	0.01	73.0 (3.5)	69.6 (3.3)	0.01
Anxiety[c]	237 (7.7)	85 (11.6)	111 (6.6)	0.173	54 (9.7)	47 (8.5)	0.044	160.5 (7.7)	170.0 (8.1)	0.014
Depression[c]	87 (2.9)	26 (3.5)	40 (2.4)	0.067	18 (3.2)	16 (2.9)	0.021	51.3 (2.5)	49.8 (2.4)	0.006
Sleep disorders[c]	62 (2.0)	22 (3.0)	27 (1.6)	0.092	15 (2.7)	13 (2.3)	0.023	45.7 (2.2)	51.4 (2.4)	0.017
Other psychiatric disorder[c]	32 (1.1)	9 (1.2)	17 (1.0)	0.019	5 (0.9)	2 (0.4)	0.068	15.1 (0.7)	17.3 (0.8)	0.011
Loss of consciousness duration, mean (SD)[c]										
Did not lose consciousness	2317 (76.0)	552 (75.0)	1274 (76.1)		427 (77.1)	420 (75.8)		1592.9 (76.3)	1603.2 (76.2)	
Lost consciousness, min				0.042			0.058			0.041
<3	395 (13.0)	101 (13.7)	206 (12.3)		71 (12.8)	68 (12.3)		275.3 (13.2)	257.8 (12.3)	
≥3	337 (11.1)	83 (11.3)	194 (11.6)		56 (10.1)	66 (11.9)		218.7 (10.5)	242.9 (11.5)	
Seizure[c]	57 (1.9)	13 (1.8)	32 (1.9)	0.011	11 (2.0)	5 (0.9)	0.091	40.8 (2.0)	40.1 (1.9)	0.003

(continued)

Table 1. Baseline Characteristics Total Sample, Unweighted Sample, Propensity Score-Matched Sample, and Inverse Probability of Treatment-Weighted Sample (continued)

Characteristic	Physical Activity at 7 Days, No. (%)									
		Unweighted Sample (n = 2413)			Propensity 1:1 Matching (n = 1108)			IPTW (n = 2099)[a,b]		
	Total (N = 3063)	No (n = 736)	Yes (n = 1677)	Standardized Mean Difference	No (n = 554)	Yes (n = 554)	Standardized Mean Difference	No (n = 645)	Yes (n = 1454)	Standardized Mean Difference
ACE[c]										
Appears dazed and confused	1504 (49.1)	425 (57.7)	773 (46.1)	0.235	312 (56.3)	290 (52.3)	0.08	1029.6 (49.3)	1058.3 (50.3)	0.019
Confused about events	755 (24.6)	211 (28.7)	387 (23.1)	0.128	151 (27.3)	135 (24.4)	0.066	491.6 (23.6)	520.7 (24.7)	0.028
Answers questions slowly	1253 (40.9)	353 (48.0)	618 (36.9)	0.226	254 (45.8)	237 (42.8)	0.062	859.3 (41.2)	844.8 (40.2)	0.021
Repeats questions	418 (13.6)	120 (16.3)	207 (12.3)	0.113	88 (15.9)	69 (12.5)	0.098	296.7 (14.2)	288.1 (13.7)	0.015
Forgetful	643 (21.0)	184 (25.0)	319 (19.0)	0.145	132 (23.8)	122 (22.0)	0.043	445.7 (21.4)	441.6 (21.0)	0.009
Sports injury[c]	2071 (67.6)	539 (73.2)	1119 (66.8)	0.141	400 (72.2)	394 (71.1)	0.024	1441.6 (69.1)	1452.9 (69.1)	0.001
Balance tandem stance[c]	1206 (40.6)	304 (42.0)	650 (39.7)	0.047	223 (40.3)	217 (39.2)	0.022	835.7 (40.0)	826.8 (39.3)	0.015
Mechanism of injury										
Sports and recreational play	2071 (67.6)	539 (73.2)	1119 (66.8)	0.183	400 (72.2)	394 (71.1)	0.164	1441.6 (69.1)	1452.9 (69.1)	0.123
Non-sports-related injury and fall	741 (24.2)	139 (18.9)	429 (25.6)		108 (19.5)	129 (23.3)		453.2 (21.7)	501.0 (23.8)	
Motor vehicle collision	55 (1.8)	18 (2.4)	28 (1.7)		13 (2.3)	9 (1.6)		51.1 (2.4)	30.7 (1.5)	
Assault	39 (1.3)	11 (1.5)	17 (1.0)		11 (2.0)	3 (0.5)		39.5 (1.9)	17.7 (0.8)	
Other	143 (4.7)	29 (3.9)	82 (4.9)		22 (4.0)	19 (3.4)		101.5 (4.9)	101.6 (4.8)	
Helmet use[c]	779 (37.6)	217 (40.3)	407 (36.4)	0.08	164 (41.0)	145 (36.8)	0.086	605.1 (42.0)	538.4 (37.1)	0.101
Mouth guard use[c]	448 (21.7)	134 (24.9)	222 (19.9)	0.12	104 (26.1)	87 (22.2)	0.091	368.3 (25.6)	302.1 (20.9)	0.113
Parent report indicators of PPCS[c,d]										
Headache	2517 (86.8)	638 (90.4)	1359 (84.9)	0.166	497 (89.7)	495 (89.4)	0.012	1844.5 (88.4)	1829.5 (87.0)	0.044
Nausea	1702 (58.8)	442 (62.6)	904 (56.6)	0.123	338 (61.0)	336 (60.6)	0.007	1241.9 (59.5)	1225.5 (58.2)	0.026
Balance problems	1265 (43.7)	363 (51.4)	660 (41.3)	0.203	278 (50.2)	262 (47.3)	0.058	934.0 (44.8)	936.2 (44.5)	0.005
Dizziness	2032 (70.2)	540 (76.5)	1069 (66.9)	0.214	413 (74.5)	411 (74.2)	0.008	1463.9 (70.2)	1471.9 (70.0)	0.004
Drowsiness	2127 (73.4)	531 (75.2)	1142 (71.5)	0.085	411 (74.2)	397 (71.7)	0.057	1557.0 (74.6)	1546.3 (73.5)	0.025
Increased sleeping	1007 (34.8)	264 (37.4)	536 (33.6)	0.081	196 (35.4)	180 (32.5)	0.061	763.5 (36.6)	737.8 (35.1)	0.032
Sensitivity to light	1136 (39.2)	326 (46.2)	578 (36.2)	0.206	242 (43.7)	236 (42.6)	0.022	839.1 (40.2)	828.9 (39.4)	0.017
Sensitivity to noise	1033 (35.7)	316 (44.8)	496 (31.0)	0.286	234 (42.2)	234 (42.2)	<0.001	754.5 (36.2)	757.8 (36.0)	0.003
Irritability	778 (26.9)	211 (29.9)	391 (24.5)	0.122	156 (28.2)	148 (6.7)	0.032	553.3 (26.5)	551.3 (26.2)	0.007
Sadness	1152 (39.8)	260 (36.8)	643 (40.2)	0.07	210 (37.9)	201 (36.3)	0.034	807.1 (38.7)	813.8 (38.7)	<0.001

(continued)

Table 1. Baseline Characteristics Total Sample, Unweighted Sample, Propensity Score-Matched Sample, and Inverse Probability of Treatment-Weighted Sample (continued)

Characteristic	Physical Activity at 7 Days, No. (%)									
		Unweighted Sample (n = 2413)			Propensity 1:1 Matching (n = 1108)			IPTW (n = 2099)[a,b]		
	Total (N = 3063)	No (n = 736)	Yes (n = 1677)	Standardized Mean Difference	No (n = 554)	Yes (n = 554)	Standardized Mean Difference	No (n = 645)	Yes (n = 1454)	Standardized Mean Difference
Nervousness	720 (24.9)	167 (23.7)	397 (24.8)	0.028	126 (22.7)	121 (21.8)	0.022	486.8 (23.3)	505.6 (24.0)	0.017
Acts more emotional	1153 (39.8)	260 (36.9)	648 (40.5)	0.074	211 (38.1)	195 (35.2)	0.06	850.8 (40.8)	832.2 (39.6)	0.025
Seems mentally foggy	1546 (53.5)	434 (61.5)	792 (49.7)	0.239	326 (58.8)	316 (57.0)	0.037	1129.4 (54.1)	1129.3 (53.7)	0.009
Poor concentration	1075 (37.2)	322 (45.7)	529 (33.1)	0.259	230 (41.5)	214 (38.6)	0.059	782.7 (37.5)	778.3 (37.0)	0.011
Forgetfulness	866 (29.9)	247 (35.0)	440 (27.5)	0.163	177 (31.9)	178 (32.1)	0.004	629.3 (30.2)	621.9 (29.6)	0.013
Visual problems	979 (33.8)	273 (38.7)	511 (32.0)	0.141	203 (36.6)	196 (35.4)	0.026	706.5 (33.9)	713.9 (33.9)	0.002
Increased fatigue	2160 (74.6)	567 (80.4)	1141 (71.4)	0.213	432 (78.0)	424 (76.5)	0.034	1574.6 (75.5)	1563.8 (74.3)	0.026
Confusion with directions or tasks	682 (23.6)	205 (29.2)	319 (19.9)	0.215	147 (26.5)	135 (24.4)	0.05	459.0 (22.0)	470.5 (22.4)	0.009
Clumsy movement	790 (27.3)	210 (29.8)	416 (26.0)	0.084	160 (28.9)	162 (29.2)	0.008	566.6 (27.2)	570.8 (27.1)	<0.001
Slower response to questions	1363 (47.1)	371 (52.6)	693 (43.4)	0.186	283 (51.1)	273 (49.3)	0.036	988.4 (47.4)	990.7 (47.1)	0.005
Site[c]										
1		78 (10.6)	227 (13.5)		61 (11.0)	59 (10.6)		246.1 (11.8)	268.1 (12.7)	
2		144 (19.6)	376 (22.4)		106 (19.1)	121 (21.8)		431.3 (20.7)	416.6 (19.8)	
3		35 (4.8)	121 (7.2)		32 (5.8)	31 (5.6)		133.8 (6.4)	147.1 (7.0)	
4		72 (9.8)	206 (12.3)		43 (7.8)	48 (8.7)		201.1 (9.6)	198.8 (9.4)	
5		112 (15.2)	88 (5.2)	0.508	71 (12.8)	61 (11.0)	0.096	185.9 (8.9)	180.9 (8.6)	0.047
6		54 (7.3)	156 (9.3)		47 (8.5)	45 (8.1)		197.4 (9.5)	191.1 (9.1)	
7		144 (19.6)	153 (9.1)		108 (19.5)	100 (18.1)		285.1 (13.7)	293.7 (14.0)	
8		49 (6.7)	200 (11.9)		43 (7.8)	43 (7.8)		218.2 (10.5)	226.3 (10.8)	
9		48 (6.5)	150 (8.9)		43 (7.8)	46 (8.3)		187.9 (9.0)	181.3 (8.6)	

Abbreviations: ACE, acute concussion evaluation; ADD, attention-deficit disorder; ADHD, attention deficit/hyperactivity disorder; IPTW, inverse probability of treatment weighting; PPCS, persistent postconcussive symptoms.

[a] Proportions and medians are weighted using IPTW.

[b] The IPTW sample only includes those with complete data on exercise and all 43 covariates included in the propensity analysis.

[c] Variables included in the propensity score.

[d] Indicators are based on patient complaint, parental observation of the patient, or a combination of both.

Sensitivity Analyses

When the total ED symptom score was replaced with the total score at day 7, only the association in the unweighted sample remained significant with similar magnitude and directionality as in the primary analyses (bivariable analyses RR, 0.75 [95% CI, 0.70-0.80]; **Table 3**).

In the second sensitivity analysis, the analytical sample was limited to children and adolescents with at least 3 symptoms at day 7 (n = 1387). Despite current guidelines strongly advocating physical rest until the patient is asymptomatic, 584 of 1387 (57.9%) participants engaged in some form of physical activity (ie, were nonadherent with current recommendations). Although the directionality of the association remained similar, the propensity score–matched analysis and IPTW analysis no longer reached statistical significance (bivariable analysis RR, 0.83 [95% CI, 0.74-0.92]); propensity score–matched RR, 0.92 [95% CI, 0.80-1.05]; IPTW RR, 0.92 [95% CI, 0.82-1.04]; Table 3). When early physical activity subcategories were distinguished within this symptomatic cohort, children and adolescents who participated in physical activity had lower risk of PPCS (light aerobic exercise absolute risk, 46.4% [RR, 0.88 {95% CI, 0.78-0.99}; ARD, 6.6% {95% CI, 0.6%-12.5%}]; moderate exercise absolute risk, 38.6% [RR, 0.77 {95% CI, 0.66-0.89}; ARD, 14.3% {95% CI, 5.9%-22.2%}]; and full exercise absolute risk, 36.1% [RR, 0.74 {95% CI, 0.63-0.86}; ARD, 16.8% {95% CI, 7.5%-25.5%}]) compared with the no activity group (absolute risk, 52.9%). Finally, there was no statistically significant interaction between age and physical activity in an unadjusted model for PPCS; for each additional year of age, RR increased by a factor of 1.01 (95% CI, 0.97-1.05; P = .52).

Discussion

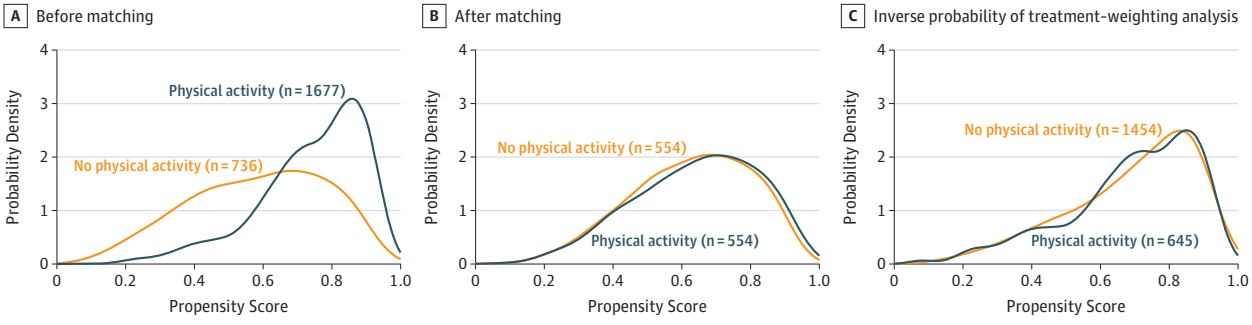
In this prospective cohort study, 69.5% of children and adolescents participated in physical activity within 7 days following an acute concussion—primarily with light aerobic exercise. The resumption of physical activity within 7 days postconcussion was associated with a lower risk of PPCS as compared with no physical activity. This finding was consistent across analytic approaches and intensity of exercise.

Evidence about the importance of physical activity in childhood for maintaining physical and cognitive health is unequivocal.[29] Physical activity is considered an effective method for improving cognitive function and brain health.[30] Compared with other conditions in which the latest insights regarding the benefits of early physical rehabilitation have been adopted, including stroke,[31] the field of pediatric concussion lags behind.[2] Overwhelming evidence supports the overall benefits of physical activity in youth [32] including better body composition,[33] skeletal health,[34] and cardiorespiratory fitness, as well as improvement of depression, anxiety, self-concept,[35] cognitive performance, and academic achievement.[36,37]

Table 2. Summary of Results of the Primary Analysis

Type Analysis	No. (Absolute Risk, %)		Absolute Risk Difference, % (95% CI)	Relative Risk (95%CI)
	Physical Activity	No Physical Activity		
Unweighted sample	1677 (24.6)	736 (43.5)	18.9 (14.7-23.0)	0.75 (0.70-0.80)
Light activity vs none (subgroup 1)	795 (31.4)	736 (43.5)	12.0 (7.2-16.8)	0.82 (0.76-0.89)
Moderate activity vs none (subgroup 2)	357 (24.4)	736 (43.5)	19.1 (13.2-24.6)	0.75 (0.69-0.81)
Full-contact activity vs none (subgroup 3)	525 (14.5)	736 (43.5)	29.0 (24.2-33.5)	0.66 (0.61-0.71)
Matched	554 (28.7)	554 (40.1)	11.4 (5.8-16.9)	0.84 (0.77-0.92)
Inverse probability of treatment weighting	1454	645	9.7 (5.7-13.5)	0.74 (0.65-0.84)

Figure 2. Distribution of Propensity Scores in the Physical Activity Group and the Rest Group



Inverse probability of treatment-weighting analysis includes only those patients with complete data on physical activity and all 43 covariates included in the propensity analysis. For intervals along the x-axis, the area under the probability density curve represents the probability of those propensity scores. Smoothing was via the kernel density estimate.[28]

Table 3. Summary of Sensitivity Analysis 1 and 2

Type Analysis	No. (Absolute Risk, %)		Absolute Risk Difference (95% CI)	Relative Risk (95% CI)
	Physical Activity	No Physical Activity		
Sensitivity analysis 1				
Unweighted sample	1667 (30.4)	736 (69.6)	18.9 (14.7 to 23.0)	0.75 (0.70 to 0.80)
Matched	519 (39.1)	519 (38.3)	−0.77 (−6.7 to 5.1)	1.01 (0.92 to 1.11)
IPTW			−0.041 (−4.1 to 4.0)	1.00 (0.88 to 1.14)
Sensitivity analysis 2				
Unweighted sample	803 (43.0)	584 (52.9)	9.9 (4.6 to 15.2)	0.83 (0.74 to 0.92)
Subgroup 1 (light activity vs none)	494 (46.4)	584 (52.9)	6.6 (0.6 to 12.5)	0.88 (0.78 to 0.99)
Subgroup 2 (moderate activity vs none)	176 (38.6)	584 (52.9)	14.3 (5.9 to 22.2)	0.77 (0.66 to 0.89)
Subgroup 3 (full exercise vs none)	133 (36.1)	584 (52.9)	16.8 (7.5 to 25.5)	0.74 (0.63 to 0.86)
Matched	388 (47.2)	388 (51.5)	4.4 (−2.6 to 11.3)	0.92 (0.80 to 1.05)
IPTW	687	507	4.0 (−1.7 to 9.7)	0.92 (0.82 to 1.04)

Abbreviation: IPTW, inverse probability of treatment weighting.

Preliminary studies in concussed adolescents found that participants engaging in moderate levels of activity reported lower symptom levels and superior neurocognitive performance compared with those with physical rest, although the optimal timing for re-introducing physical activity remains undetermined.[6,11] Available evidence suggests that gradual resumption of physical activity should begin as soon as tolerated following an acute concussion,[3,11] with the exception of activities likely to increase the risk of re-injury.[6,9,11] Rest exceeding 3 days postinjury was similarly or less effective than treatment regimens allowing for earlier participation in physical activity following a concussion[11,38]; if prolonged, rest may predispose to secondary symptoms of fatigue, reactive depression, physiological deconditioning, and delayed recovery.[7,8] Also in symptomatic adolescents, pilot evidence suggests that gradual resumption of aerobic physical activities results in superior symptom recovery from concussion compared with complete rest.[9,10]

A proposed mechanism by which exercise may improve recovery is through the promotion of neuroplasticity mechanisms and from possible effects on cardioregulatory mechanisms, possibly leading to improved cerebral blood flow.[39] This is of particular importance in pediatric concussion, since autoregulatory dysfunction and abnormal cerebral blood flow regulation have been associated with PPCS in school-aged children.[40,41] Controlled aerobic exercise may improve recovery by restoring normal cerebral blood flow regulation[10] with the rate of symptom improvement relating directly to the exercise intensity achieved.[10] Conversely, physical inactivity may predispose patients to PPCS through an activity restriction cascade model; it has been theorized that the psychological consequences of removal from life-validating activities, combined with physical deconditioning, may contribute to the development of PPCS after mild traumatic brain injury in youth.[3]

The results of this study should be considered in the context of study limitations. Because of the observational design, the authors cannot account for unmeasured confounding due to factors that may have been associated with physical activity shortly after concussion, nor can causation be determined. Although potential confounding by observed baseline characteristics was accounted for by conducting a propensity analysis,[20] unmeasured confounders and intermediaries may have influenced the results. Because the lowest odds of PPCS were observed in children participating in full exercise at day 7, children who simply felt better may have started physical activity earlier and subsequently resumed full competition despite still having symptoms. This possibility was examined through sensitivity analyses in which 1-week symptoms replaced ED symptoms and the inclusion of only those children with 3 or more symptoms at day 7. Given the limitation of possible confounding variables, a well-designed and adequately powered randomized clinical trial is needed to confirm the benefits of early return to physical activity.

Second, physical activity was rated via self-report questionnaires. Although direct measures of physical activity have greater precision, no single criterion standard exists and self-rated measures remain the most common and feasible method of measuring physical activity in large settings due to their practicality, low cost, low participant burden, and general acceptance.[42]

Third, because objective data on physical activity (eg, actigraphy) was not collected, information regarding duration and frequency of physical activity is limited; uncertainty remains as to whether exercise intensity exacerbates symptoms and how total activity load may be associated with PPCS risk.

Fourth, it is possible that patients who did not resume physical activity may have participated in more cognitive activity and therefore may have been potentially more symptomatic. Measuring cognitive rest is challenging because it is poorly defined across the literature and difficult to objectively measure. Since objective cognitive activity data were not collected, conclusions regarding the benefit or detriment of cognitive rest were unattainable.

Fifth, because all participants received care as usual by treating physicians, rest and activity recommendations likely varied across sites and clinicians both in the ED and subsequent follow-up.

Sixth, the influence of interim activity (eg, participation in physical activities between days 7 and 28 postenrollment) was not considered.

Divergence from conservative rest recommendations following pediatric concussion toward early active physical rehabilitation would be a new approach in concussion management, potentially affecting the well-being of millions of children and families worldwide. Early physical activity could mitigate the undesired effects of physical and mental deconditioning associated with prolonged rest. Regardless of potential benefit, caution in the immediate postinjury period is prudent; participation in activities that might introduce risk for collision (eg, resumption of contact sports) or falls (eg, skiing, skating, bicycling) should remain prohibited until clearance by a health professional to reduce the risk for a potentially more serious second concussion during a period of increased vulnerability. To be noted, results of this study do not infer any evidence of benefit or harm in association with return to practice or play.

Conclusions

Among children and adolescents aged 5 to 18 years with acute concussion, participation in physical activity within 7 days of acute injury compared with no physical activity was associated with lower risk of PPCS at 28 days. A well-designed randomized clinical trial is needed to determine the benefits of early physical activity following concussion.

ARTICLE INFORMATION

Author Contributions: Dr Zemek had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Author Affiliations: Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada (Grool, Aglipay, Momoli, Barrowman, Ledoux); Sports Concussion Clinic, Boston Children's Hospital, Boston, Massachusetts (Meehan); Department of Pediatrics, Alberta Children's Hospital, Alberta Children's Hospital Research Institute, University of Calgary, Alberta, Canada (Freedman); Department of Psychology, Alberta Children's Hospital Research Institute, and Hotchkiss Brain Institute, University of Calgary, Alberta, Canada (Yeates); Department of Pediatrics, Hospital Ste Justine, University of Montreal, Montreal, Quebec, Canada (Gravel); Department of Pediatrics, Montreal Children's Hospital, McGill University Health Center, Montreal, Quebec, Canada (Gagnon); Department of Pediatrics, Hospital for Sick Children, Toronto, Ontario, Canada (Boutis); Sport Injury Prevention Research Centre, Faculty of Kinesiology, University of Calgary, Calgary, Alberta, Canada (Meeuwisse); Department of Pediatrics, Children's Hospital of Eastern Ontario, University of Ottawa, Ottawa, Ontario, Canada (Osmond, Zemek).

Concept and design: Grool, Aglipay, Meehan III, Freedman, Gravel, Gagnon, Boutis, Meeuwisse, Barrowman, Osmond, Zemek.
Acquisition, analysis, or interpretation of data: Grool, Aglipay, Momoli, Meehan III, Freedman, Yeates, Gravel, Gagnon, Boutis, Barrowman, Ledoux, Osmond, Zemek.
Drafting of the manuscript: Grool, Aglipay, Barrowman, Zemek.
Critical revision of the manuscript for important intellectual content: Grool, Aglipay, Momoli, Meehan III, Freedman, Yeates, Gravel, Gagnon, Boutis, Meeuwisse, Ledoux, Osmond, Zemek.
Statistical analysis: Aglipay, Momoli, Barrowman, Zemek.
Obtained funding: Meehan, Freedman, Gravel, Gagnon, Boutis, Meeuwisse, Osmond, Zemek.
Administrative, technical, or material support: Ledoux.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Meehan reports receipt of royalties from ABC-Clio and from Wolders Kluwer for sales of authored publications; contracts with ABC-Clio and Springer International for a future book; and research funding, in part, by a grant from the National Football League Players Association and by philanthropic support from the National Hockey League Alumni Association through the Corey C. Griffin Pro-Am Tournament. Dr Freedman reports receipt of support from the Alberta Children's Hospital Foundation Professorship in Child Health and Wellness. Dr Zemek reports receipt of support by the University of Ottawa Brain and Mind Research Institute as a clinical research chair in pediatric concussion. No other disclosures were reported.

Funding/Support: This study was supported by a Canadian Institutes of Health Research (CIHR) operating grant (MOP 126197); a CIHR–Ontario Neurotrauma Foundation Mild Traumatic Brain Injury team grant (TM1 127047); and CIHR planning grant (MRP 119829).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Group Information: Additional PERC Predicting Persistent Post-Concussive Problems in Pediatrics (5P) Concussion Team members: Candice McGahern, BA (Children's Hospital of Eastern Ontario, Ontario); Gurinder Sangha, MD (Department of Pediatrics, London Children's Hospital, London, Ontario); Darcy Beer, MD (Department of Pediatrics, Manitoba Children's Hospital; Winnipeg, Manitoba); William Craig, MDCM (Department of Pediatrics, Stollery Children's Hospital, Edmonton, Alberta); Emma Burns, MD (Department of Pediatrics, IWK Health Centre, Halifax, Nova Scotia); Ken J. Farion, MD (Department of Pediatrics; Children's Hospital of Eastern Ontario); Angelo Mikrogianakis, MD (Department of Pediatrics, Alberta Children's Hospital); Karen Barlow, MD (Department of Pediatrics and Clinical Neurosciences, Alberta's Children's Hospital, Calgary, Alberta); Alexander S. Dubrovsky, MDCM, MSc (Department of Pediatrics, Montreal Children's Hospital, Montreal, Québec;); Gerard Gioia, PhD (Department of Neuropsychology, Children's National Health System, George Washington University School of Medicine, Rockville, Maryland); Miriam H. Beauchamp, PhD (Ste Justine Research Center, University of Montreal, Montreal, Quebec); Yael Kamil, BSc (Children's Hospital of Eastern Ontario); Blaine Hoshizaki, PhD (Department of Kinesiology, University of Ottawa); Peter Anderson, PhD (Department of Psychology, Children's Hospital of Eastern Ontario); Brian L. Brooks, PhD (Alberta Children's Hospital Research Institute, Calgary, Alberta); Michael Vassilyadi, MDCM, MSc (Department of Neurosurgery, Children's Hospital of Eastern Ontario); Terry Klassen, MD (Department of Pediatrics, Manitoba Children's Hospital, Winnipeg, Manitoba); Michelle Keightley, PhD (Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, Ontario); Lawrence Richer, MD (Department of Neurology, Stollery Children's Hospital, Edmonton, Alberta); Carol DeMatteo, MSc (School of Rehabilitation Science, McMaster University, Hamilton, Ontario). Mss McGahern and Kamil received compensation in association with their contributions to this article. None of the other aforementioned individuals received compensation.

Previous Presentation: Preliminary results of this study were presented as an abstract at the 11th Congress on Brain Injury, The Hague, the Netherlands, March 2016; the Pediatric Academic Societies Meeting, Baltimore, Maryland, May 2016; and the Fifth International Consensus Conference on Concussion in Sport, Berlin, Germany, October 2016.

Disclaimer: This is a substudy of the Predicting and Preventing Postconcussive Problems in Pediatrics (5P) study (recently published in *JAMA*), and while it includes necessary overlap in baseline patient characteristic data, the primary and secondary outcomes are unique.

Additional Contributions: We thank the parents and children who enrolled in this study and acknowledge the research coordinators and research assistants across the 9 sites responsible for patient recruitment, enrollment, and follow-up. Student volunteers at Children's Hospital of Eastern Ontario, Alberta Children's Hospital, Le Centre hospitalier universitaire Sainte-Justine (CHUSJ), and the Hospital for Sick Children provided assistance in patient screening at the emergency department. We appreciate the collaboration and assistance of all the treating physicians of the emergency departments across the sites. We are grateful to Pediatric Emergency Research Canada for making this study possible. None of those mentioned were compensated for their contribution.

REFERENCES

1. McCrory P, Meeuwisse WH, Aubry M, et al. Consensus statement on concussion in sport. *Br J Sports Med*. 2013;47(5):250-258.

2. Zemek R, Duval S, Dematteo C. *Guidelines for Diagnosing and Managing Pediatric Concussion*. Toronto, Canada: Ontario Neurotrauma Foundation; 2014.

3. DiFazio M, Silverberg ND, Kirkwood MW, Bernier R, Iverson GL. Prolonged activity restriction after concussion. *Clin Pediatr (Phila)*. 2016;55(5):443-451.

4. Cancelliere C, Hincapié CA, Keightley M, et al. Systematic review of prognosis and return to play after sport concussion. *Arch Phys Med Rehabil*. 2014;95(3)(suppl):S210-S229.

5. Schneider KJ, Iverson GL, Emery CA, McCrory P, Herring SA, Meeuwisse WH. The effects of rest and treatment following sport-related concussion. *Br J Sports Med*. 2013;47(5):304-307.

6. Majerske CW, Mihalik JP, Ren D, et al. Concussion in sports. *J Athl Train*. 2008;43(3):265-274.

7. Thomas DG, Apps JN, Hoffmann RG, McCrea M, Hammeke T. Benefits of strict rest after acute concussion. *Pediatrics*. 2015;135(2):213-223.

8. Willer B, Leddy JJ. Management of concussion and post-concussion syndrome. *Curr Treat Options Neurol*. 2006;8(5):415-426.

9. Gagnon I, Grilli L, Friedman D, Iverson GL. A pilot study of active rehabilitation for adolescents who are slow to recover from sport-related concussion. *Scand J Med Sci Sports*. 2016;26(3):299-306.

10. Leddy JJ, Kozlowski K, Donnelly JP, Pendergast DR, Epstein LH, Willer B. A preliminary study of subsymptom threshold exercise training for refractory post-concussion syndrome. *Clin J Sport Med*. 2010;20(1):21-27.

11. Silverberg ND, Iverson GL. Is rest after concussion "the best medicine?". *J Head Trauma Rehabil*. 2013;28(4):250-259.

12. Zemek R, Osmond MH, Barrowman N, et al. Predicting and preventing postconcussive problems in paediatrics (5P) study. *BMJ Open*. 2013;3(8):e003550.

13. Zemek R, Barrowman N, Freedman SB, et al; Pediatric Emergency Research Canada (PERC) Concussion Team. Clinical risk score for persistent postconcussion symptoms among children with acute concussion in the ED. *JAMA*. 2016;315(10):1014-1025.

14. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381.

15. Gioia GA, Collins M, Isquith PK. Improving identification and diagnosis of mild traumatic brain injury with evidence. *J Head Trauma Rehabil*. 2008;23(4):230-242.

16. Sady MD, Vaughan CG, Gioia GA. Psychometric characteristics of the postconcussion symptom inventory in children and adolescents. *Arch Clin Neuropsychol*. 2014;29(4):348-363.

17. McCrory P, Meeuwisse WH, Aubry M, et al. Child SCAT3. *Br J Sports Med*. 2013;47(5):263-266.

18. Steindel SJ. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J Am Med Inform Assoc*. 2010;17(3):274-282.

19. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med*. 1998;17(8):857-872.

20. Haukoos JS, Lewis RJ. The propensity score. *JAMA*. 2015;314(15):1637-1638.

21. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399-424.

22. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*. 2006;163(3):262-270.

23. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156.

24. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.1093/biomet/70.1.41

25. Williamson EJ, Forbes A. Introduction to propensity scores. *Respirology*. 2014;19(5):625-635.

26. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35.

27. Rosenbaum P, Rubin D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39:33-38. doi:10.1017/CBO9780511810725.019

28. Scott D. *Multivariate Density Estimation: Theory, Practice and Visualization. Wiley Series in Probability and Mathematical Statistics*. New York, NY: John Wiley & Sons; 1992.

29. Khan NA, Hillman CH. The relation of childhood physical activity and aerobic fitness to brain function and cognition. *Pediatr Exerc Sci*. 2014;26(2):138-146.

30. Hillman CH, Erickson KI, Kramer AF. Be smart, exercise your heart. *Nat Rev Neurosci*. 2008;9(1):58-65.

31. Veerbeek JM, van Wegen E, van Peppen R, et al. What is the evidence for physical therapy poststroke? *PLoS One*. 2014;9(2):e87987.

32. Longmuir PE, Colley RC, Wherley VA, Tremblay MS. Canadian Society for Exercise Physiology position stand: benefit and risk for promoting childhood physical activity. *Appl Physiol Nutr Metab*. 2014;39(11):1271-1279.

33. Rush E, Simmons D. Physical activity in children. *Med Sport Sci*. 2014;60:113-121.

34. McKay HA, Petit MA, Schutz RW, Prior JC, Barr SI, Khan KM. Augmented trochanteric bone mineral density after modified physical education classes. *J Pediatr*. 2000;136(2):156-162.

35. Annesi JJ. Correlations of depression and total mood disturbance with physical activity and self-concept in preadolescents enrolled in an after-school exercise program. *Psychol Rep*. 2005;96(3 Pt 2):891-898.

36. Howie EK, Pate RR. Physical activity and academic achievement in children. *J Sport Health Sci*. 2012;1(3):160-169. doi:10.1016/j.jshs.2012.09.003

37. Chaddock-Heyman L, Hillman CH, Cohen NJ, Kramer AF III. The importance of physical activity and aerobic fitness for cognitive control and memory in children. *Monogr Soc Res Child Dev*. 2014;79(4):25-50.

38. Moor HM, Eisenhauer RC, Killian KD, et al. The relationship between adherence behaviors and recovery time in adolescents after a sports-related concussion. *Int J Sports Phys Ther*. 2015;10(2):225-233.

39. Perrey S. Promoting motor function by exercising the brain. *Brain Sci*. 2013;3(1):101-122.

40. Farquhar WB, Greaney JL. Autonomic exercise physiology in health and disease. *Auton Neurosci*. 2015;188:1-2.

41. Leddy JJ, Kozlowski K, Fung M, Pendergast DR, Willer B. Regulatory and autoregulatory physiological dysfunction as a primary characteristic of post concussion syndrome. *NeuroRehabilitation*. 2007;22(3):199-205.

42. Adamo KB, Prince SA, Tricco AC, Connor-Gorber S, Tremblay M. A comparison of indirect versus direct measures for assessing physical activity in the pediatric population. *Int J Pediatr Obes*. 2009;4(1):2-27.

DOI: 10.1377/hlthaff.2019.00466
HEALTH AFFAIRS 39,
NO. 1 (2020): 58–66

By Amol S. Navathe, Ezekiel J. Emanuel, Atheendar S. Venkataramani, Qian Huang, Atul Gupta, Claire T. Dinh, Eric Z. Shan, Dylan Small, Norma B. Coe, Erkuan Wang, Xinshuo Ma, Jingsan Zhu, Deborah S. Cousins, and Joshua M. Liao

Spending And Quality After Three Years Of Medicare's Voluntary Bundled Payment For Joint Replacement Surgery

Amol S. Navathe (amol.navathe@gmail.com) is a core investigator at the Corporal Michael J. Cresencz Veterans Affairs Medical Center, in Philadelphia, and an assistant professor in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, and a senior fellow at the Leonard Davis Institute of Health Economics, both at the University of Pennsylvania.

Ezekiel J. Emanuel is the Diane V. S. Levy and Robert M. Levy University Professor, chair of the Department of Medical Ethics and Health Policy, and vice provost for global initiatives, all at the University of Pennsylvania.

Atheendar S. Venkataramani is an assistant professor of medical ethics and of health policy at the Perelman School of Medicine, University of Pennsylvania.

Qian Huang is a statistical analyst in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Atul Gupta is an assistant professor in the Department of Health Care Management, Wharton School, University of Pennsylvania.

ABSTRACT **Medicare has reinforced its commitment to voluntary bundled payment by building upon the Bundled Payments for Care Improvement (BPCI) initiative via an ongoing successor program, the BPCI Advanced Model. Although lower extremity joint replacement (LEJR) is the highest-volume episode in both BPCI and BPCI Advanced, there is a paucity of independent evidence about its long-term impact on outcomes and about whether improvements vary by timing of participation or arise from patient selection rather than changes in clinical practice. We found that over three years, compared to no participation, participation in BPCI was associated with a 1.6 percent differential decrease in average LEJR episode spending with no differential changes in quality, driven by early participants. Patient selection accounted for 27 percent of episode savings. Our findings have important policy implications in view of BPCI Advanced and its two participation waves.**

Bundled payment has the potential to improve the value of care by holding hospitals accountable for episode-specific quality and costs. For example, participation in bundled payment programs from the Centers for Medicare and Medicaid Services (CMS) has been associated with financial savings and stable quality for lower extremity joint replacement (LEJR) surgery episodes.[1]

CMS has demonstrated particular commitment to voluntary bundled payment. In 2013 the agency scaled voluntary bundled payment for LEJR nationwide via the Bundled Payments for Care Improvement (BPCI) initiative.[1] One peer-reviewed evaluation conducted by CMS and its contractors estimated that there had been a 3.9 percent reduction in LEJR episode spending without adverse effects on quality after one year of BPCI.[2] This finding was further corroborated at longer durations by subsequent non-peer-reviewed evaluations conducted by CMS contractors.[3,4] Consequently, CMS elected to continue voluntary bundled payment in October 2018 through the national five-year BPCI Advanced model, in which participating hospitals and physician groups accept bundled payment for up to thirty-seven clinical episodes that correspond to various surgical procedures and medical conditions. BPCI Advanced is the only new advanced alternative payment model established under the administration of President Donald Trump. LEJR is the most commonly selected episode in BPCI Advanced, which has already enrolled nearly 1,300 participants in the first of its two participation waves.[5]

Despite growing participation in voluntary bundled payment for LEJR, two policy salient issues remain unaddressed. First, to our knowledge, there are no studies of BPCI's nationwide impact beyond evidence produced by CMS and its contractors, which has resulted in an overall

paucity of evidence about the long-term impact of BPCI on spending and patient outcomes—particularly based on the timing of organizations' entry into BPCI. This issue is relevant to BPCI Advanced because the program enrolls organizations over time in two waves, separated by fifteen months. Furthermore, evaluations of other payment reforms suggest that benefits can require several years to emerge.[6,7] However, no independent peer-reviewed evaluations have addressed the effects of voluntary bundled payment for LEJR over multiple years of follow-up, and none have examined the timing of participation and benefits as key policy concerns.

Second, voluntary programs such as BPCI and BPCI Advanced are susceptible to patient selection—that is, changes in the case-mix of patients cared for after organizations begin program participation. Consequently, evaluations of voluntary bundled payment must account for the possibility that cost savings associated with participation may arise from the selection of healthier patients rather than practice improvements.[8] Unfortunately, existing evaluations have been limited in accounting for such selection.

Given these critical policy issues, we used a novel analytic approach to address patient selection while conducting the first independent multiyear evaluation of the association between timing of participation in voluntary bundled payment for LEJR and changes in episode spending and quality.

Study Data And Methods

STUDY PERIODS We used January 2011–September 2013 as the baseline (pre-BPCI) period and October 2013–December 2016 as the intervention (entire BPCI) period. To evaluate performance over time, we divided the intervention period into an early BPCI period (October 2013–June 2015) and a late BPCI period (July 2015–December 2016). In the late BPCI period, data through September 2016 were used to define LEJR episodes to allow for a ninety-day postdischarge period between October and December 2016.

MARKETS AND HOSPITALS BPCI consisted of four separate models of care. Among these, model 2 was the most popular and largest in terms of enrollment. Hospitals in model 2 accepted bundled payment for episodes beginning with hospitalization and spanning up to ninety days of postacute care. We identified hospitals that participated in LEJR bundled payment under model 2 ("BPCI hospitals") by compiling publicly available quarterly BPCI participant lists.[2] This allowed us to define BPCI participation per quarter, which reflected both the time-varying nature of program enrollment and the potential for entry into and exit from the program over time. Markets were defined by hospital referral region[9] and categorized as those that contained at least one BPCI hospital at any time in the program ("BPCI markets") or those that contained no BPCI hospitals ("non-BPCI markets"). Our comparison group of hospitals consisted of those that never participated in model 2 and were located in non-BPCI markets ("non-BPCI hospitals").

To evaluate BPCI performance based on duration of participation, we divided BPCI hospitals into "early entrants" (those that began participating in the early BPCI period) and "late entrants" (those that began participating in the late BPCI period). We obtained hospital and market characteristics using data from Medicare claims, the American Hospital Association, Hospital Compare, and the 2017 CMS Improving Medicare Post-Acute Care Transformation (IMPACT) Act file, as done in prior work.[10–12]

PATIENT POPULATION AND EPISODE CONSTRUCTION We identified patients admitted to BPCI hospitals ("BPCI patients") and non-BPCI hospitals ("non-BPCI patients") for major hip and knee joint replacement or reattachment of lower extremity with and without major complicating or comorbid condition (Medicare Severity Diagnosis-Related Groups 469 and 470, respectively). We used a 100 percent sample of Medicare beneficiaries admitted to BPCI hospitals in 2011–16 and a 20 percent national sample of beneficiaries over the same period to identify patients admitted to non-BPCI hospitals. We excluded people not continuously enrolled in fee-for-service Medicare for the episode and a 180-day lookback period, those with end-stage renal disease, and those who died during the initial hospitalization for LEJR surgery. To increase sample homogeneity,[1,13] we also excluded patients younger than age sixty-five or older than age ninety and those using hospice.

Patient characteristics used to control for baseline health included age; sex; race; dual eligibility status (that is, eligibility for Medicare and Medicaid); Elixhauser comorbidities;[14] and prior use of an acute care hospital, skilled nursing facility (SNF), or inpatient rehabilitation facility (IRF). Consistent with prior literature,[10,13,15] LEJR episodes were constructed beginning with hospitalization for Medicare Severity Diagnosis-Related Group 469 or 470 and spanning ninety days after hospital discharge, using all Medicare data for inpatient, outpatient, SNF, and other claims. In line with BPCI program rules, we excluded overlapping or repeat episodes.[2] We also excluded episodes covered by the Comprehensive Care

Claire T. Dinh is a medical student at Harvard Medical School, in Boston, Massachusetts. She was a research coordinator in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania, when this work was completed.

Eric Z. Shan is a research assistant in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Dylan Small is a professor in the Department of Statistics, University of Pennsylvania.

Norma B. Coe is an associate professor in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Erkuan Wang is a data analyst in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Xinshuo Ma is a data analyst in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Jingsan Zhu is associate director of data analytics in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Deborah S. Cousins is a project manager in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Joshua M. Liao is medical director of payment strategy, director of the Value and Systems Science Lab, and an assistant professor in the Department of Medicine, all at the University of Washington, in Seattle, and an adjunct senior fellow at the Leonard Davis Institute of Health Economics, University of Pennsylvania.

for Joint Replacement model.[16]

OUTCOMES Our primary spending outcome was the average Medicare payment per LEJR episode. Secondary spending outcomes included the proportion of average episode spending attributable to the index hospitalization, readmissions, SNF care, IRF care, home health agency services, professional services, and all remaining components. Spending estimates were standardized following methods used previously and transformed into 2016 dollars.[10–12,17,18] Quality outcomes included ninety-day postdischarge risk-standardized mortality, unplanned readmission, and emergency department visit rates, as well as LEJR-specific complication rates.[19]

STATISTICAL ANALYSES We compared patient, hospital, and market characteristics between BPCI and non-BPCI hospitals and markets. Chi-square tests were used to compare categorical variables, while *t*-tests and Wilcoxon rank-sum tests were used to compare continuous variables.

We conducted unadjusted analyses that compared changes in outcomes across study periods. In adjusted analyses we used a difference-in-differences method to estimate differential changes in spending and quality outcomes for BPCI hospitals compared to a 1:1 propensity-matched set of non-BPCI hospitals in the pre-BPCI and BPCI periods.[20,21] We used baseline hospital and market characteristics and matched BPCI and non-BPCI hospitals based on propensity score values with identical first two digits. Because hospitals that did not participate in BPCI but were located in BPCI markets could have been affected by the program (for example, as a result of changes in market share or referral patterns after it began), only non-BPCI hospitals in non-BPCI markets were used in propensity score matching (for details about our propensity score matching approach, see the appendix methods 1 section and appendix exhibit 1, both in the online appendix).[22]

Episode spending was analyzed using generalized linear models with a log link and gamma distribution, while secondary spending outcomes and binary quality outcomes were analyzed using ordinary least squares regression. In adjusted analyses all models controlled for patient demographics, clinical conditions (for a list of conditions used in statistical models, see appendix exhibit 2),[22] and time-varying market characteristics; included hospital and quarter fixed effects; and used bootstrapped standard errors.[23,24]

Evaluations of voluntary LEJR bundled payment are susceptible to selection (that is, changes in outcomes may arise from changes in the types of patients receiving LEJR under bundled payment rather than changes in care delivery processes at hospitals after participation starts), which could lead to spurious inferences about program-related savings and quality improvements. Such selection—in which bundled-payment participants shift toward lower-risk patients based partially on characteristics unobservable in claims data, and ostensibly to increase financial gain—has been previously described under BPCI.[25]

Such selection may occur for several reasons. For example, physicians who maintain operating privileges at multiple hospitals could preferentially "sort" lower-risk patients to BPCI hospitals and higher-risk patients to non-BPCI hospitals. Hospitals could also engage in strategies that drive preferential selection, such as practice acquisition or network expansion.

To address such selection and mitigate its effects, we used an instrumental variable together with our difference-in-differences method. The instrumental variable used historical hospital referral patterns before BPCI began, to identify whether in the absence of the BPCI program, a given patient would have been hospitalized for LEJR at a hospital that would later become a BPCI hospital versus one that would remain a non-BPCI hospital (for details about our instrumental variable approach, see the appendix methods 2 section and appendix exhibits 3 and 4).[22] Specifically, we used hospital referral patterns in 2011 to generate the predicted probability of hospitalization for LEJR at an eventual BPCI hospital.

In the first stage of a two-stage least squares instrumental variable regression, we used this predicted probability as an instrument for actual hospitalization for LEJR at a BPCI hospital. In the second stage, instrumented admission to a BPCI hospital was related to spending and quality outcomes. Though selection based on unobservable characteristics could also have occurred before BPCI, changes in patient selection as a result of subsequent BPCI participation could not have been influenced by hospital referral patterns in that period specifically. Therefore, this instrumental variable approach helped mitigate effects of any patient selection that occurred after hospitals began participating in BPCI based on characteristics unobservable in our data.

We evaluated changes in outcomes associated with BPCI participation for the BPCI period overall, as well as by early versus late BPCI periods and early- versus late-entrant hospitals. We also tested the parallel trends assumption. Statistical tests were two-tailed and considered significant at $\alpha = 0.05$. Robust standard errors were corrected for heteroscedasticity.[26]

Analyses were performed using SAS, version

9.4, or Stata, version 15.1. This study was approved by the University of Pennsylvania Institutional Review Board, with a waiver of informed consent.

SENSITIVITY ANALYSES We extended methods from prior CMS contractor evaluations to conduct analyses without an instrumental variable or hospital fixed effects.[1] To account for hospital practice changes that occurred in anticipation of BPCI participation, we also conducted analyses that excluded January–September 2013 from the baseline period. Additionally, we conducted analyses that used a less stringent optimal propensity score matching approach and a caliper of 0.05 that varied the match ratio from 1:1 to 1:3 (for details about our propensity score matching approach, see the appendix methods 1 section in the online appendix),[22] as well as analyses that excluded hospitals that participated in the Comprehensive Care for Joint Replacement model from propensity score matching and subsequent analysis.

LIMITATIONS This study had several limitations. First, it was an observational study, and thus our results could be confounded by omitted variables and both patient and hospital selection. However, these concerns were mitigated by the quasi-experimental design that incorporated a robust set of patient, hospital, and market characteristics. In contrast to prior BPCI evaluations, this analysis also incorporated time-varying participation (that is, hospitals served as controls prior to BPCI participation), hospital fixed effects to address hospital selection based on time-invariant unobservable factors, and an instrumental variable to reduce potential confounding from patient selection.

Second, results from our instrumental variable approach applied only to beneficiaries who received LEJR at BPCI hospitals regardless of the BPCI model, not to all beneficiaries who received LEJR.

Third, while our use of an instrumental variable accounted for unobserved selection in ways that prior studies did not, more work is needed to ensure that all sources of selection are accounted for in policy evaluations. This is particularly true because our analytic approach did not account for all forms of selection—such as the selection of healthier patients for LEJR who would not have received the procedure without bundled payment (that is, in the event that BPCI hospitals induced demand for LEJR).

Fourth, secondary outcomes were analyzed with ordinary least squares regression, in part because of the need to include large numbers of fixed effects. While changes in mean estimates rather than predicted values were of interest, this approach might not have accounted for the skewness of data.

Fifth, results might not be generalizable to medical condition episodes or episodes initiated by physician group practices.

Sixth, this study focused only on model 2 in BPCI. However, that model was the largest in terms of enrollment and the basis for BPCI Advanced.

Seventh, the study findings might not apply to mandatory programs such as the Comprehensive Care for Joint Replacement model.[10]

Study Results

Our sample included 244 BPCI hospitals in 123 BPCI markets during the BPCI period (exhibit 1), with 10,757 BPCI patients per quarter in the early BPCI period and 10,949 in the late BPCI period (data not shown). Our matched comparison group consisted of 244 non-BPCI hospitals operating in 98 non-BPCI markets during the BPCI period. There were 2,819 non-BPCI patients per quarter at non-BPCI hospitals in the early BPCI period and 2,375 in the late BPCI period. Wald tests did not indicate divergent secular trends between BPCI and non-BPCI patients during the pre-BPCI period (for details about our tests of parallel trends, see the appendix methods 3 section in the online appendix).[22]

PATIENT, HOSPITAL, AND MARKET CHARACTERISTICS A number of characteristics differed by study period among both BPCI and non-BPCI patients (exhibit 2; see also appendix exhibit 5).[22] For example, in both cases, the mean Elixhauser comorbidity index of patients was significantly lower in the late BPCI period than the early BPCI or pre-BPCI periods. There were several other patient characteristics with small but significant differences by study period, although the changes did not exhibit meaningful differential trends across BPCI and non-BPCI patients (appendix exhibit 6).[22]

BPCI and non-BPCI hospitals varied with respect to a number of characteristics, with differences similar to those described in prior reports.[8,27] For example, compared to non-BPCI hospitals, BPCI hospitals were larger and more likely to be urban, not for profit, and teaching hospitals (exhibit 1). In both BPCI and non-BPCI markets, Medicare Advantage penetration and the proportion of markets with physician group practices were greater in the late BPCI period compared to the early BPCI and pre-BPCI periods (appendix exhibit 7).[22] Differences between non-BPCI and BPCI hospitals were small after propensity score matching (appendix exhibit 1).[22]

EPISODE SPENDING For BPCI patients, mean episode spending was $23,552 in the pre-BPCI period and $22,129 in the BPCI period (a reduc-

EXHIBIT 1

Baseline characteristics of hospitals that participated in the Bundled Payments for Care Improvement (BPCI) initiative in 2011 and those that did not, after propensity score matching

Characteristic	Non-BPCI hospitals (n = 244)	BPCI hospitals (n = 244)	Standardized difference
HOSPITAL ADMISSIONS			
Mean annual admissions for top 10 BPCI episodes (%)[a]	22.7	22.6	–0.03
Median no. of annual admissions for LEJR	170	179	0.01
Mean share of discharges to the highest-volume SNF (%)	28.9	29.5	0.04
Median share of discharges to the highest-volume IRF (%)	90.0	80.0	–0.10
Median 90-day readmission rate	11.0	11.0	0.12
Median 90-day LEJR episode spending ($)	23,936	24,355	0.09
HOSPITAL ORGANIZATIONS			
Median no. of beds	235	261	0.06
Ownership status (%)			
For profit	17.2	16.4	–0.02
Not for profit	77.1	78.7	0.04
Government owned	5.7	4.9	–0.04
Member of a system (%)	79.1	78.7	–0.01
Teaching status (%)[b]			
Major teaching	13.5	12.7	–0.02
Minor teaching	31.6	34.4	0.06
Nonteaching	54.9	52.9	–0.04
Median ratio of interns and residents to beds	0.0	0.0	0.08
Median DSH payments ($)[c]	2,162,258	2,472,058	0.02
Urban location (%)	99.2	99.2	0.00
Mean Medicare days (% of total patient days)	52.9	52.0	–0.07
Median market share (%)	9.2	9.2	0.03
MARKETS			
Median no. of beneficiaries	1,068,113	1,321,591	0.14
Median income ($)	52,303	53,089	0.05
Median SNF beds	6,124	6,308	–0.01
Median IRF beds	122	134	0.05
Medicare Advantage penetration[d]	24.7	25.4	0.05
Hospital market concentration (HHI)	2,216	2,181	–0.02

SOURCE Authors' analysis of 2011 data on hospital characteristics from the American Hospital Association (AHA) Annual Survey and 2011 Medicare claims data. **NOTES** The exhibit shows the characteristics of hospitals used in propensity score matching. Standardized difference refers to the difference in the mean values of a characteristic between the non-BPCI hospital and BPCI hospital groups, divided by an estimate of the standard deviation of the values of that characteristic. Medians are provided where the data are skewed. A fuller version of this exhibit, showing standardized differences before and after propensity score matching, is in appendix exhibit 1 (see note 22 in text). LEJR is lower extremity joint replacement. SNF is skilled nursing facility. IRF is inpatient rehabilitation facility. HHI is Herfindahl-Hirschman Index. [a]Major joint replacement of the lower extremity; double joint replacement of the lower extremity; revision of the hip or knee; hip and femur procedures except major joint; lower extremity and humerus procedure except hip, foot, and femur; coronary artery bypass graft; acute myocardial infarction; congestive heart failure; simple pneumonia and respiratory infections; and chronic obstructive pulmonary disease or bronchitis or asthma. [b]The AHA Annual Survey defines major teaching hospitals as those that are members of the Council of Teaching Hospitals (COTH), minor teaching hospitals as non-COTH members that reported a medical school affiliation to the American Medical Association, and nonteaching hospitals as all other hospitals. [c]Disproportionate share hospital (DSH) payment percentages are derived from the 2017 CMS Improving Medicare Post-Acute Care Transformation (IMPACT) file. [d]The number of Medicare Advantage enrollees in a market divided by the number of Medicare beneficiaries in that market.

tion of $1,423; $p < 0.001$) (appendix exhibit 8).[22] Mean episode spending among non-BPCI patients was $22,834 in the pre-BPCI period and $22,073 in the BPCI period (a reduction of $761; $p < 0.001$). In unadjusted analyses of raw means, the differential change in mean episode spending for BPCI versus non-BPCI patients was –2.7 percent (a differential reduction of approximately $662; $p < 0.001$).

In an adjusted difference-in-differences analysis that used the instrumental variable, LEJR among patients admitted to BPCI hospitals was associated with a 1.6 percent differential decrease (of approximately $377) in episode spending before versus after hospitals initiated BPCI participation, as compared to changes for patients admitted to non-BPCI hospitals (exhibit 3). In the early BPCI period, BPCI participation was associated with a 1.8 percent differential decrease in mean episode spending. In contrast, BPCI participation was not associated with significant changes in mean episode spending for

EXHIBIT 2

Sample and patient characteristics, by hospital participation in the Bundled Payments for Care Improvement (BPCI) initiative and program period

	Non-BPCI patients			BPCI patients		
Characteristic	**Pre-BPCI**	**Early BPCI**	**Late BPCI**	**Pre-BPCI**	**Early BPCI**	**Late BPCI**
SAMPLE						
Markets	98	98	98	123	123	123
Hospitals	244	244	241	244	244	244
Beneficiaries	20,497	19,734	11,876	75,614	75,297	54,744
PATIENTS						
Mean age (years)	73.4	73.4	73.0	73.3**	73.0**	73.0**
Black race (%)[a]	8.7	8.3	7.9	7.6	7.2	7.1
Female (%)	64.4	63.4	62.2	66.1****	64.2****	64.2****
Dually eligible (%)[b]	13.3	12.5	11.8	15.5**	14.2**	13.5**
Mean Elixhauser comorbidity index[c,d]	5.5****	5.3****	4.4****	5.4****	4.8****	4.3****
Prior acute care hospital use (%)[d]	18.9***	17.5***	16.3***	17.9****	15.7****	15.3****
Prior IRF use (%)[d]	1.3	1.6	1.4	1.4	1.2	1.2
Prior SNF use (%)[d]	5.2	5.3	4.9	5.1**	4.7**	4.6**

SOURCE Authors' analysis of Medicare claims data for 2011–16. **NOTES** Patient characteristics are those of people who received lower extremity joint replacement (LEJR), drawn from a 20 percent Medicare claims sample for non-BPCI hospitals and from a 100 percent sample for BPCI hospitals. The pre-BPCI period was January 2011–September 2013. The early BPCI period was October 2013–June 2015, and the late BPCI period was July 2015–December 2016 (data through September 2016 were used to define LEJR episodes to allow for a ninety-day postdischarge period between October and December 2016). Wilcoxon rank-sum tests or *t*-tests were used to test the differences in continuous variables, and chi-square tests were used for categorical variables. Appendix exhibit 6 contains a fuller version of this exhibit, with differential changes for BPCI and non-BPCI patients (see note 22 in text). IRF is inpatient rehabilitation facility. SNF is skilled nursing facility. [a]Race was broken out as black versus others because of existing disparities in access to LEJR among black patients specifically. [b]Eligible for Medicare and Medicaid (as an indicator of low socioeconomic status). [c]The Elixhauser comorbidity score, an index of severity, ranges from −32 to 92, with increasing scores highly correlated with increased probability of in-hospital death. [d]Calculated using data from the year before LEJR hospitalization. **$p < 0.05$ ***$p < 0.01$ ****$p < 0.001$

the late BPCI period.

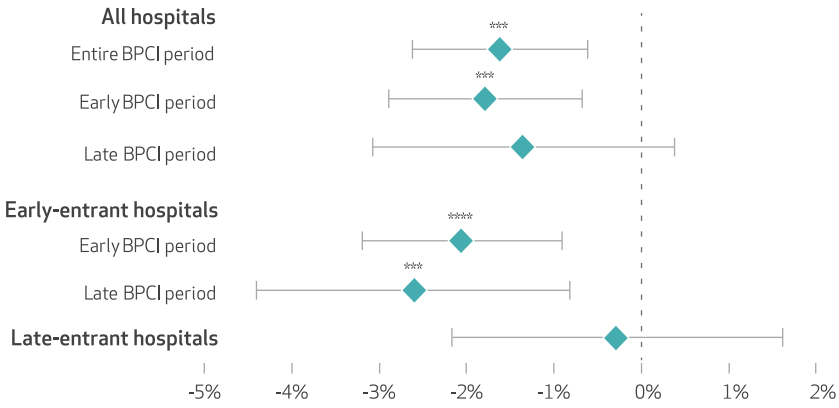
In a secondary analysis, BPCI participation was associated with decreased episode spending for early-entrant hospitals in both the early and late BPCI periods (−2.1 percent and −2.6 percent differential changes, respectively). In contrast, participation was not associated with changes in episode spending among late-entrant hospitals (−0.3 percent differential change).

EPISODE SPENDING COMPONENTS The raw values of several spending components varied by study period (appendix exhibit 9).[22] In adjusted analyses BPCI participation was associated with a differential decrease in episode spending attributable to SNF care (−0.4 percent; $p = 0.02$) and IRF care (−0.7 percent; $p < 0.001$) for BPCI versus non-BPCI patients (appendix exhibit 10).[22] In contrast, spending attributable to index hospitalizations (0.7 percent; $p = 0.001$), home health agency services (0.2 percent; $p = 0.02$), and professional services (0.2 percent; $p = 0.03$) increased differentially for BPCI patients. These associations were generally larger for early versus late BPCI periods and early- versus late-entrant hospitals.

QUALITY OUTCOMES We observed secular trends for a number of quality measures among both BPCI and non-BPCI patients (appendix ex-

EXHIBIT 3

Percent changes in episode spending associated with hospital participation in the Bundled Payments for Care Improvement (BPCI) initiative, by program period and timing of participation



SOURCE Authors' analysis of Medicare claims data for 2011–16. **NOTES** This exhibit shows the results from a series of difference-in-differences models that used an instrumental variable to evaluate the association between participation in BPCI and differential changes in mean episode spending. Separate models were used to evaluate associations between BPCI participation and changes in mean episode spending for the overall cohort and study period (that is, all hospitals and the entire BPCI period, October 2013–December 2016), all hospitals by program period (early versus late BPCI period, defined in the notes to exhibit 2), early-entrant hospitals (those that began participating in the early BPCI period) by program period, and late-entrant hospitals (those that began participating in the late BPCI period) in the late BPCI period. All spending estimates were standardized and transformed into 2016 dollars. Negative estimates indicate savings. The error bars indicate 95% confidence intervals. ***$p < 0.01$ ****$p < 0.001$

hibit 8).[22] In adjusted analyses there were no significant differential changes between BPCI and non-BPCI patients in ninety-day risk-standardized mortality (−0.15 percent; $p = 0.35$), unplanned readmission (0.15 percent; $p = 0.67$), emergency department visit (−0.19 percent; $p = 0.63$), or LEJR-specific complication (0.12 percent; $p = 0.59$) rates (exhibit 4). BPCI participation was not associated with changes in quality outcomes when evaluated specifically by program period or duration of hospital participation.

SENSITIVITY ANALYSES When we extended the approach from prior CMS evaluations, we found that BPCI participation was associated with a 2.2 percent differential decrease in mean episode spending (appendix exhibit 12),[22] a value that was 27 percent lower than estimates from our main analyses (1.6 percent differential decrease in mean episode spending). Analyses that excluded data for January–September 2013, used less stringent propensity matching, and removed hospitals that participated in the Comprehensive Care for Joint Replacement model yielded generally similar results (appendix exhibits 13–15).[22]

Discussion

In this long-term study, participation in the Bundled Payments for Care Improvement initiative was associated with a 1.6 percent decrease in average lower extremity joint replacement episode spending, driven by the performance of early participants, with no changes in quality. This finding has five important implications in light of CMS's decision to continue implementing voluntary bundles through BPCI Advanced.

First, early reductions in episode spending accounted for the savings observed over three years under voluntary LEJR bundled payment. Along with early work that evaluated the impact of mandatory CMS bundled payment programs for LEJR,[11,28] our findings fill an important knowledge gap about bundled payment policy. In particular, this study provides the first rigorous evidence about savings from voluntary LEJR bundled payment that, albeit lower than the 3.9 percent savings demonstrated in early program evaluations and 21 percent savings achieved by high performers,[13] were nonetheless sustained through three years. Longer-term savings support Medicare's decision to expand voluntary bundles through BPCI Advanced.

Second, early 1.8 percent savings came from early entrants that sustained savings over time. These results suggest that observed decreases in SNF and IRF spending continued in the late BPCI period and align with evidence from other value-based payment arrangements such as accountable care organizations[29] and the Alternative Quality Contract[30] that organizations can achieve savings over longer time periods.

Third, despite sustained savings among early entrants, the lack of observed overall LEJR episode savings later in BPCI suggests that savings might not be generalizable across all participants. This may be because of differences between hospitals in the initial and later participation waves. Later participants may have been less able to influence episode spending by systematically changing discharge practices to institutional postacute care or driving practice improvement in activities that require greater time and resource investment (for example, the establishment of clinical protocols for handling potential complications among preferred SNFs). While this study did not evaluate strategies employed by participants, the results collectively suggest that opportunities to achieve savings vary by

EXHIBIT 4

Changes in quality outcomes associated with hospital participation in the Bundled Payments for Care Improvement (BPCI) initiative, by program period and timing of participation

		Program period		Timing of participation		
				Early-entrant hospitals		Late-entrant hospitals
Outcome	Overall	Early BPCI	Late BPCI	Early BPCI	Late BPCI	Late BPCI
Mortality rate	−0.15%	−0.03%	−0.34%	0.01%	−0.27%	−0.37%
Unplanned readmission rate	0.15	0.21	0.03	0.24	−0.10	0.18
ED visit rate	−0.19	−0.22	−0.08	−0.25	0.15	0.54
LEJR-specific complication rate[a]	0.12	0.21	−0.03	0.24	0.003	−0.09

SOURCE Authors' analysis of Medicare claims data for 2011–16. **NOTES** The exhibit shows results from difference-in-differences models that evaluated the association between participation in BPCI and differential changes in quality outcomes, with changes displayed for the overall cohort and intervention period (entire BPCI period) as well as for program period (early and late BPCI, defined in the notes to exhibit 2) and timing of participation (early- and late-entrant hospitals, defined in the notes to exhibit 3). Negative estimates indicate reductions in rates (that is, quality improvements). Emergency department (ED) visits are those without a hospitalization. LEJR is lower extremity joint replacement. Appendix exhibit 11 contains a fuller version of this exhibit (see note 22 in text). [a]Defined by Hospital Compare.

timing of participation without diminishing over time for those able to achieve them—critical insights, given the two participation waves in BPCI Advanced.

Fourth, the 1.6 percent episode savings observed in our primary instrumental variable analyses and the 2.2 percent episode savings observed in our non–instrumental variable sensitivity analyses suggest that while patient selection exists in LEJR bundled payment, it does not fully account for associated savings. The fact that both estimates were smaller than those reported in prior peer-reviewed (−3.8 percent) and CMS-contracted (−3.9 percent) analyses of BPCI over one to three years highlights the ways in which our analysis differs methodologically from those evaluations.[1,4] For instance, CMS contractor evaluations did not account for time-varying entry into BPCI or unobserved differences in characteristics between BPCI and non-BPCI hospitals (for example, they did not include hospital fixed effects). Additional reasons why those results vary in magnitude from ours may include differences in study samples (for example, we excluded episodes covered by the Comprehensive Care for Joint Replacement model) and variables (for example, we incorporated hospital characteristics from the American Hospital Association).

A comparative strength of our analytic approach and use of the instrumental variable is that it mitigated the effects of changes in unobserved patient characteristics. In particular, our instrumental variable analysis yielded estimates that were 27 percent lower than those of our sensitivity analysis without the variable (1.6 percent versus 2.2 percent). While more work is needed to ensure that policy evaluations account for unobserved selection, our results provide early estimates that reflect the fact that even in the absence of observable selection (that is, selection based on characteristics captured in our data), unobserved selection can still occur. For instance, while physicians and hospitals might not select patients for LEJR based on sex or clinical comorbidities, as observed in our analysis, they might still preferentially avoid patients perceived to have poor social support or low treatment adherence. Notably, estimates from our analysis were similar in magnitude to those from evaluations of Medicare's mandatory joint replacement program, which are not as susceptible to patient selection.[28,31]

Taken together, our results corroborate the presence of unobserved patient selection in voluntary bundled payment and demonstrate associated savings in spite of it. These are important insights as Medicare continues to emphasize voluntary bundled payment through BPCI Advanced, whose results are far more likely to mirror those of BPCI than those of mandatory programs—because of organizations' ability to choose whether and when to participate.

Fifth, several measures of quality did not appear to change under LEJR bundled payment, even among hospitals that participated in BPCI for close to two years. On the one hand, these findings indicate that despite generating financial savings, early participants were unable to redesign practice or coordinate care in ways that reduced mortality, complication, readmission, or emergency department visit rates. On the other hand, this analysis suggests that financial savings did not appear to come at the expense of quality. Future work should evaluate other measures of quality, such as patient-reported outcomes, that might not appear in administrative data.

Conclusion

Participation in bundled payment for lower extremity joint replacement under the Bundled Payments for Care Improvement initiative was associated with episode savings over three years, though savings were driven by early participants, and there were no associated changes in quality. These results have important policy implications as CMS continues expanding voluntary bundled payment via BPCI Advanced. ■

Earlier versions of the analyses reported in this article were presented at the AcademyHealth Annual Research Meeting in Washington, D.C., June 5, 2019, and at the Society of General Internal Medicine Annual Meeting in Washington, D.C., May 9, 2019. This work was funded by the Commonwealth Fund and the Leonard Davis Institute of Health Economics at the University of Pennsylvania. Amol Navathe received personal fees from Navvis Healthcare Inc., the National University Health System of Singapore, and Agathos Inc.; personal fees and equity from NavaHealth; personal fees for service as a commissioner of the Medicare Payment Advisory Commission; speaking fees from the Cleveland Clinic; and an honorarium from Elsevier Press, none of which were related to this article. He also serves without compensation as a board member of Integrated Services Inc. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US government.

NOTES

1 Centers for Medicare and Medicaid Services. Bundled Payments for Care Improvement (BPCI) initiative: general information [Internet]. Baltimore (MD): CMS; [last updated 2019 Apr 17; cited 2019 Sep 23]. Available from: https://innovation.cms.gov/initiatives/bundled-payments/

2 Dummit LA, Kahvecioglu D, Marrufo G, Rajkumar R, Marshall J, Tan E, 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):1267–78.

3 Lewin Group. CMS Bundled Payments for Care Improvement (BPCI) initiative models 2–4: year 2 evaluation and monitoring annual report [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; 2016 Aug [cited 2019 Sep 23]. Available from: https://innovation.cms.gov/Files/reports/bpci-models2-4-yr2evalrpt.pdf

4 Lewin Group. CMS Bundled Payments for Care Improvement (BPCI) initiative models 2–4: year 5 evaluation and monitoring annual report [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; 2018 Oct [cited 2019 Sep 23]. Available from: https://downloads.cms.gov/files/cmmi/bpci-models2-4-yr5evalrpt.pdf

5 CMS.gov. BPCI Advanced [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; [last updated 2019 Sep 16; cited 2019 Sep 23]. Available from: https://innovation.cms.gov/initiatives/bpci-advanced

6 Bleser WK, Muhlestein D, Saunders RS, McClellan MB. Half a decade in, Medicare accountable care organizations are generating net savings: part 1. Health Affairs Blog [blog on the Internet]. 2018 Sep 20 [cited 2019 Sep 23]. Available from: https://www.healthaffairs.org/do/10.1377/hblog20180918.957502/full/

7 Muhlestein D, Saunders RS, Richards R, McClellan MB. Recent progress in the value journey: growth of ACOs and value-based payment models in 2018. Health Affairs Blog [blog on the Internet]. 2018 Aug 14 [cited 2019 Sep 23]. Available from: https://www.healthaffairs.org/do/10.1377/hblog20180810.481968/full/

8 Navathe AS, Liao JM, Dykstra SE, Wang E, Lyon ZM, Shah Y, et al. Association of hospital participation in a Medicare bundled payment program with volume and case mix of lower extremity joint replacement episodes. JAMA. 2018;320(9):901–10.

9 Dartmouth Institute. Dartmouth atlas of health care, data by region [Internet]. Lebanon (NH): Trustees of Dartmouth College; [cited 2019 Dec 9]. Available from: https://atlasdata.dartmouth.edu/static/supp_research_data/#crosswalks

10 Navathe AS, Liao JM, Polsky D, Shah Y, Huang Q, Zhu J, et al. Comparison of hospitals participating in Medicare's voluntary and mandatory orthopedic bundle programs. Health Aff (Millwood). 2018;37(6):854–63.

11 Navathe AS, Liao JM, Shah Y, Lyon Z, Chatterjee P, Polsky D, et al. Characteristics of hospitals earning savings in the first year of mandatory bundled payment for hip and knee surgery. JAMA. 2018;319(9):930–2.

12 Liao JM, Emanuel EJ, Polsky DE, Huang Q, Shah Y, Zhu J, et al. National representativeness of hospitals and markets in Medicare's mandatory bundled payment program. Health Aff (Millwood). 2019;38(1):44–53.

13 Navathe AS, Troxel AB, Liao JM, Nan N, Zhu J, Zhong W, et al. Cost of joint replacement using bundled payment models. JAMA Intern Med. 2017;177(2):214–22.

14 Van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Med Care. 2009;47(6):626–33.

15 Liao JM, Holdofski A, Whittington GL, Zucker M, Viroslav S, Fox DL, et al. Baptist Health System: succeeding in bundled payments through behavioral principles. Healthc (Amst). 2017;5(3):136–40.

16 CMS.gov. Comprehensive Care for Joint Replacement model [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; [last updated 2019 Sep 18; cited 2019 Sep 23]. Available from: https://innovation.cms.gov/initiatives/CJR

17 Tsai TC, Joynt KE, Wild RC, Orav EJ, Jha AK. Medicare's bundled payment initiative: most hospitals are focused on a few high-volume conditions. Health Aff (Millwood). 2015;34(3):371–80.

18 Joynt KE, Gawande AA, Orav EJ, Jha AK. Contribution of preventable acute care spending to total spending for high-cost Medicare patients. JAMA. 2013;309(24):2572–8.

19 Medicare.gov. Hospital Compare datasets [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; [last updated 2019 Oct 30; cited 2019 Nov 19]. Available from: https://data.medicare.gov/data/hospital-compare

20 Meyer BD. Natural and quasi-experiments in economics. J Bus Econ Stat. 1995;13(2):151–61.

21 Card D, Krueger A. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. Am Econ Rev. 1994;84(4):772–93.

22 To access the Appendix, click on the Details tab of the article online.

23 Angrist JD, Pischke J-S. Mostly harmless econometrics: an empiricist's companion. Princeton (NJ): Princeton University Press; 2009.

24 Woolridge JM. Econometric analysis of cross section and panel data. Cambridge (MA): MIT Press; 2002.

25 Alexander D. How do doctors respond to incentives? Unintended consequences of paying doctors to reduce costs [dissertation]. Princeton (NJ): Princeton University; 2006.

26 MacKinnon JG, White H. Some heteroskedasticity-consistnet covariance matrix estimators with improved finite sample properties. J Econom. 1985;29(3):305–25.

27 Joynt Maddox KE, Orav EJ, Zheng J, Epstein AM. Participation and dropout in the Bundled Payments for Care Improvement initiative. JAMA. 2018;319(2):191–3.

28 Finkelstein A, Ji Y, Mahoney N, Skinner J. 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.

29 McWilliams JM, Hatfield LA, Landon BE, Hamed P, Chernew ME. Medicare spending after 3 years of the Medicare Shared Savings Program. N Engl J Med. 2018;379(12):1139–49.

30 Song Z, Safran DG, Landon BE, Landrum MB, He Y, Mechanic RE, et al. The "Alternative Quality Contract," based on a global budget, lowered medical spending and improved quality. Health Aff (Millwood). 2012;31(8):1885–94.

31 Barnett ML, Wilcock A, McWilliams JM, Epstein AM, Joynt Maddox KE, Orav EJ, et al. Two-year evaluation of mandatory bundled payments for joint replacement. N Engl J Med. 2019;380(3):252–62.

Navathe AS, Emanuel EJ, Venkataramani AS, Huang Q, Gupta A, Dinh CT, et al.
Spending and quality after three years of Medicare's voluntary bundled payment for joint replacement surgery.
Health Aff (Millwood). 2020;39(1).

Appendix

Contents

Appendix Exhibit 11. Changes in quality outcomes associated with BPCI participation by program period and timing of hospital participation, 2012-2016 (Full Table)

Appendix Exhibit 12. Sensitivity analysis extending methods from prior CMS evaluations

Appendix Exhibit 13. Sensitivity analysis excluding January to September 2013 from the baseline period

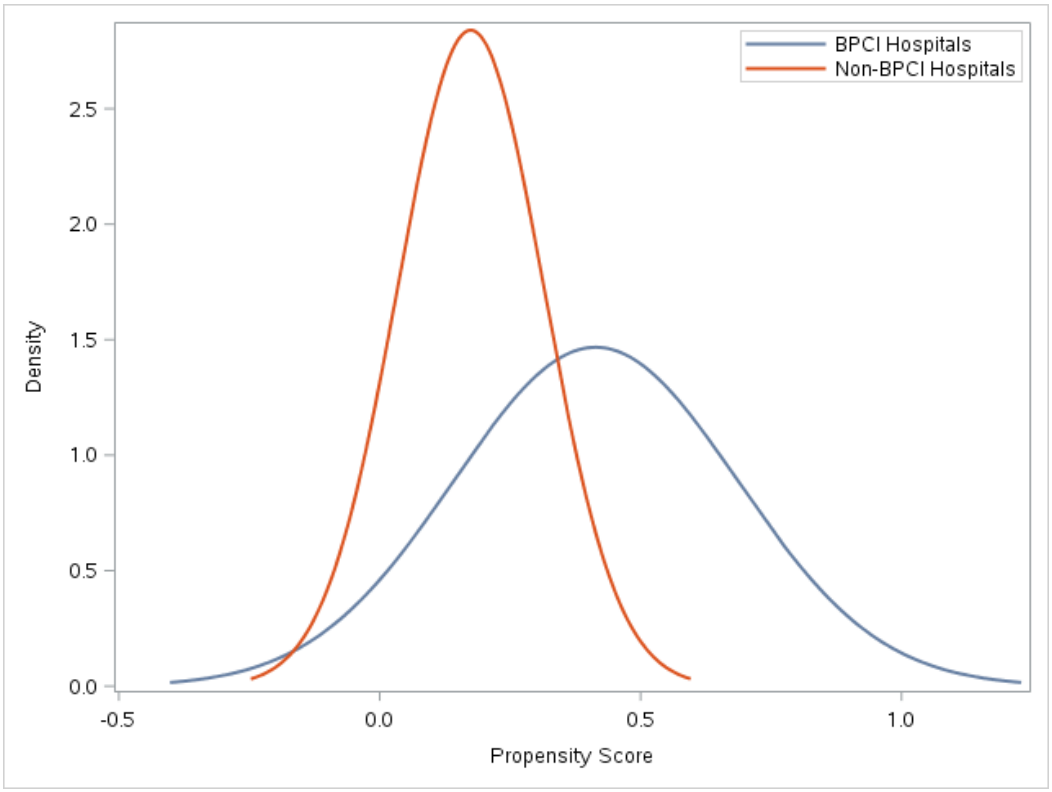
Appendix Exhibit 14. Sensitivity analysis using less stringent propensity matching

Appendix Exhibit 15. Sensitivity analysis excluding CJR hospitals from propensity score matching and analyses

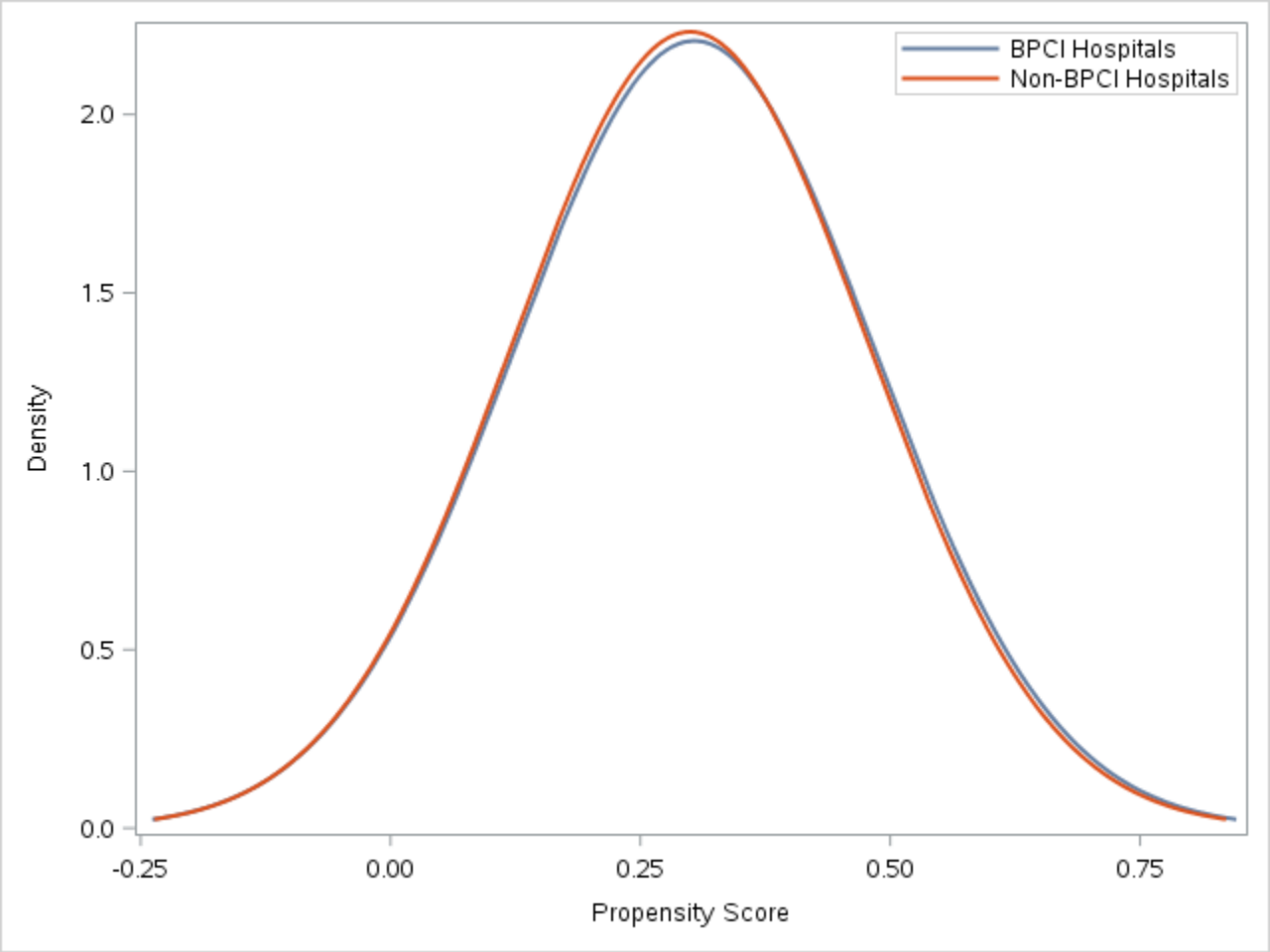
Appendix Methods 1. Propensity score matching of BPCI hospitals to Non-BPCI hospitals

To decrease differences between BPCI and Non-BPCI hospitals, we used propensity score matching on a variables drawn from Medicare claims and AHA Annual Survey data (see Appendix Exhibit 3 below for all the variables). Hospitals not participating in BPCI but located in BPCI markets could have been affected by program (e.g., from changes in market share or referral patterns after BPCI began). Thus, they were not used in the Non-BPCI Hospital group throughout our analyses to avoid potential confounding. Our propensity model used logistic regression with a binary dependent variable of participation in BPCI during any quarter in the BPCI period for the LEJR episode and independent variables listed in Appendix Exhibit 1. Using a minimum of a 2 digit propensity score match, we were able to match 244 BPCI hospitals with 244 Non-BPCI hospitals in Non-BPCI markets in a 1:1 match.[1] Comparison of hospital characteristics between Non-BPCI and BPCI hospitals before and after matching is presented in Appendix Exhibit 1 below. As a sensitivity analysis, we used a less stringent optimal propensity match method by varying the match ratio from 1:1 to 1:3 and using a caliper=0.05 to match 225 BPCI with 628 Non-BPCI hospitals.[2]

Density plot: propensity score distribution of all BPCI and all Non-BPCI Hospitals in Non-BPCI Markets, before matching

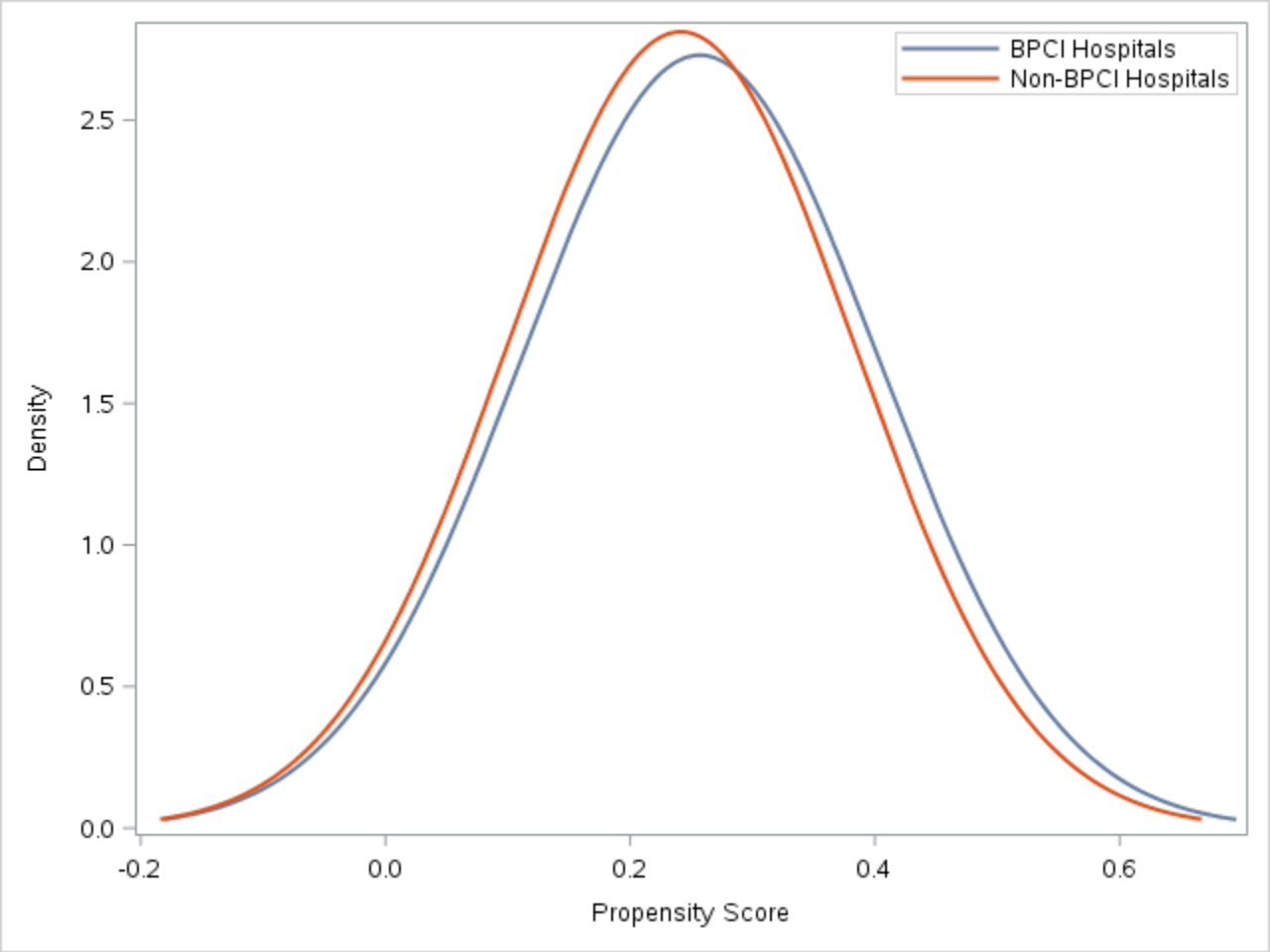


Density plot: propensity score distribution of BPCI (n=244) and Non-BPCI (n=244) Hospitals, after matching



These density plot figures illustrate the area of common support and the distribution of propensity scores before and after matching using a minimum of a 2 digit propensity score match in a 1:1 match.

Density plot: less stringent propensity score distribution of BPCI (n=225) and Non-BPCI (n=628) Hospitals, after matching



These density plot figures illustrate the area of common support and the distribution of propensity scores before and after matching using a less stringent optimal propensity match method by varying the match ratio from 1:1 to 1:3 and using a caliper=0.05. This resulted in 225 BPCI hospitals matched to 628 Non-BPCI Hospitals.

Appendix Exhibit 1. Comparison of baseline characteristics of BPCI hospitals and non-BPCI hospitals before and after propensity matching, 2011

Patient Characteristics						
	Before Propensity Matching			**After Propensity Matching**		
	Non-BPCI Hospitals	**BPCI Hospitals**	**Standardized Difference**	**Non-BPCI Hospitals**	**BPCI Hospitals**	**Standardized Difference**
Annual admissions for top 10 BPCI episodes, mean % (SD)[a]	24.8 (6.7)	22.1 (5.0)	-0.46	22.7 (5.2)	22.6 (5.0)	-0.03
Annual admissions for LEJR, median (IQR)[b]	118 (55 to 229)	175 (94 to 296)	0.39	170 (90 to 314)	179 (96 to 293)	0.01
Proportion of discharges to highest-volume SNF, mean % (SD)	35.7 (19.5)	28.2 (17.0)	-0.41	28.9 (14.8)	29.5 (17.6)	0.04
Proportion of discharges to highest-volume IRF, median % (IQR)[b]	50.0 (0.0 to 100)	80.0 (0.0 to 99.6)	0.22	90.0 (0.0 to 100.0)	80.0 (0.0 to 100.0)	-0.10
90-day readmission rate, median (IQR)[b]	10.5 (5.7 to 16.7)	11.3 (9.1 to 15.3)	0.25	7.1 (16.3)	11.0 (9.0 to 15.0)	0.12
90-day LEJR episode spending, median $ (IQR)[b]	22,996 (20,134 to 26,813)	24,606 (22,343 to 27,523)	0.39	23,936 (21,092 to 27,930)	24,355 (21,977 to 27,309)	0.09
Hospital Characteristics						
	Before Propensity Matching			**After Propensity Matching**		
	Non-BPCI Hospitals	**BPCI Hospitals**	**Standardized Difference**	**Non-BPCI Hospitals**	**BPCI Hospitals**	**Standardized Difference**
Number of beds, median (IQR)[b]	176 (94 to 313)	268 (164 to 402)	0.53	235 (148 to 401)	261 (154 to 391)	0.06
Ownership status, %						
For-profit	19.0	18.2	0.36	17.2	16.4	0.05
Not-for-profit	66.7	78.0		77.1	78.7	
Government	14.4	3.8		5.7	4.9	
Member of a system, %	63.8	80.2	0.37	79.1	78.7	-0.01
Teaching status, %[c]						
Major teaching	7.9	14.7	0.3	13.5	12.7	0.05
Minor teaching	30.1	36.1		31.6	34.4	
Non-teaching	62.0	49.2		54.9	52.9	

Intern and resident to bed ratio, median (IQR)[b]	0.0 (0.0 to 0.03)	0.0 (0.0 to 0.08)	0.33	0.0 (0.0 to 0.06)	0.0 (0.0 to 0.06)	0.08
Disproportionate share hospital payments, median $ (IQR)[d]	1,464,209 (422,560 to 4,464,053)	2,723,772 (521,608 to 7,198,542)	0.22	2,162,258 (390,499 to 6,420,029)	2,472,058 (492,948 to 6,560,881)	0.02
Urban status, %	94.4	99.4	0.29	99.2	99.2	0
Medicare days as % of total patient days, mean (SD)	52.0 (12.8)	51.7 (10.7)	-0.03	52.9 (11.8)	52.0 (10.7)	0.07
Market share, median % (IQR)[b]	7.7 (3.2 to 19.6)	7.9 (3.4 to 18.4)	-0.004	9.2 (3.7 to 23.8)	9.2 (4.4 to 19.7)	0.03
Market characteristics						
	Before Propensity Matching			**After Propensity Matching**		
Total number of beneficiaries, median (IQR)[b]	821,288 (449,198 to 1,523,905)	1,617,708 (796,342 to 2,765,863)	0.61	1,068,113 (591,261 to 2,695,204)	1,321,591 (678,746 to 2,075,609)	0.14
Median income, $ (SD)	48,143 (11,293)	56,104 (16,608)	0.56	52,303 (14,697)	53,089 (14,543)	0.05
SNF beds, median (IQR)[b]	5,026 (2,752 to 9,607)	7,652 (4,043 to 15,361)	0.43	6,124 (3,420 to 13,670)	6,308 (3,578 to 9,646)	-0.01
IRF beds, median (IQR)[b]	97 (54 to 163)	180 (82 to 388)	0.57	122 (64 to 270)	134 (60 to 253)	0.05
MA penetration, % (SD)	24.2 (13.7)	24.8 (11.5)	0.05	24.7 (15.1)	25.4 (11.5)	0.05
Hospital concentration, HHI (SD)	2,471 (1,762)	1,990 (1,564)	-0.29	2,216 (1,735)	2,181 (1,653)	-0.02

[a]Major joint replacement of the lower extremity, Double joint replacement of the lower extremity, Revision of the hip or knee, Hip and femur procedures except major joint, Lower extremity and humerus procedure except hip, foot, and femur, Coronary artery bypass graft, Acute myocardial infarction, Congestive heart failure, Simple pneumonia and respiratory infections, Chronic obstructive pulmonary disease, bronchitis/asthma. [b]Median (IQR) provided where data are skewed. [c]From the AHA Annual Survey, major teaching hospitals are those that are members of the Council of Teaching Hospitals (COTH), minor teaching hospitals are non-COTH members that had a medical school affiliation reported to the American Medical Association, and nonteaching hospitals are all other institutions. [d]Disproportionate share hospital payment percentage derived from the FY2017 CMS IMPACT file.

Appendix Exhibit 2. Clinical conditions used in statistical models

Acquired immune deficiency syndrome
Alcohol abuse
Chronic blood loss anemia
Chronic pulmonary disease
Coagulopathy
Congestive heart failure
Deficiency anemias
Depression
Diabetes with chronic complications
Diabetes without chronic complications
Drug abuse
Fluid and electrolyte disorders
Hypertension with complication
Hypothyroidism
Liver disease
Lymphoma
Metastatic cancer
Obesity
Other neurological disorders
Paralysis
Peptic ulcer disease excluding bleeding
Peripheral vascular disease
Psychoses
Pulmonary circulation disease
Renal failure
Rheumatoid arthritis/collagen vascular disease
Solid tumor without metastasis
Valvular disease
Weight loss

Appendix Methods 2. Instrumental variable approach

We introduce a new instrumental variable (IV) to mitigate confounding from selection of beneficiaries for hospitalization based on unobservable characteristics. This IV is an adaptation of instrumental variables from outside of healthcare.[3-4] Our approach uses historical hospital referral patterns before the beginning of the BPCI program (i.e., to which hospitals patients were hospitalized for LEJR prior to October 2013) to predict which patients would be hospitalized for LEJR at hospitals later participating in the BPCI program (i.e., which patients would be hospitalized at BPCI Hospitals in the BPCI period based on historical hospitalization patterns prior to the BPCI period). Specifically, we used a set of patient characteristics -- age, race, gender, zip code of residence, clinical conditions (listed in Appendix Exhibit Table 2), MS-DRG, as well as prior hospital, SNF, and IRF use -- from 2011 to predict the probability that a given patient would be hospitalized for LEJR at a hospital that later participates in BPCI (BPCI Hospital).

Importantly, we could not use other standard IVs used in health care services research such as the distance to hospital. This is because many hospital systems in bundled payment exerted substantial effort to attract preferred (i.e., lower risk or healthier) types of patients, many of whom were at greater distance from the hospital. These efforts including buying hospitals in outlying suburban areas, buying physician practices, or contracting with surgeons in areas with a greater density of preferred patients.

Because historical hospital referral patterns for elective conditions such as LEJR are not correlated with changes in patient selection after hospital participation in BPCI (conditional on variables we can observe in our data such as patient demographics), they serve as a

reasonable IV for our analysis. In tests of the instrumental variable with respect to observed treatment at a BPCI hospital, we found a strong association (F-measure of 11,835) between predicted and observed BPCI exposure, controlling for time-varying hospital, market, and patient characteristics and including hospital, market, and time fixed-effects. This confirmed that we had a strong instrument that was uncorrelated with the confounder of unobservable patient selection, but highly predictive of our treatment, i.e., admission to a BPCI Hospital (in the BPCI program period). The instrument was constructed using 2011 data and the formula $BPCI_{ever,h} = Zip_{pt} + Cov_{pt} + \varepsilon$, where $BPCI_{ever,h}$ is an indicator describing whether a Medicare beneficiary was admitted for LEJR at a hospital that later joins BPCI (at any point), Zip_{pt} is the beneficiary's ZIP code of residence, Cov_{pt} is a vector of characteristics of the beneficiary (including demographics and clinical comorbidities) and ε is the error term.

In the first-stage regression, we used pre-BPCI period hospital referral patterns in 2011 to generate the predicted probability of hospitalization for LEJR at an eventual BPCI Hospital ($\bar{\pi}_{pt}$), and used that probability as an instrument for actual hospitalization for LEJR at a BPCI Hospital. The first stage is $BPCI\_observed_{pt,h,t} = \bar{\pi}_{pt} + HRR_{FE} + Time_{FE} + Cov_{pt} + Cov_{HRR} + \varepsilon$, where the observed BPCI status for a beneficiary receiving LEJR (whether the hospital the beneficiary was admitted to participated in BPCI in that market-quarter) was regressed on the IV ($\bar{\pi}_{pt}$, the predicted probability for that beneficiary of going to a BPCI hospital based on historical patterns), market (based on patient ZIP code of residence) and time fixed effects, patient characteristic covariates, market time-varying covariates, and an error term. We use the first stage as part of a differences-in-differences design, but to do so the $\bar{\pi}_{pt}$ is set to 0 in the pre-period (i.e., there is no BPCI treatment before BPCI began). Thus, in the second stage regression, we then used

hospitalization to a BPCI hospital as an instrument for observed "treatment" at a BPCI or Non-BPCI hospital in the period after BPCI began as part of the difference-in-differences design, relating treatment to spending and quality outcomes. $y_{pt,h,t} = \alpha + \beta * BPCI\_observed_{pt,h,t} + \gamma * Hosp_{FE} + HRR_{FE} + Time_{FE} + \delta * Cov_{pt} + \theta * Cov_{HRR} + \varepsilon$, where the coefficient of interest is β and captures the average effect of BPCI on outcome y (note that $BPCI\_observed_{pt,h,t}$ is time-varying and is 1 for BPCI episodes in the post-period only and 0 otherwise, which is similar to the interaction term in a usual difference-in-difference model. Together with hospital and time fixed effects, the coefficient β on $BPCI\_observed_{pt,h,t}$ gives the difference-in-differences estimate of the effect of BPCI on outcome *y*. While we show these equations separately, this was estimated simultaneously using 2 stage least squares (2SLS) and not in 2 steps. This instrumental variable approach allowed us to measure the effect of BPCI among patients who received LEJR at a BPCI Hospital regardless of BPCI's existence.

Notably, because Non-BPCI hospitals were also likely affected by BPCI when in the same market as a BPCI Hospital (a BPCI Market), we only use beneficiaries admitted to Non-BPCI Hospitals (propensity-matched per Appendix A in our primary analysis) located in Non-BPCI Markets as the comparison group. We also conducted a Hausman test evaluating for endogeneity, finding $p<0.001$ and therefore rejecting the null hypothesis of equivalence (i.e., no endogeneity from unobserved confounding). This result provided additional rationale for the need for this IV to mitigate confounding from unobserved selection. We used bootstrapped standard errors to account for the fact that the IV is an estimated quantity.[5]

Finally, in response to reviewer suggestions, we directly examined potential selection. In our sample, 2,096 surgeons (5.1% of all surgeons) operated at least 1 BPCI and 1 Non-BPCI hospital. Further, among the 6,712 surgeons who work at multiple hospitals, the 2,096 surgeons corresponded to the 31.2% who operated at BPCI and Non-BPCI hospitals, as opposed to only BPCI or only Non-BPCI hospitals. These values suggest that while not widespread, the dynamic of operating at multiple hospitals exists in our sample and could give rise to selection.

To further examine and illustrate changes in hospital referral patterns, we evaluated how hospital referral patterns changed with BPCI participation for surgeons working at BPCI and Non-BPCI hospitals. Specifically, for the sample of patients receiving LEJR from surgeons operating at both BPCI and Non-BPCI hospitals, we examined differential changes at BPCI hospitals (those that become a BPCI hospital in the BPCI period) vs. non-BPCI hospitals with respect to the mean predicted probability of admission to a BPCI hospital. This difference-in-differences analysis provides insight about whether, among surgeons who operate at both BPCI and Non-BPCI in both the pre-BPCI and BPCI periods, the probability of referring a patient to a BPCI vs Non-BPCI hospital changes after the surgeon begins participating in the program. We found that there was indeed a differential change in likelihood of operating on patients in BPCI vs Non-BPCI hospitals (diff-in-diff estimate 6.6%, p=0.0127) associated with BPCI participation.

More broadly, there are multiple forms of selection that could occur under BPCI. In particular, hospitals can also ‘sort’ patients and drive selection through practice acquisition, development of clinically integrated networks, hiring physicians, and other activities. To

examine selection more broadly (and in a fashion more germane to BPCI and Medicare vis-à-vis the policy question), we conducted an additional difference-in-differences analysis evaluating the effect of BPCI participation on changes in hospital referral patterns across all episodes in our sample. We observed differential changes in the probability of being admitted to a BPCI hospital coincident with hospital participation in BPCI (diff-in-diff estimate 1.1%, $p<0.0001$).

Taken together, these analyses suggest that while the magnitude of selection is not expansive, it does occur on the margin. Our findings also corroborate that big shifts in referral patterns for large amounts of patients may be difficult and perhaps undesirable for BPCI hospitals. Regardless, these results highlight that while the magnitude of selection is small, it has a significant and outsized impact on BPCI program outcomes.

Appendix Exhibit 3. Covariate balance between Non-BPCI and BPCI hospital groups by values of the instrumental variable, using all Non-BPCI episodes

	Original Observed Data			Quartile 1			Quartile 2			Quartile 3			Quartile 4		
	Non-BPCI	BPCI	SMD	Non-BPCI	BPCI	SMD	Non-BPCI	BPCI	SMD	Non-BPCI	BPCI	SMD	Non-BPCI	BPCI	SMD
Episodes, No.	243,880	177,756	NA	161,135	6,685	NA	120,051	47,769	NA	73,244	94,576	NA	36,601	131,219	NA
Age, mean year (SD)	73.0*	73.0*	-0.001	73.0*	72.8*	-0.03	73.2***	72.9***	-0.03	73.3***	73.2***	-0.02	73.0*	73.2*	0.02
Elixhauser co-morbidity index, mean (SD)	4.4***	4.2***	-0.02	4.4*	4.9*	0.05	4.6**	4.8**	0.02	4.8***	4.4***	-0.04	4.6***	4.2***	-0.04
Black, %	5.6***	6.6***	0.04	5.9	5.6	-0.01	4.7**	5.0**	0.02	6.1***	7.0***	0.04	7.4	7.2	-0.01
Female, %	62.8***	63.9***	0.02	63.2	62.5	-0.02	63.0**	62.3**	-0.02	63.9**	64.7**	0.02	63.0***	64.9***	0.04
Dual-eligible, %	11.3***	10.5***	-0.03	12.2*	11.3*	-0.03	11.4**	11.9**	0.02	11.2	11.0	-0.01	11.3***	10.4***	-0.03
Prior acute care hospital use, %	16.3***	14.6***	-0.05	16.5	16.6	0.003	16.6	16.5	-0.002	18.1***	15.5***	-0.07	18.4***	14.5***	-0.11
Prior IRF use, %	1.4***	1.2***	-0.01	1.3	1.4	0.02	1.4	1.4	-0.0002	1.4***	1.2***	-0.02	1.7***	1.3***	-0.03
Prior SNF use, %	4.5***	4.0***	-0.03	4.5***	5.6***	0.05	4.4	4.4	0.001	5.0***	4.2***	-0.04	4.8***	3.8***	-0.05

This table SMD=Standardized Mean Difference. *p<0.05, **p<0.01, ***p<0.001

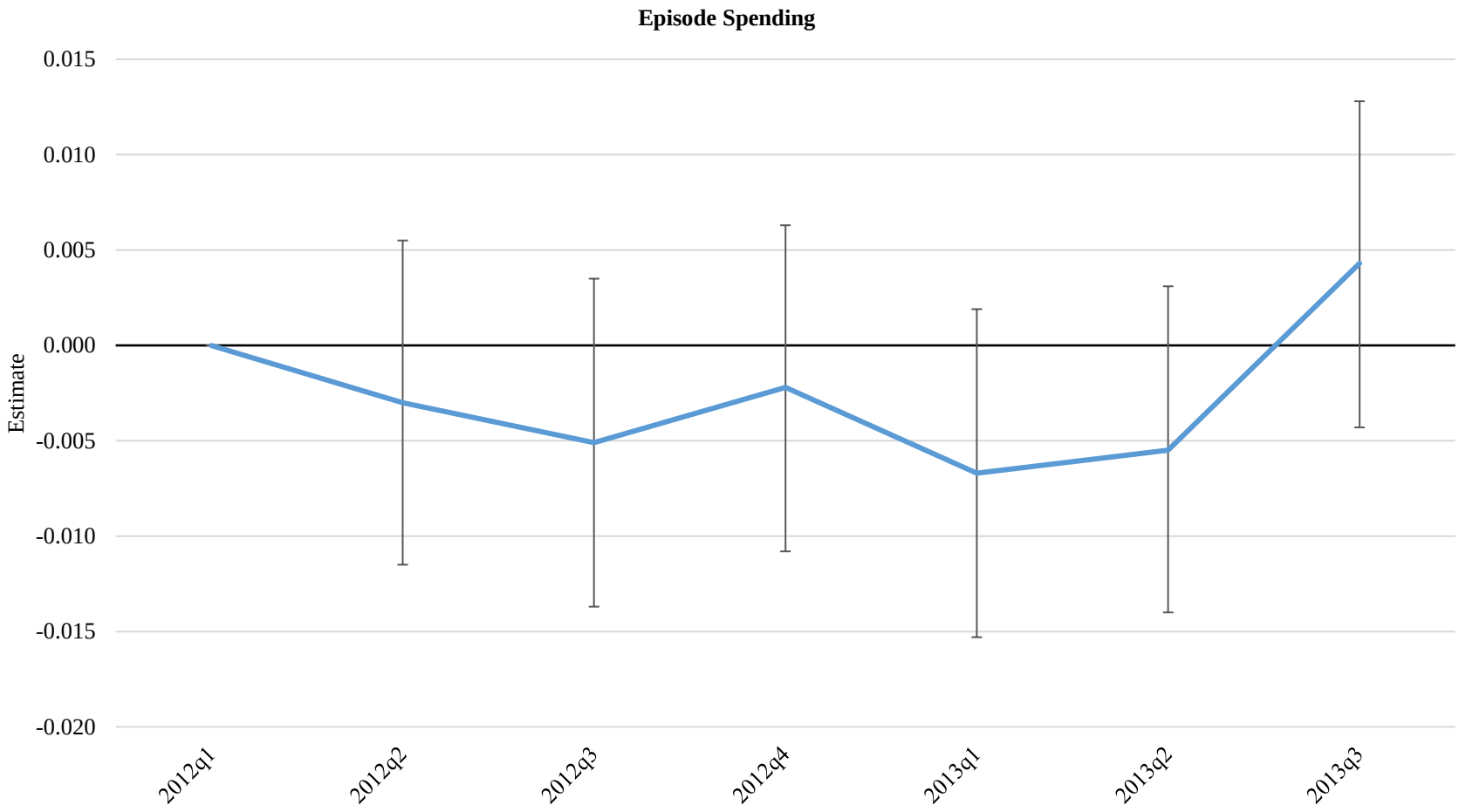
Appendix Exhibit 4. Covariate balance between Non-BPCI and BPCI hospital groups across values of the instrumental variable, using only Non-BPCI episodes in BPCI markets

	Original Observed Data			Quartile 1			Quartile 2			Quartile 3			Quartile 4		
	Non-BPCI	BPCI	SMD	Non-BPCI	BPCI	SMD	Non-BPCI	BPCI	SMD	Non-BPCI	BPCI	SMD	Non-BPCI	BPCI	SMD
Episodes, No.	131,505	177,756	NA	54,014	23,301	NA	38,309	39,006	NA	25,543	51,773	NA	13,639	63,676	NA
Age, mean year (SD)	73.2***	73.0***	-0.02	73.2***	72.6***	-0.07	73.2***	72.9***	-0.03	73.3*	73.1*	-0.01	73.0	73.1	0.01
Elixhauser co-morbidity index, mean (SD)	4.6***	4.2***	-0.04	4.6	4.6	-0.005	4.5*	4.2*	-0.03	4.8***	4.2***	-0.06	4.5***	4.0***	-0.05
Black, %	5.4***	6.6***	0.05	4.9*	5.4*	0.02	4.5***	5.4***	0.04	7.4	7.6	0.01	6.3***	7.1***	0.03
Female, %	62.9***	63.9***	0.02	62.8***	61.4***	-0.03	62.7*	63.4*	0.01	63.3*	64.2*	0.02	62.4***	64.8***	0.05
Dual-eligible, %	11.4***	10.5***	-0.03	12.3***	11.2***	-0.04	10.2	10.6	0.01	11.9***	11.0***	-0.03	10.3*	9.8*	-0.02
Prior acute care hospital use, %	17.1***	14.6***	-0.07	16.5*	15.9*	-0.02	16.8***	14.7***	-0.06	18.4***	14.9***	-0.09	17.6***	13.7***	-0.11
Prior IRF use, %	1.5***	1.2***	-0.02	1.5	1.4	-0.01	1.4***	1.0***	-0.03	1.5***	1.2***	-0.03	1.8***	1.3***	-0.04
Prior SNF use, %	4.8***	4.0***	-0.04	4.7	4.5	-0.01	4.5***	4.0***	-0.03	5.5***	4.2***	-0.06	4.8***	3.6***	-0.06

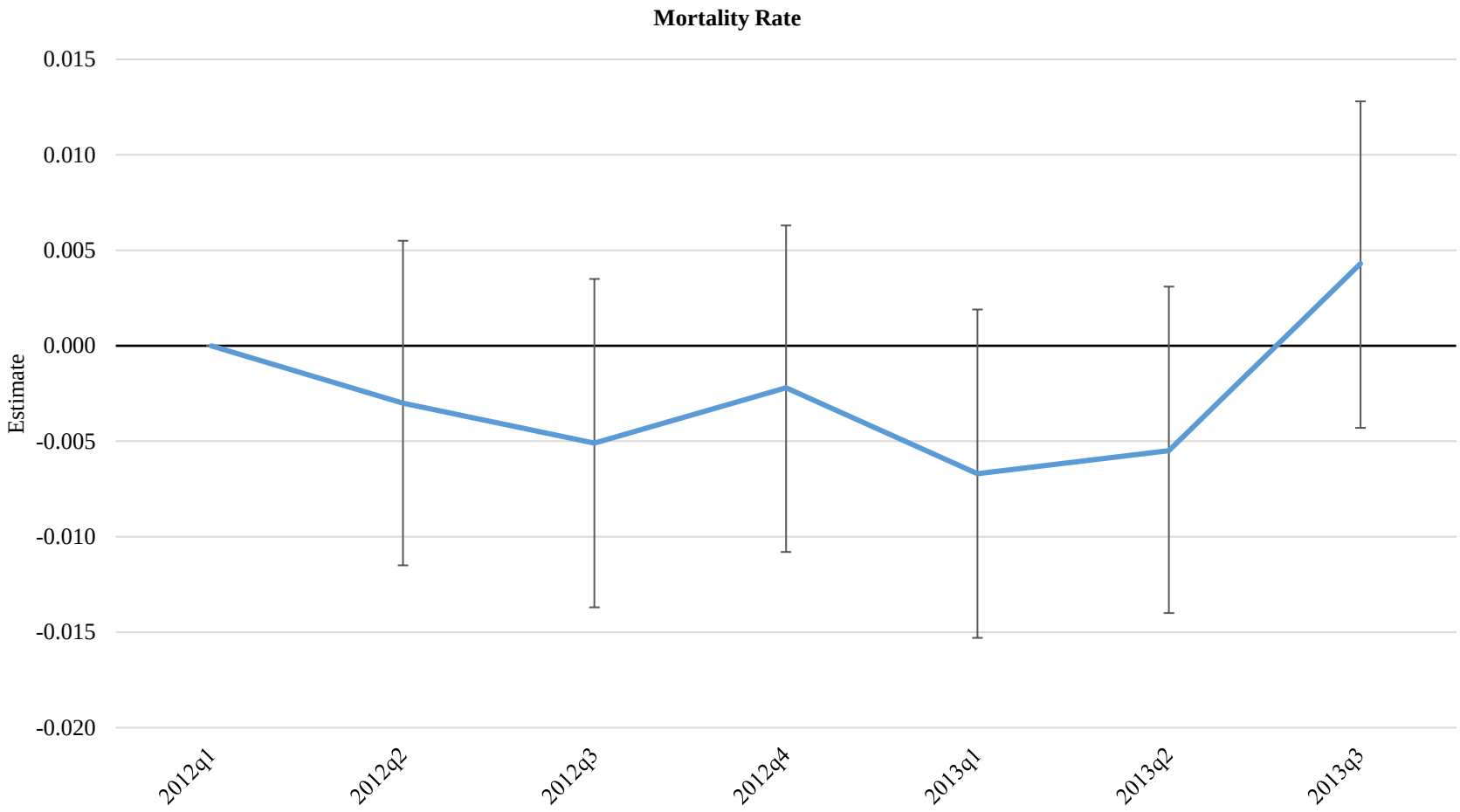
SMD=Standardized Mean Difference. *p<0.05, **p<0.01, ***p<0.001

Appendix Methods 3. Tests of parallel trends between BPCI and Non-BPCI hospitals for spending and quality outcomes

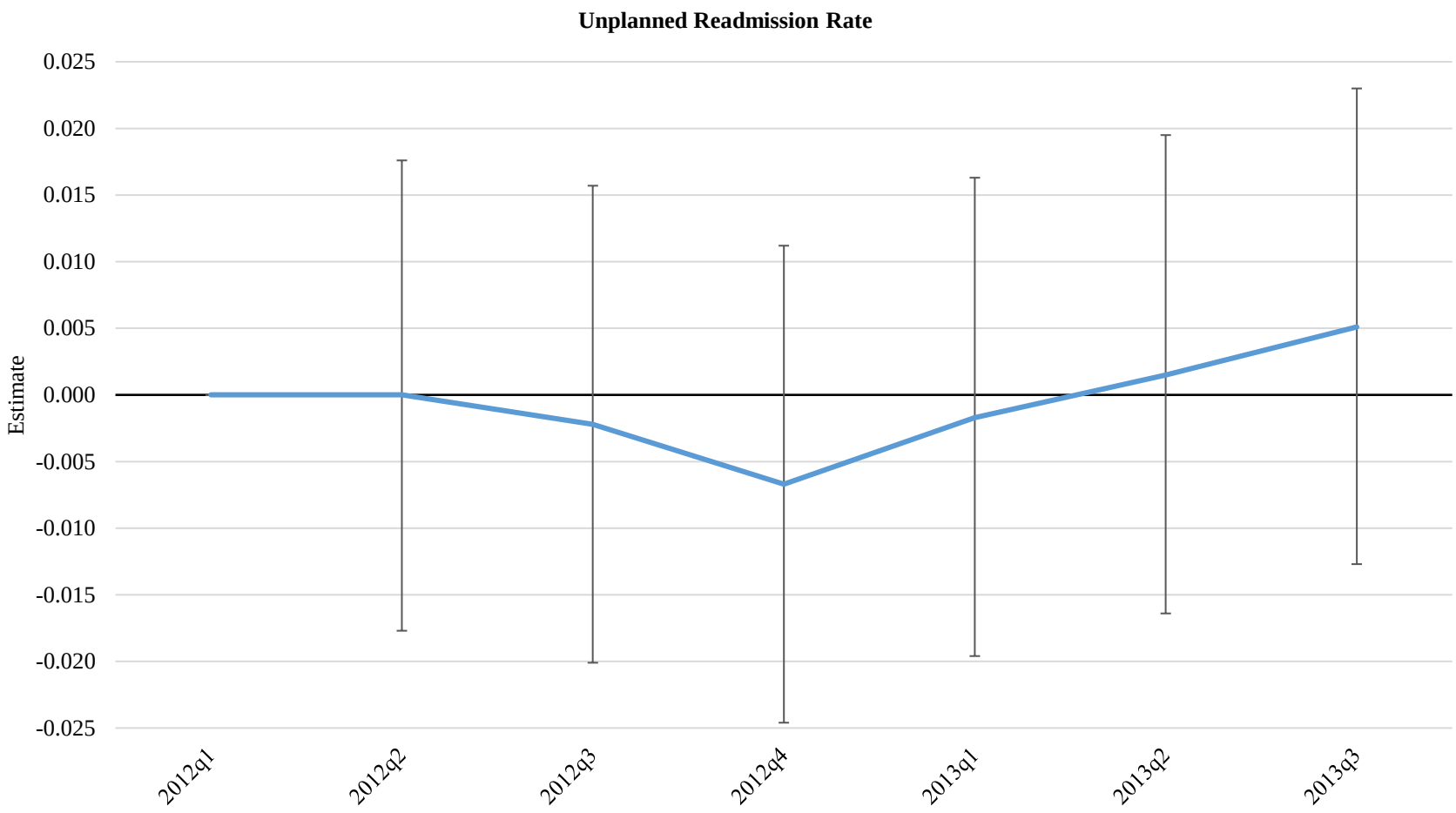
This series of graphs shows the results of generalized linear regression models for each outcome as the dependent variable, and independent variables of a time (quarter) fixed effects, BPCI Hospital indicator variable, and the interaction. The estimates plotted show that the interaction term coefficients are not statistically significant, indicating no divergent trends in the pre-period for any outcome variable. The estimates plotted show that the interaction term coefficients are not statistically significant, indicating a lack of divergent trends in the pre-period for any outcome variable. Similarly, tests conducted using a linear, instead of categorical, time variable demonstrated no statistically significant differential trends.



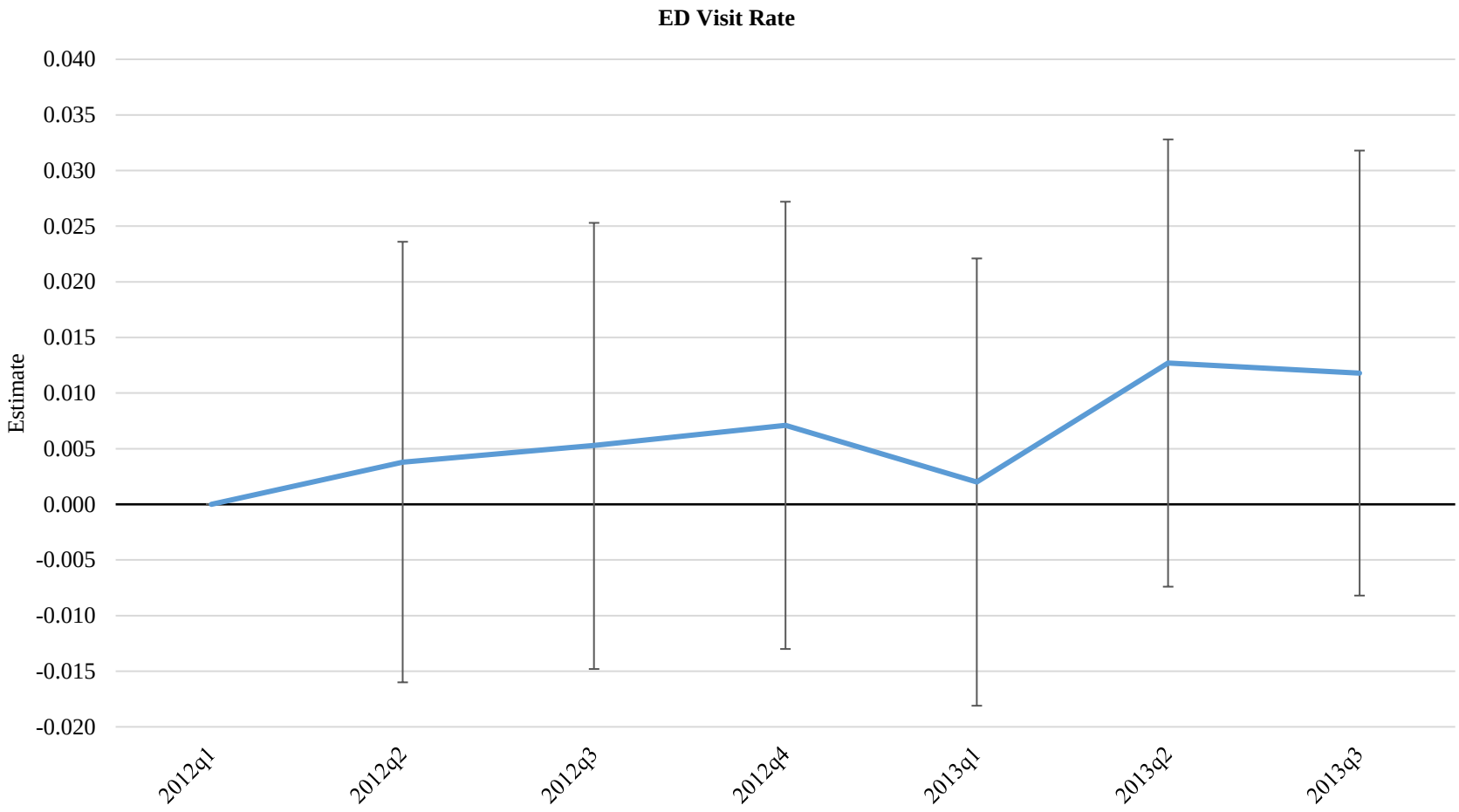
p-values	.	0.85	0.78	0.47	0.62	0.24	0.93



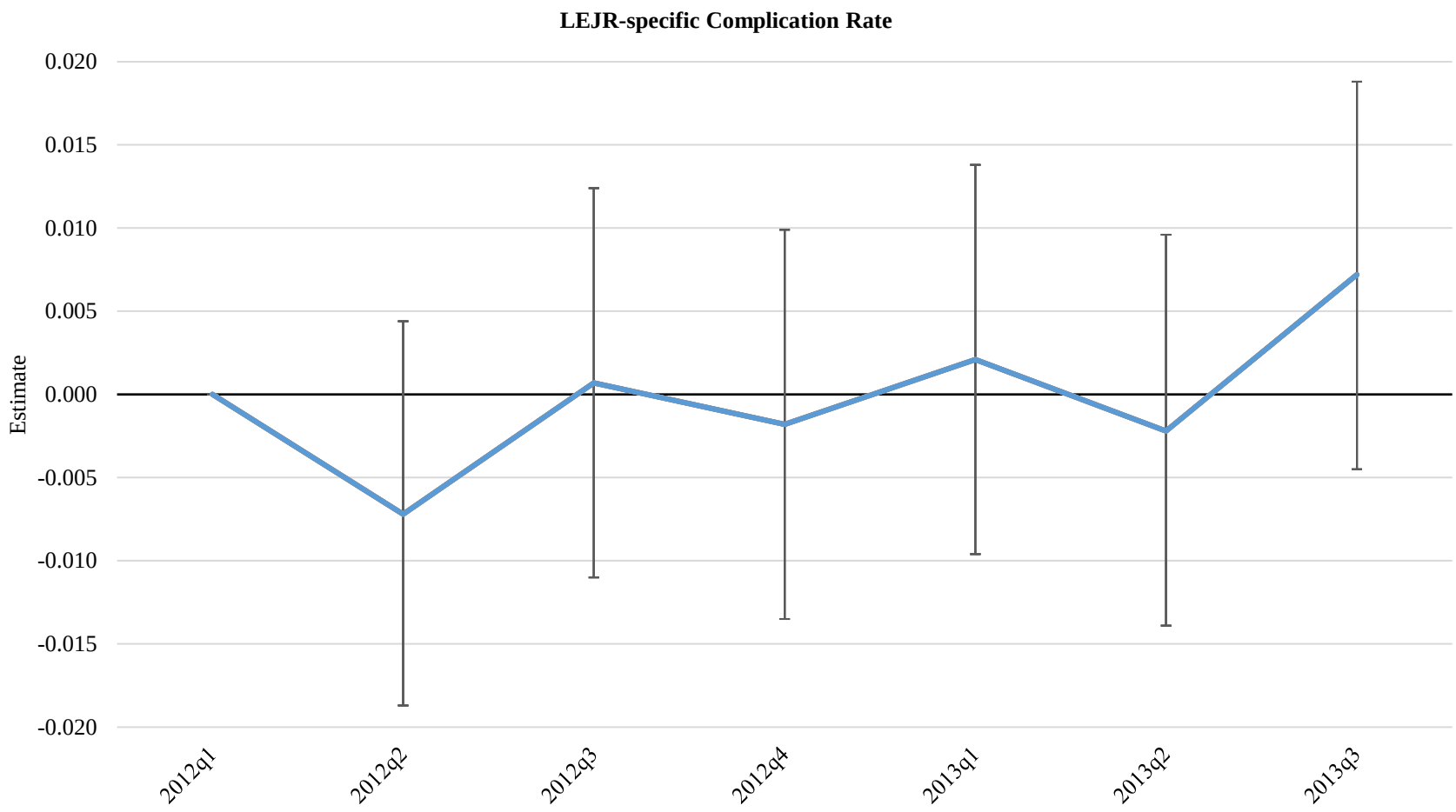
p-values	.	0.49	0.25	0.61	0.13	0.21	0.33



p-values	.	1.00	0.81	0.47	0.86	0.87	0.57



p-values	.	0.71	0.61	0.49	0.84	0.22	0.25



p-values	.	0.22	0.91	0.76	0.73	0.72	0.23

Appendix Exhibit 5. Patient characteristics by BPCI participation and program period, 2012-2016 (Full Table)

Sample Characteristics[a]						
	Non-BPCI Patients			BPCI Patients		
	Pre-BPCI	Early BPCI	Late BPCI	Pre-BPCI	Early BPCI	Late BPCI
Markets, No.	98	98	98	123	123	123
Hospitals, No.	244	244	241	244	244	244
Beneficiaries, No.	20,497	19,734	11,876	75,614	75,297	54,744
Patient Characteristics						
	Non-BPCI Patients			BPCI Patients		
	Pre-BPCI	Early BPCI	Late BPCI	Pre-BPCI	Early BPCI	Late BPCI
Age, mean	73.4 (2.6)	73.4 (2.7)	73.0 (3.3)	73.3* (1.9)	73.0* (1.7)	73.0* (1.8)
Black race, %[b]	8.7 (14.5)	8.3 (13.1)	7.9 (13.6)	7.6 (10.7)	7.2 (10.6)	7.1 (10.9)
Female, %	64.4 (8.5)	63.4 (10.3)	62.2 (13.9)	66.1*** (5.6)	64.2*** (4.8)	64.2*** (5.9)
Dual-eligible, %[c]	13.3 (10.9)	12.5 (11.0)	11.8 (11.0)	15.5* (11.2)	14.2* (10.5)	13.5* (10.3)
Elixhauser comorbidity index, mean[d,e]	5.5*** (3.1)	5.3*** (2.9)	4.4*** (3.3)	5.4*** (2.8)	4.8*** (2.1)	4.3*** (2.1)
Prior acute care hospital use, %[e]	18.9** (10.1)	17.5** (8.9)	16.3** (10.9)	17.9*** (7.1)	15.7*** (4.6)	15.3*** (4.6)
Prior IRF use, %[e,f]	1.3 (2.5)	1.6 (3.2)	1.4 (3.1)	1.4 (1.8)	1.2 (1.6)	1.2 (1.6)
Prior SNF use, %[e,g]	5.2 (5.6)	5.3 (5.2)	4.9 (8.0)	5.1* (3.8)	4.7* (2.7)	4.6 (3.1)

This table describes patient characteristics in the pre-BPCI and BPCI periods for patients receiving LEJR at BPCI Hospitals and Non-BPCI Hospitals. Differential changes for BPCI and Non-BPCI Patients are shown in Appendix Exhibit 3. Pre-Period=January 2011 to September 2013. BPCI Period=October 2013 to December 2016. Early BPCI=October 2013 to June 2015. Late BPCI=July 2015 to December 2016; however, data presented were drawn from LEJR episodes occurring between July 2015 and September 2016 in order to allow for 90-day post-discharge period. Wilcoxon rank-sum or t-tests were used to test the differences in continuous variables and Chi-square tests for categorical variables. *p<0.05, **p<0.01, ***p<0.001. [a]Characteristics for Non-BPCI hospitals and patients were drawn from a 20% Medicare claims sample while characteristics for BPCI hospitals and patients were drawn from a 100% sample. [b]Race was broken out as black versus others because of existing disparities in access to LEJR among black patients specifically. [c]Dual eligible indicates eligibility for both the Medicare and Medicaid programs as an indicator of low socioeconomic status. [d]The Elixhauser comorbidity score is an index of severity with a range of −32 to +92 with increasing scores highly correlated with increased probability of in-hospital death. [e]Calculated using data from the year prior to LEJR hospitalization. [f]Inpatient Rehabilitation Facility. [g]Skilled Nursing Facility.

Appendix Exhibit 6. Changes in patient and market characteristics by BPCI participation and program period, 2012-2016

Characteristics of Beneficiaries Receiving LEJR[a]								
	Non-BPCI Hospitals			BPCI Hospitals				
	Pre-BPCI	BPCI	Difference	Pre-BPCI	BPCI	Difference	DiD[b]	DiD P-Value[b]
Patient Characteristics								
Age, mean year (SD)	73.4 (2.6)	73.2 (2.4)	-0.2	73.3 (1.9)	73.0 (1.6)	-0.3	-0.1	0.61
Black, % (SD)[c]	8.7 (14.5)	8.1 (12.8)	-0.5	7.6 (10.7)	7.2 (10.8)	-0.4	0.1	0.93
Female, % (SD)	64.4 (8.5)	63.0 (8.9)	-1.3	66.1 (5.6)	64.2 (4.1)	-1.8	-0.5	0.61
Dual-eligible, % (SD)[d]	13.3 (10.9)	12.0 (8.9)	-1.3	15.5 (11.2)	13.9 (10.2)	-1.6	-0.3	0.80
Elixhauser comorbidity index, mean (SD)[e,f]	5.5 (3.1)	5.0 (2.6)	-0.5	5.4 (2.8)	4.5 (1.9)	-0.9	-0.3	0.32
Prior acute care hospital use, % (SD)[f]	18.9 (10.1)	16.9 (7.2)	-2.0	17.9 (7.1)	15.6 (3.8)	-2.4	-0.4	0.70
Prior IRF use, % (SD)[f,g]	1.3 (2.5)	1.6 (2.6)	0.2	1.4 (1.8)	1.2 (1.4)	-0.2	-0.5	0.09
Prior SNF use, % (SD)[f,h]	5.2 (5.6)	5.0 (4.1)	-0.3	5.1 (3.8)	4.7 (2.3)	-0.5	-0.2	0.67
Market Structure Characteristics[i]								
	Non-BPCI Hospitals			BPCI Hospitals				
	Pre-BPCI	BPCI	Difference	Pre-BPCI	BPCI	Difference	DiD[b]	DiD P-Value[b]
Quarterly LEJR volume, mean (SD)	303.4 (258.0)	326.5 (281.9)	23.0	375.6 (316.2)	408.2 (348.7)	32.5	9.5	0.87
Hospital beds, mean (SD)	3,124.2 (2,870.4)	3,153.9 (2,887.9)	29.7	4,283.4 (4,633.9)	4,320.5 (4,670.5)	37.1	7.4	0.99
SNF beds, mean (SD)[h]	5,145.8 (4,239.6)	5,118.2 (4,244.8)	-27.6	6,758.1 (6,259.5)	6,753.3 (6,243.9)	-4.8	22.8	0.98
MA penetration, % (SD)[j]	24.2 (13.8)	28.1 (13.6)	3.9	26.8 (12.9)	30.7 (13.4)	3.9	0.0	1.00
ACO penetration, % (SD)[k]	3.8 (4.4)	14.8 (11.5)	11.1	4.7 (4.8)	21.7 (13.7)	17.0	5.9	0.001

Hospital concentration, HHI (SD)[l,m]	3,237.0 (1,950.8)	3,256.8 (1,893.8)	19.8	2,697.4 (1,788.7)	2,748.6 (1,824.0)	51.2	31.4	0.93
SNF concentration, HHI (SD)[h,m]	1,398.8 (1,074.0)	1,239.3 (860.2)	-159.5	1,172.5 (822.6)	1,129.3 (788.9)	-43.1	116.4	0.49
Markets with PGP, %[n]	0.0	46.9	46.9	0.0	53.7	53.7	6.72	0.32

This table describes patient and market characteristics in the pre-BPCI and BPCI periods for patients receiving LEJR at BPCI Hospitals and Non-BPCI Hospitals. Pre-Period=January 2011 to September 2013. BPCI Period=October 2013 to December 2016. Early BPCI=October 2013 to June 2015. Late BPCI=July 2015 to December 2016; however, data presented were drawn from LEJR episodes occurring between July 2015 and September 2016 in order to allow for 90-day post-discharge period. Wilcoxon rank-sum or t-tests were used to test the differences in continuous variables and Chi-square tests for categorical variables. Diff-in-diff Estimate=differences-in-differences estimate. [a]Characteristics for Non-BPCI hospitals and patients were drawn from a 20% Medicare claims sample while characteristics for BPCI hospitals and patients were drawn from a 100% sample. [b]DiD=difference-in-differences estimate. [c]Race was broken out as black versus others because of existing disparities in access to LEJR among black patients specifically. [d]Dual eligible indicates eligibility for both the Medicare and Medicaid programs as an indicator of low socioeconomic status. [e]The Elixhauser comorbidity score is an index of severity with a range of −32 to +92 with increasing scores highly correlated with increased probability of in-hospital death. [f]Calculated using data from the year prior to LEJR hospitalization. [g]Inpatient Rehabilitation Facility. [h]Skilled Nursing Facility. [i]Market characteristics are calculated based on total procedural (episode) volume using a 100% sample rather than based on unique patients as in the rest of the table. [j]Medicare Advantage penetration was determined using the 100% Medicare Beneficiary Summary File and computed at the market–quarter level for the proportion of Medicare beneficiaries enrolled in Medicare Advantage at any time during that quarter. [k]ACO penetration (number of beneficiaries in a hospital referral region attributed to a Medicare ACO out of all Medicare beneficiaries) was determined using data from a random 20% sample of fee-for-service beneficiaries and the CMS ACO Provider-level Research Identifiable File on a yearly basis. [l]Herfindahl-Hirschman Index. [m]Hospital and skilled nursing facility concentration was determined using the Herfindahl-Hirschman index.[6] [n]Physician Practice Group. Markets with Physician Group Practice indicates markets with a physician group practice participating in BPCI for the LEJR condition.

Appendix Exhibit 7. Market characteristics by BPCI participation and program period, 2012-2016

Market Characteristics[a]	Non-BPCI Markets			BPCI Markets		
	Pre-BPCI	Early BPCI	Late BPCI	Pre-BPCI	Early BPCI	Late BPCI
Quarterly LEJR volume, mean	303 (258)	315 (273)	342 (294)	376 (316)	394 (337)	428 (367)
Hospital beds, mean	3,124 (2,870)	3,151 (2,887)	3,158 (2,890)	4,283 (4,634)	4,311 (4,657)	4,334 (4,690)
SNF beds, mean	5,146 (4,240)	5,129 (4,247)	5,104 (4,243)	6,758 (6,260)	6,770 (6,266)	6,730 (6,215)
MA penetration, %[b]	24.2* (13.8)	27.4* (13.6)	29.1* (13.8)	26.8* (12.9)	29.9* (13.2)	31.8* (13.6)
ACO penetration, %[c]	3.8*** (4.4)	13.1*** (11.4)	17.4*** (12.8)	4.7*** (4.8)	19.0*** (13.7)	25.3*** (15.3)
Hospital concentration, HHI[d,e]	3,237 (1,951)	3,253 (1,904)	3,262 (1,885)	2,697 (1,789)	2,738 (1,818)	2,763 (1,841)
SNF concentration, HHI[e,f]	1,399 (1,074)	1,270 (906)	1,197 (818)	1,173 (823)	1,134 (787)	1,124 (806)
Markets with PGP, %[g]	0.0***	26.5***	46.9***	0.0***	27.6***	53.7***

This table describes market characteristics in the pre-BPCI and BPCI periods for patients receiving LEJR at BPCI Hospitals and Non-BPCI Hospitals. Differential changes for BPCI and Non-BPCI Patients are shown in Appendix Exhibit 3. Pre-Period=January 2011 to September 2013. BPCI Period=October 2013 to December 2016. Early BPCI=October 2013 to June 2015. Late BPCI=July 2015 to December 2016; however, data presented were drawn from LEJR episodes occurring between July 2015 and September 2016 in order to allow for 90-day post-discharge period. Wilcoxon rank-sum or t-tests were used to test the differences in continuous variables and Chi-square tests for categorical variables. *p<0.05, **p<0.01, ***p<0.001. [a]Market characteristics are calculated based on total procedural (episode) volume using a 100% sample rather than based on unique patients as in the rest of the table. [b]Medicare Advantage penetration was determined using the 100% Medicare Beneficiary Summary File and computed at the market–quarter level for the proportion of Medicare beneficiaries enrolled in Medicare Advantage at any time during that quarter. [c]ACO penetration (number of beneficiaries in a hospital referral region attributed to a Medicare ACO out of all Medicare beneficiaries) was determined using data from a random 20% sample of fee-for-service beneficiaries and the CMS ACO Provider-level Research Identifiable File on a yearly basis. [d]Herfindahl-Hirschman Index. [e]Hospital and skilled nursing facility concentration was determined using the Herfindahl-Hirschman index (Rhoades SA. The Herfindahl-Hirschman index. *Fed. Res. Bull*. 1993;79:188). [f]Skilled Nursing Facility. [g]Physician Practice Group. Markets with Physician Group Practice indicates markets with a physician group practice participating in BPCI for the LEJR condition.

Appendix Exhibit 8. Unadjusted changes in spending and quality outcomes by BPCI participation, 2012-2016

	Non-BPCI Patients				BPCI Patients					
	Pre-BPCI	**BPCI**	**Difference, %**	**p-value[a]**	**Pre-BPCI**	**BPCI**	**Difference, %**	**p-value[a]**	**Differential change[b]**	**p- value[c]**
Episode spending, $ (SD)	22,834 (13,369)	22,073 (13,177)	-3.3	<0.001	23,552 (13,974)	22,129 (13,752)	-6.0	<0.001	-2.7	<0.001
Mortality rate, %	1.9	1.7	-10.0	0.08	2.0	1.7	-14.8	<0.001	-4.8	0.51
Unplanned readmission rate, %	9.0	8.3	-7.6	0.006	9.5	8.2	-13.2	<0.001	-5.6	0.05
ED visit rate, %[d]	4.8	4.6	-4.4	0.26	13.8	13.7	-0.6	0.6	3.8	0.68
LEJR-specific complica-tion rate, %[e]	3.8	3.8	1.4	0.77	3.8	3.6	-5.8	0.01	-7.2	0.16

Pre-BPCI=January 2012 to September 2013; BPCI=October 2013 to December 2016. All spending estimates were standardized and adjusted for inflation and transformed into 2016 dollars. [a]Obtained from Wilcoxon rank sum tests. [b]Calculated by subtracting the difference between pre-BPCI and BPCI periods among Non-BPCI patients from the difference between pre-BPCI and BPCI periods among BPCI patients. [c]Obtained from two-way ANOVA, with p-value reflecting statistical significance of the interaction term measuring differential change. [d]Emergency Department (ED) visits without hospitalization. [e]Defined by Hospital Compare.

Appendix Exhibit 9. Unadjusted spending components by program period, 2012-2016

Episode Spending Components			
	Pre-BPCI	**Early BPCI**	**Late BPCI**
Index hospitalization, mean % (SD)	64.2 (20.4)	65.7 (19.9)	67.9 (18.8)
Readmissions, mean % (SD)	2.4 (8.0)	2.2 (7.9)	2.1 (7.9)
SNF, mean % (SD)[a]	7.7 (16.1)	7.0 (15.4)	5.1 (13.2)
IRF, mean % (SD)[b]	4.6 (14.1)	3.6 (12.5)	2.3 (9.6)
HHA, mean % (SD)[c]	10.0 (8.6)	10.2 (8.7)	10.6 (9.2)
Professional services, mean % (SD)	7.3 (6.3)	7.6 (6.5)	8.0 (6.8)
Other, mean % (SD)[d]	3.7 (5.7)	3.7 (5.6)	3.9 (5.9)

Pre-BPCI=January 2012 to September 2013; Early BPCI=October 2013 to June 2015; Late BPCI=July 2015 to December 2016. All spending estimates were standardized and adjusted for inflation and transformed into 2016 dollars. [a]Skilled Nursing Facility. [b]Inpatient Rehabilitation Facility. [c]Home Health Agency. [d]Includes durable medical equipment and other outpatient facility and non-professional payments.

Appendix Exhibit 10. Changes in the proportion of episode spending attributable to specific components associated with BPCI participation by program period and timing of hospital participation, 2012-2016

	Overall		Program Period				Timing of Participation					
							Early Entrant Hospitals				Late Entrant Hospitals	
			Early BPCI		Late BPCI		Early BPCI		Late BPCI		Late BPCI	
	DiD[a]	P-Value	DiD	P-Value	DiD	P-Value	DiD	P-Value	DiD	P-Value	DiD	P-Value
Episode spending components												
Index hospitalization, mean % (SD)	0.7 (0.3 to 1.1)	0.001	0.9 (0.5 to 1.4)	<0.001	0.3 (-0.5 to 1.1)	0.45	1.1 (0.6 to 1.6)	<0.001	1.1 (0.2 to 1.9)	0.01	-0.2 (-1.0 to 0.5)	0.56
Re-admissions, mean % (SD)	-0.02 (-0.2 to 0.2)	0.85	-0.02 (-0.2 to 0.2)	0.87	-0.03 (-0.4 to 0.3)	0.84	-0.01 (-0.2 to 0.2)	0.91	-0.1 (-0.4 to 0.3)	0.76	0.02 (-0.3 to 0.4)	0.90
SNF, mean % (SD)[b]	-0.4 (-0.7 to -0.1)	0.02	-0.5 (-0.9 to -0.1)	0.01	-0.3 (-0.9 to 0.3)	0.30	-0.5 (-0.9 to -0.1)	0.01	-0.3 (-0.9 to 0.3)	0.28	-0.3 (-0.9 to 0.3)	0.37
IRF, mean % (SD)[c]	-0.7 (-1.0 to -0.4)	<0.001	-0.5 (-0.9 to -0.2)	0.002	-0.9 (-1.4 to -0.5)	<0.001	-0.7 (-1.0 to -0.3)	<0.001	-1.6 (-2.1 to -1.1)	<0.001	-0.3 (-0.7 to 0.2)	0.28
HHA, mean % (SD)[d]	0.2 (0.04 to 0.4)	0.02	-0.2 (-0.4 to 0.002)	0.05	1.0 (0.7 to 1.4)	<0.001	-0.3 (-0.5 to -0.1)	0.01	0.8 (0.4 to 1.1)	<0.001	1.0 (0.6 to 1.3)	<0.001
Professional services, mean % (SD)	0.2 (0.01 to 0.3)	0.03	0.2 (0.1 to 0.4)	0.01	0.04 (-0.2 to 0.3)	0.76	0.3 (0.1 to 0.5)	0.002	0.1 (-0.1 to 0.4)	0.31	-0.2 (-0.4 to 0.1)	0.24
Other, mean % (SD)[e]	0.002 (-0.1 to 0.1)	0.97	0.1 (-0.1 to 0.2)	0.43	-0.1 (-0.3 to 0.2)	0.48	0.1 (-0.1 to 0.2)	0.42	0.01 (-0.2 to 0.3)	0.95	-0.1 (-0.3 to 0.2)	0.53

This table shows results from difference-in-differences models evaluating the association between BPCI participation and differential changes in episode spending attributable to specific components. Changes are displayed for the overall cohort and study period (Overall) as well as program period (Early BPCI and Late BPCI) and timing of participation (Early Entrant and Late Entrant hospitals). Negative estimates indicate decreases in the proportion of episode spending for the relevant component. Early BPCI=October 2013 to June 2015; Late BPCI=July 2015 to December 2016. BPCI=Early BPCI + Late BPCI (October 2013 to December 2016). All spending estimates were standardized and transformed into 2016 dollars. [a]DiD=differences-in-differences estimate. [b]Skilled Nursing Facility. [c]Inpatient Rehabilitation Facility. [d]Home Health Agency. [e]Includes durable medical equipment and other outpatient facility and non-professional payments.

Appendix Exhibit 11. Changes in quality outcomes associated with BPCI participation by program period and timing of hospital participation, 2012-2016 (Full Table)

	Overall		Program Period				Timing of Participation					
							Early Entrant Hospitals				Late Entrant Hospitals	
			Early BPCI		Late BPCI		Early BPCI		Late BPCI		Late BPCI	
	DiD[a]	95% CI	DiD	95% CI	DiD	95% CI	DiD	95% CI	DiD	95% CI	DiD	95% CI
Mortality rate, %	-0.15	-0.5 to 0.2	-0.03	-0.4 to 0.3	-0.34	-3.1 to 0.4	0.01	-0.4 to 0.4	-0.27	-0.8 to 0.3	-0.37	-0.9 to 0.2
Unplanned re-admission rate, %	0.15	-0.5 to 0.8	0.21	-0.5 to 1.0	0.03	-0.9 to 0.2	0.24	-0.5 to 1.0	-0.10	-1.3 to 1.1	0.18	-1.0 to 1.4
ED visit rate, %[b]	-0.19	-1.0 to 0.6	-0.22	-1.1 to 0.7	-0.08	-1.1 to 1.1	-0.25	-1.2 to 0.6	0.15	-1.1 to 1.4	0.54	-0.8 to 1.9
LEJR-specific complica-tion rate, %[c]	0.12	-0.3 to 0.6	0.21	-0.3 to 0.7	-0.03	-1.3 to 1.1	0.24	-0.2 to 0.7	0.003	-0.8 to 0.8	-0.09	-0.9 to 0.7

This table shows results from difference-in-differences models evaluating the association between BPCI participation and differential changes in quality outcomes, with changes displayed for the overall cohort and study period (Overall) as well as program period (Early BPCI and Late BPCI) and timing of participation (Early Entrant and Late Entrant hospitals). Negative estimates indicate reductions in rates (i.e., quality improvements). Early BPCI=October 2013 to June 2015; Late BPCI=July 2015 to December 2016. [a]DiD =differences-in-differences estimate. [b]Emergency Department (ED) visits without hospitalization. [c]Defined by Hospital Compare.

Appendix Exhibit 12. Sensitivity analysis extending methods from prior CMS evaluations

Adjusted results				
	Sensitivity analysis (without IV or hospital fixed effects)		Primary analysis (with IV[a] and hospital fixed effects)	
	DiD[b] (95% CI)	p-value	DiD[b] (95% CI)	p-value
Episode spending, $ (%)	-2.2 (-2.9 to -1.5)	<0.001	-1.6 (-2.6 to -0.6)	<0.001
Mortality rate, %	-0.3 (-0.5 to 0.0)	0.04	-0.2 (-0.5 to 0.2)	0.35
Unplanned readmission rate, %	-0.2 (-0.7 to 0.4)	0.59	0.1 (-0.5 to 0.8)	0.67
ED visit rate, %[c]	0.2 (-0.5 to 0.8)	0.65	-0.2 (-1.0 to 0.6)	0.63
LEJR-specific complication rate, %[d]	0.2 (-0.2 to 0.5)	0.43	0.1 (-0.3 to 0.6)	0.59

[a]Instrumental variable approach. [b]DiD=difference-in-differences estimate. [c]Emergency Department (ED) visits without hospitalization. [d]Defined by Hospital Compare.

Appendix Exhibit 13. Sensitivity analysis excluding January to September 2013 from the baseline period

	Early BPCI		Late BPCI		Overall BPCI	
	DiD[a] (95% CI)	**p-value**	**DiD[a] (95% CI)**	**p-value**	**DiD[a] (95% CI)**	**p-value**
Episode spending, %	-1.6 (-2.7 to -0.5)	0.01	-0.9 (-2.5 to 0.7)	0.27	-1.4 (-2.3 to -0.4)	0.01
Mortality rate, %	-0.01 (-0.4 to 0.4)	0.97	-0.4 (-0.9 to 0.1)	0.16	-0.2 (-0.5 to 0.2)	0.34
Unplanned readmission rate, %	0.2 (-0.6 to 1.0)	0.60	0.2 (-0.9 to 1.4)	0.70	0.2 (-0.5 to 0.9)	0.57
ED visit rate, %[b]	-0.3 (-1.2 to 0.6)	0.57	0.002 (-1.3 to 1.3)	>0.99	-0.2 (-1.0 to 0.6)	0.63
LEJR-specific complication rate, %[c]	0.2 (-0.3 to 0.7)	0.34	0.1 (-0.7 to 0.9)	0.81	0.2 (-0.3 to 0.6)	0.41

Pre-BPCI=January 2012 to September 2013; Early BPCI=October 2013 to June 2015; Late BPCI=July 2015 to December 2016. All spending estimates were standardized and adjusted for inflation and transformed into 2016 dollars. [a]DiD=difference-in-differences estimate. [b]Emergency Department (ED) visits without hospitalization. [c]Defined by Hospital Compare.

Appendix Exhibit 14. Sensitivity analysis using less stringent propensity matching

	Early BPCI		Late BPCI		Overall BPCI	
	DiD[a] (95% CI)	p-value	DiD[a] (95% CI)	p-value	DiD[a] (95% CI)	p-value
Episode spending, %	-2.4 (-3.6 to -1.2)	<0.001	-1.9 (-3.2 to -0.7)	<0.001	-2.1 (-3.1 to -1.2)	<0.001
Mortality rate, %	-0.1 (-0.5 to 0.2)	0.41	-0.3 (-0.7 to 0.04)	0.08	-0.3 (-0.5 to 0.03)	0.08
Unplanned readmission rate, %	-0.1 (-0.9 to 0.6)	0.72	-0.2 (-1.0 to 0.6)	0.63	-0.2 (-0.8 to 0.4)	0.60
ED visit rate, %[b]	-0.3 (-1.2 to 0.6)	0.48	0.1 (-0.9 to 1.0)	0.87	-0.1 (-0.8 to 0.6)	0.76
LEJR-specific complication rate, %[c]	0.1 (-0.5 to 0.5)	0.85	-0.4 (-0.9 to 0.2)	0.18	-0.2 (-0.6 to 0.2)	0.38

Pre-BPCI=January 2012 to September 2013; Early BPCI=October 2013 to June 2015; Late BPCI=July 2015 to December 2016. All spending estimates were standardized and adjusted for inflation and transformed into 2016 dollars. [a]DiD=difference-in-differences estimate. [b]Emergency Department (ED) visits without hospitalization. [c]Defined by Hospital Compare.

Appendix Exhibit 15. Sensitivity analysis excluding CJR hospitals from propensity score matching and analyses

	Early BPCI		Late BPCI		Overall BPCI	
	DiD[a] (95% CI)	p-value	DiD[a] (95% CI)	p-value	DiD[a] (95% CI)	p-value
Episode spending, %	-1.5 (-2.7 to -0.2)	0.02	-1.4 (-3.0 to 0.3)	0.10	-1.5 (-2.5 to -0.4)	0.01
Mortality rate, %	-0.1 (-0.5 to 0.3)	0.70	-0.4 (-0.8 to 0.1)	0.16	-0.2 (-0.5 to 0.1)	0.23
Unplanned readmission rate, %	0.4 (-0.4 to 1.2)	0.33	-0.3 (-1.4 to 0.8)	0.60	0.1 (-0.6 to 0.8)	0.79
ED visit rate, %[b]	-0.1 (-1.0 to 0.9)	0.92	0.1 (-1.2 to 1.3)	0.90	-0.02 (-0.9 to 0.8)	0.97
LEJR-specific complication rate, %[c]	0.1 (-0.4 to 0.7)	0.62	-0.5 (-1.2 to 0.3)	0.20	-0.1 (-0.6 to 0.4)	0.60

Pre-BPCI=January 2012 to September 2013; Early BPCI=October 2013 to June 2015; Late BPCI=July 2015 to December 2016. All spending estimates were standardized and adjusted for inflation and transformed into 2016 dollars. [a]DiD=difference-in-differences estimate. [b]Emergency Department (ED) visits without hospitalization. [c]Defined by Hospital Compare.

References

1. Parsons LS. Performing a 1:N case control match on propensity score. Poster presented at: SAS Users Group International 29; May, 2004, Montreal, Canada.
2. Rosenbaum PR. Optimal Matching for Observational Studies. *J Am Stat Assoc.* 1989;84(408):1024-1032.
3. Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. *Econometrica.* 1994;62(2):467-475.
4. Angrist JD, Krueger AB. Instrumental variables and the search for identification: from supply and demand to natural experiments. *J Econ Perspect.* 2001;15(4):69-85.
5. Angrist J, Pischke J. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press; 2008.
6. Rhoades SA. The Herfindahl-Hirschman index. *Fed Res Bull*. 1993;79:188.